

UNIVERSIDAD DE GRANADA
E.T.S. DE INGENIERÍAS INFORMÁTICA Y DE
TELECOMUNICACIÓN



ugr | Universidad
de Granada

Departamento de Ciencias de la Computación
e Inteligencia Artificial

Programa de Doctorado en
Tecnologías de la Información y la Comunicación

**Modelización e Inferencia Bayesiana en Problemas de
Reconstrucción y Clasificación de Imágenes**

**Bayesian Modeling and Inference in Image Recovery and
Classification Problems**

Tesis Doctoral

Pablo Ruiz Matarán

Directores: Rafael Molina Soriano, Javier Mateos Delgado y Aggelos K. Katsaggelos

Editorial: Universidad de Granada. Tesis Doctorales
Autor: Pablo Ruiz Matarán
ISBN: 978-84-9125-241-2
URI: <http://hdl.handle.net/10481/40870>

La memoria titulada “Modelización e Inferencia Bayesianas en problemas de Reconstrucción y Clasificación de Imágenes”, que presenta Don Pablo Ruiz Matarán, para optar al grado de Doctor, ha sido realizada en el Departamento de Ciencias de la Computación e Inteligencia Artificial de la Universidad de Granada bajo la dirección de los Doctores Rafael Molina Soriano y Javier Mateos Delgado y el Profesor Aggelos K. Katsaggelos de la Universidad de Northwestern (Illinois).

El doctorando Pablo Ruiz Matarán y los directores de la tesis Doctores Rafael Molina Soriano, Javier Mateos Delgado y Aggelos K. Katsaggelos, garantizamos, al firmar esta tesis doctoral, que el trabajo ha sido realizado por el doctorando bajo la dirección de las directores de la tesis y hasta donde nuestro conocimiento alcanza, en la realización del trabajo, se han respetado los derechos de otros autores a ser citados, cuando se han utilizado sus resultados o publicaciones.

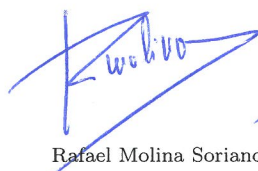
Granada, Junio de 2015

El doctorando

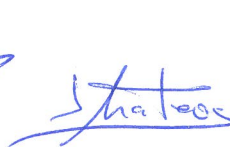


Pablo Ruiz Matarán

Los directores



Rafael Molina Soriano



Javier Mateos



Aggelos K. Katsaggelos

Agradecimientos

Quiero dedicar esta Tesis a todas aquellas personas que me han apoyado durante este tiempo y que de algún modo u otro, me han ayudado a hacerla posible.

Han sido varios años de trabajo, que por fin, pueden resumirse en esta memoria, y por ello, quiero agradecer a mis directores: Rafael Molina, Javier Mateos y Aggelos Katssagelos, que confiaran en mí para realizar esta Tesis, el esfuerzo que han dedicado a mi formación durante todo este tiempo y su capacidad para resolver todos los problemas que nos hemos encontrado por el camino. Muchas gracias por todo, para mí es un privilegio trabajar con vosotros.

También quiero agradecer, a todos los compañeros y compañeras con los que hemos colaborado para llevar a cabo los trabajos que presentamos: Nicolás Pérez de la Blanca, Xu Zhou, Derin Babacan, Gustavo Camps Valls, Hiram Madero, Concepción Cárdenas, Shinichi Nakajima y Li Gao.

Por supuesto no quiero olvidarme de agradecer a toda mi familia y amigos que me han apoyado durante todos estos años, y aunque me haya perdido algunas reuniones con vosotros, siempre habéis seguido estando ahí, sobre todo se lo quiero agradecer a Gloria, que sabe perfectamente de lo que hablo.

Por último, quiero agradecerles a todos mis compis del CITIC, que desde que llegué a la ETSIIT siempre han estado dispuestos a echarme una mano, y con los que he compartido muchos momentos que han hecho que estos años sean inolvidables.

¡Muchas gracias a todos!

Pablo

Contents

Resumen y Conclusiones	ix
Summary and Conclusions	xvii
1 Introduction	1
1.1 Introduction	1
1.1.1 Image Restoration and Blind Image Deconvolution	2
1.1.2 Multispectral Image Classification and Active Learning	3
1.1.3 Light Field Acquisition	4
1.1.4 Video Retrieval	5
1.2 Outline	5
2 Problem Formulation and Objectives of the Thesis	7
2.1 Bayesian Modeling	7
2.1.1 Prior models	8
2.1.2 Observation Model	9
2.1.3 <i>Hyperprior</i> Models	9
2.2 Bayesian Inference	10
2.2.1 Variational Inference	11
2.3 Objectives of the Ph.D. Thesis	12
2.3.1 Image Restoration and Blind Deconvolution	12
2.3.2 Multispectral Image Classification and Active Learning	13
2.3.3 Other Related problems (Light Field Acquisition and Video Retrieval	13
3 Image Restoration and Blind Deconvolution	15
3.1 Image Restoration	15
3.1.1 Image Deblurring Combining Poisson Singular Integral and Total Variation Prior Models	15
3.1.2 Combining Poisson Singular Integral and Total Variation prior models in Image Restoration	23
3.2 Blind Deconvolution	57

4	Multispectral Image Classification (I). Image Processing for Classification	89
4.1	Interactive Classification Oriented Superresolution of Multispectral Images	89
4.2	Learning Filters in Gaussian Process Classification Problems	97
5	Multispectral Image Classification (II). Active Learning	105
5.1	A Bayesian Active Learning Framework for a Two-Class Classification Problem	105
5.2	Bayesian Active Remote Sensing Image Classification	119
5.3	Bayesian Classification and Active Learning Using ℓ_p -Priors. Application to Image Segmentation	131
6	Other Related Problems	139
6.1	Light field Acquisition	139
6.1.1	Compressive Light Field Sensing	139
6.1.2	Light field acquisition from blurred observations using a programmable coded aperture camera	169
6.2	Video Retrieval	177
6.2.1	Video Retrieval Using Sparse Bayesian Reconstruction	177
6.2.2	Retrieval of Video Clips with Missing Frames using Sparse Bayesian Reconstruction	185
7	Conclusions and Future Works	195
7.1	Conclusions	195
7.1.1	Image Restoration and Blind Deconvolution	195
7.1.2	Multispectral Image Classification Problems	196
7.1.3	Other Related Problems (Light Field Acquisition and Video Retrieval)	197
7.2	Future Works	198
7.2.1	Image Restoration and Blind Deconvolution	198
7.2.2	Multispectral Image Classification Problems	198
7.2.3	Other Related Problems	198
	Bibliography	201

Resumen y Conclusiones

Introducción

Una buena parte de las investigaciones y aplicaciones en reconstrucción y clasificación de imágenes se centran en la resolución de problemas inversos, esto es, a partir de un evento observado, encontrar las causas más probables que lo han provocado. Estos problemas han sido abordados siguiendo numerosas aproximaciones. Las estructuras espectrales y espaciales de las imágenes, provocan una alta correlación entre los píxeles, que puede ser explotada explícitamente por métodos probabilísticos (véase por ejemplo, [1, 2, 3, 4]).

Problemas de reconstrucción y clasificación de imágenes como restauración de imágenes [5, 6, 7], deconvolución ciega [8, 9, 10], super-resolución [11, 12, 13], adquisición del campo de luz [14, 15, 16], *pansharpening* [17], clasificación de imágenes multiespectrales [18, 19, 20, 21], aprendizaje activo [22, 23, 24], recuperación de vídeo [25, 26, 27], compresión de vídeo [28], vídeo vigilancia [29], reconocimiento de caras [30, 31, 32], registrado de imágenes [33] y el tratamiento de imágenes médicas [34] entre otros, pueden abordarse usando modelización e inferencia bayesiana.

Un principio fundamental de la filosofía bayesiana es considerar todos los parámetros y variables no observadas como cantidades estocásticas, asignándoles distribuciones de probabilidad basadas en creencias. Por ejemplo, a veces en reconstrucción de imágenes, la imagen original, el ruido de la observación, e incluso la función que define el proceso de adquisición de la observación, son tratadas como variables aleatorias, asignándoles funciones de densidad que modelan el conocimiento disponible sobre la naturaleza de las imágenes y el proceso de formación de la imagen observada.

Dentro de la inferencia bayesiana, los métodos variacionales bayesianos (VB) han atraído el interés de la comunidad estadística bayesiana, la dedicada al aprendizaje automático, así como la de otras áreas relacionadas. La mayor desventaja de los métodos de inferencia tradicionales, como máxima verosimilitud o máximo a posteriori, es que no hacen uso de la información que aporta la distribución a posteriori. El algoritmo EM requiere un conocimiento completo de la distribución a posteriori que, en muchas ocasiones, no puede ser calculada. Los métodos de simulación consiguen obtener la distribución a posteriori pero en la práctica son métodos computacional-

mente muy costosos. Los métodos VB [35, 36, 37, 1, 38, 39, 40] consiguen superar estas limitaciones aproximando la distribución a posteriori desconocida por una distribución más simple y analíticamente tratable y, por tanto, extender la aplicabilidad de la inferencia bayesiana a un mayor rango de modelizaciones. Por ejemplo, permite usar con facilidad distribuciones a priori más complejas (las cuales son muchas veces necesarias en problemas de procesamiento de imágenes) consiguiendo mejorar la exactitud de las estimaciones.

En esta tesis doctoral, exploramos la aplicación de la modelización e inferencia bayesiana a los siguientes problemas: restauración de imágenes, deconvolución ciega, clasificación de imágenes multiespectrales, aprendizaje activo, adquisición de campos de luz y recuperación de vídeo. De esta forma demostramos la amplia aplicabilidad de esta metodología para resolver un amplio rango de problemas de procesamiento de imágenes y clasificación. La memoria contiene contribuciones generales y específicas. Incluye una revisión de los modelos bayesianos que se han aplicado al problema de deconvolución ciega, así como contribuciones en problemas muy específicos. Este formato de tesis abordando una amplia gama de problemas hace que sea particularmente útil para cualquiera que esté interesado en aprender sobre modelización e inferencia bayesiana.

Estructura de la Tesis Doctoral

El principal objetivo de esta tesis doctoral es el estudio de la modelización e inferencia bayesiana y su aplicación a problemas de reconstrucción y clasificación de imágenes, que hemos agrupado en tres bloques: restauración de imágenes y deconvolución ciega, clasificación de imágenes multiespectrales y aprendizaje activo y otros problemas relacionados (adquisición de campos de luz y recuperación de vídeo). La tesis se presenta en la modalidad de “compendio” y a continuación citamos las contribuciones en cada uno de los bloques.

Bloque I: Restauración de Imágenes y Deconvolución Ciega

- H. Madero-Orozco, **P. Ruiz**, J. Mateos, R. Molina, y A.K. Katsaggelos, “Image Deblurring Combining Poisson Singular Integral and Total Variation Prior Models” en *21th European Signal Processing Conference (EUSIPCO 2013)*, 1569744251, Marrakech (Marruecos), Septiembre 2013.
- **P. Ruiz**, H. Madero-Orozco, J. Mateos, O.O. Vergara-Villegas, R. Molina, y A.K. Katsaggelos, “Combining Poisson Singular Integral and Total Variation Prior Models in Image Restoration”, *Signal Processing*, vol. 103, 296-308, Octubre 2014.
- **P. Ruiz**, X. Zhou, J. Mateos, R. Molina, y A.K. Katsaggelos, “Variational

bayesian Blind Image Deconvolution: A Review”, Digital Signal Processing, 2015. doi:10.1016/j.dsp.2015.04.012 (Aceptado para publicación. Disponible online desde 4 Mayo 2015)

Bloque II: Clasificación de Imágenes Multiespectrales y Aprendizaje Activo

- **P. Ruiz**, J.V. Talens, J. Mateos, R. Molina, y A.K. Katsaggelos, “Interactive Classification Oriented Superresolution of Multispectral Images” en *7th International Workshop Data Analysis in Astronomy (DAA2011)*, editado por Livio Scarsi and Vito Di Gesù, 77-85, Erice (Italy), Abril 2011.
- **P. Ruiz**, J. Mateos, R. Molina, y A.K. Katsaggelos, “Learning Filters in Gaussian Process Classification Problems” en *IEEE International Conference on Image Processing (ICIP 2014)*, 2913-2917, Paris (Francia), Octubre 2014.
- **P. Ruiz**, J. Mateos, R. Molina, y A.K. Katsaggelos, “A bayesian Active Learning Framework for a Two-Class Classification Problem” en *MUSCLE International Workshop on Computational Intelligence for Multimedia Understanding*, editado por Emanuele Salerno, A. Enis Çetin y Ovidio Salvetti, vol. LNCS-7252, 42-53, Pisa (Italia), 2012.
- **P. Ruiz**, J. Mateos, G. Camps-Valls, R. Molina, y A.K. Katsaggelos, “bayesian Active Remote Sensing Image Classification”, IEEE Transactions on Geoscience and Remote Sensing, vol. 52, no. 4, 2186-2196, Abril 2014.
- **P. Ruiz**, N. Pérez de la Blanca, R. Molina, y A.K. Katsaggelos, “bayesian Classification and Active Learning Using ℓ_p -Priors. Application to Image Segmentation” en *22th European Signal Processing Conference (EUSIPCO 2014)*, 1183-1187, Lisboa (Portugal), Septiembre 2014.

Bloque III: Otros Problemas Relacionados (Adquisición de Campos de Luz y Recuperación de Video)

En el problema de adquisición de campos de luz se publicaron las siguientes contribuciones:

- S.D. Babacan, R. Ansorge, M. Luessi, **P. Ruiz**, R. Molina, y A. K. Katsaggelos, “Compressive Light Field Sensing”, IEEE Transaction on Image Processing, vol. 21, no. 12, 4746-4757, Diciembre 2012.
- **P. Ruiz**, J. Mateos, C. Cárdenas, S. Nakajima, R. Molina, y A.K. Katsaggelos, “Light Field Acquisition from Blurred Observations Using a Programmable

Coded Aperture Camera” en *21th European Signal Processing Conference (EUSIPCO 2013)*, 1569743131, Marrakech (Marruecos), Septiembre 2013.

Para el problema de recuperación de video se presentaron las siguientes publicaciones:

- **P. Ruiz**, S.D. Babacan, L. Gao, Z. Li, R. Molina, y A.K. Katsaggelos, “Video Retrieval Using Sparse bayesian Reconstruction” en *IEEE International Conference on Multimedia and Expo (ICME2011)*, 1-6, Barcelona (España), Julio 2011.
- **P. Ruiz**, S.D. Babacan, R. Molina, y A.K. Katsaggelos, “Retrieval of Video Clips with Missing Frames using Sparse bayesian Reconstruction” en *7th International Symposium on Image and Signal Processing and Analysis (ISPA 2011)*, 443-448, Dubrovnik (Croacia), Septiembre 2011.

Conclusiones

En esta tesis doctoral hemos aplicado la modelización e inferencia bayesiana a problemas de recuperación y clasificación de imágenes. Hemos demostrado que los problemas de restauración de la imagen, deconvolución ciega imagen, clasificación de imágenes multiespectrales, *pansharpening*, aprendizaje activo, adquisición de campos de luz y recuperación de vídeo se pueden modelar dentro del marco bayesiano, y la inferencia bayesiana nos ha permitido estimar la solución de estos problemas. En algunos casos, se han utilizado estimadores puntuales para reducir los problemas de inferencia a problemas de optimización. En otros casos, la inferencia variacional nos ha permitido aproximar la distribución a posteriori y estimar los parámetros del modelo. En las secciones experimentales, los métodos propuestos han demostrado ser muy precisos y eficientes y, en casi todos los casos, han llegado a superar los métodos de última generación. A continuación detallamos las conclusiones específicas para cada bloque.

Bloque I: Restauración de Imágenes y Deconvolución Ciega

- Hemos presentado un nuevo método de restauración de imágenes que utiliza Modelización e Inferencia Bayesiana para combinar dos modelos a priori: el modelo de variación total (TV), que realiza las fronteras y suaviza las regiones planas, y el modelo de la integral singular de Poisson (PSI) que es capaz de conservar texturas. El producto final es un algoritmo de restauración que combina las ventajas de ambos métodos. Además se ha llevado a cabo un estudio sobre los modelos TV y PSI, y los parámetros que controlan su forma, y hemos concluido que ni el modelo TV ni el PSI por separado pueden conseguir restauraciones que a la vez recuperen textura y controlen el ruido. Finalmente

se ha llevado a cabo un conjunto de experimentos donde el método propuesto ha sido comparado con otros modelos clásicos y del estado del arte. Los experimentos llevados a cabo demuestran que, en imágenes con detalles y regiones planas, el modelo de restauración propuesto, que combina los modelos TV y PSI, obtienen los mejores resultados.

- También hemos realizado una revisión sobre los métodos bayesianos de deconvolución de imágenes que existen en la literatura (BID). Hay dos sucesos que marcan la historia reciente de BID: el creciente interés de la comunidad de visión por computador en resolver problemas de BID y el dominio de la inferencia bayesiana como herramienta para resolverlos. El uso de métodos VB en combinación con modelos de imagen como los basados en las representaciones de Super Gaussianas y Mezcla Escalada de Gaussianas ha conducido a herramientas muy generales y potentes que consiguen muy buenas restauraciones a partir de las imágenes emborronadas. También se han aportado ejemplos de restauraciones con métodos en el estado del arte y se han discutido problemas que marcarán el futuro cercano de las investigaciones en BID.

Bloque II: Clasificación de Imágenes Multiespectrales

- Hemos demostrado que las técnicas de *pansharpening* pueden ser usadas para mejorar el rendimiento de los métodos de clasificación en imágenes multiespectrales. Para ello se ha abordado el problema de modificar adaptativamente los parámetros de los métodos de pansharpening con el objetivo de mejorar las medidas de clasificación sobre una clase dada sin deteriorar el rendimiento sobre las otras clases. La validez de la técnica propuesta ha sido demostrada usando un imagen real de Quickbird.
- También hemos presentado un método que filtra y clasifica imágenes de forma conjunta. Usando la modelización bayesiana y la inferencia variacional hemos desarrollado un algoritmo iterativo que estima los parámetros del clasificador y un banco de filtros óptimo de forma conjunta. En la sección experimental demostramos que los filtros estimados ayudan a mejorar el rendimiento en clasificación. El método propuesto se comparó con otras aproximaciones de clasificación/filtrado, y los resultados experimentales demostraron que el método propuesto es más preciso y eficiente.
- Hemos presentado una aproximación bayesiana no-paramétrica basada en núcleos para clasificación de imágenes multiespectrales. A partir de la información proporcionada por el clasificador se han desarrollado técnicas de aprendizaje activo que mostraron un rendimiento comparable a técnicas de aprendizaje activo recientes, que utilizan máquinas de vectores soporte (SVM). Los tres métodos desarrollados fueron: máxima diferencia de entropías, mínima

distancia a la frontera de decisión y mínima distancia normalizada. La estimación de los parámetros se hace de forma automática. La aproximación propuesta fue probada en varios escenarios para resolver el problema de la monitorización urbana con imágenes multispectrales y datos de radar SAR. Se observó que, aunque el rendimiento en clasificación era muy similar a SVM, en aprendizaje activo los métodos propuestos consiguen una mejora importante.

- También hemos abordado el problema de clasificación imponiendo que los coeficientes adaptativos tengan mínima pseudo norma ℓ_p . Así los coeficientes adaptativos correspondientes a las características no relevantes se harán cero, lo que nos permite identificarlos. Se usó la inferencia variacional bayesiana para estimar todos los parámetros del modelo y además se probó la relación con distribuciones a priori gaussianas independientes. También se calculó la distribución predictiva de las clases, lo que nos permitió desarrollar dos nuevas técnicas de aprendizaje activo para este clasificador: máxima entropía y mínima probabilidad. En la sección experimental, los resultados demostraron que el uso de las distribuciones ℓ_p permiten al clasificador seleccionar las características discriminatorias y descartar la componentes que no son relevantes. La aproximación propuesta ha demostrado ser más precisa que los métodos SVM en problemas de clasificación y aprendizaje activo.

Bloque III: Otros Problemas Relacionados (Adquisición de Campos de Luz y Recuperación de Vídeo)

- En adquisición de campos de luz, hemos presentado un nuevo prototipo de cámara que usa apertura codificada para captar el campo de luz de una escena. Este prototipo fue desarrollado en colaboración con el Instituto Andaluz de Astrofísica (IAA). En [15] se colaboró para desarrollar un sistema que usa la teoría de muestreo compresivo para obtener el campo de luz tomando muchas menos observaciones que vistas del campo de luz. Además en [16], abordamos el problema de recuperar el campo de luz a partir de observaciones emborronadas, que aparecen debido a la profundidad de campo limitada de las cámaras. Hemos desarrollado un método para deconvolucionar el campo de luz y obtener imágenes nítidas a partir de observaciones emborronadas, que ha funcionado sobre imágenes sintéticas y reales.
- En recuperación de vídeo hemos desarrollado un sistema robusto y eficiente, basado en el uso de representaciones ralas, muestreo compresivo y modelización bayesiana del problema de recuperación de vídeo. Los resultados experimentales han demostrado que el método propuesto funciona mejor que los métodos existentes en el estado del arte. También hemos demostrado que el modelo

propuesto es muy efectivo y robusto a ruido y fotogramas perdidos, y no requiere métodos sofisticados de extracción de características. Además, el modelo propuesto tiene un menor coste computacional que algunos de los métodos de recuperación de vídeo en el estado del arte consiguiendo, al mismo tiempo, una muy alta precisión en la recuperación.

Summary and Conclusions

Introduction

A good part of the research and applications on image recovery and image classification deals with inverse problems, that is, moving from known events back to their most probable causes. Solutions to these problems have been derived using numerous approaches. Spatial and spectral image structures lead to high correlation among pixels, which can be explicitly exploited by probabilistic methods (see for instance [1, 2, 3, 4]).

Image recovery and classification problems like image restoration [5, 6, 7], blind image deconvolution [8, 9, 10], super-resolution [11, 12, 13], light field acquisition [14, 15, 16], pansharpening [17], multispectral image classification [18, 19, 20, 21], active learning [22, 23, 24], video retrieval [25, 26, 27], video compression [28], video surveillance [29], face recognition [30, 31, 32], image alignment [33] and medical imaging [34], among many others can be approached using Bayesian modeling and inference.

A fundamental principle of the Bayesian philosophy is to regard all parameters and unobservable variables of a given problem as unknown stochastic quantities, assigning probability distributions based on beliefs. For instance, in an image recovery problem, the original image(s), the observation noise, and even the function(s) defining the acquisition process can all be treated as samples of random variables, with corresponding prior Probability Density Functions (PDFs) that model the available knowledge on the nature of images and the imaging process.

Within Bayesian inference, Variational Bayesian (VB) methods have attracted a lot of interest in Bayesian statistics, machine learning and related areas. A major disadvantage of traditional methods like Maximum Likelihood or Maximum a Posteriori, is that they do not make use of the information provided by the posterior. Expectation Maximization requires the complete knowledge of some posterior probabilities which cannot, in many cases, be calculated. Simulation methods aim at obtaining the true posterior; however, they usually are very time consuming. Variational Bayesian methods [35, 36, 37, 1, 38, 39, 40] overcome these limitations by approximating the unknown posterior distributions with simpler, analytically tractable distributions, and therefore extend the applicability of Bayesian inference

to a much wider range of modeling options. For instance the use of more complex priors (which are very often needed in image processing problems) modelling the unknowns can be utilized with ease, resulting in improved estimation accuracy.

In this dissertation we explore the application of Bayesian modeling and inference to the following problems: image restoration, blind image deconvolution, multispectral image classification, active learning, light field acquisition and video retrieval. By doing so we prove the applicability of the Bayesian framework to solve a wide range of image processing and classification tasks.

We believe it is important to indicate that the dissertation contains very broad as well as very specific contributions. Its content includes a Bayesian review of the very interesting blind image deconvolution problem and also includes contributions on very specific problems. This wide range format makes it particularly useful to anyone wanting to learn about Bayesian modeling and inference.

Structure of the Ph.D. Thesis

The main goal of this dissertation is to study the application of Bayesian modeling and inference to image recovery and classification problems. Its contents have been grouped into three blocks: image restoration and blind deconvolution, multispectral image classification and other related problems (light field acquisition and video retrieval). The dissertation is presented in the modality of “compendium”, and the scientific contributions on each block are cited below.

Block I: Image Restoration and Blind Image Deconvolution

- H. Madero-Orozco, **P. Ruiz**, J. Mateos, R. Molina, and A.K. Katsaggelos, “Image Deblurring Combining Poisson Singular Integral and Total Variation Prior Models” in *21th European Signal Processing Conference (EUSIPCO 2013)*, 1569744251, Marrakech (Morocco), September 2013.
- **P. Ruiz**, H. Madero-Orozco, J. Mateos, O.O. Vergara-Villegas, R. Molina, and A.K. Katsaggelos, “Combining Poisson Singular Integral and Total Variation Prior Models in Image Restoration”, *Signal Processing*, vol. 103, 296-308, October 2014.
- **P. Ruiz**, X. Zhou, J. Mateos, R. Molina, and A.K. Katsaggelos, “Variational Bayesian Blind Image Deconvolution: A Review”, *Digital Signal Processing*, 2015. doi:10.1016/j.dsp.2015.04.012 (Accepted for publication. Available online since 4 may 2015)

Block II: Multispectral Image Classification and Active Learning

- **P. Ruiz**, J.V. Talens, J. Mateos, R. Molina, and A.K. Katsaggelos, “Interactive Classification Oriented Superresolution of Multispectral Images” in *7th International Workshop Data Analysis in Astronomy (DAA2011)*, edited by Livio Scarsi and Vito Di Gesù, 77-85, Erice (Italy), April 2011.
- **P. Ruiz**, J. Mateos, R. Molina, and A.K. Katsaggelos, “Learning Filters in Gaussian Process Classification Problems” in *IEEE International Conference on Image Processing (ICIP 2014)*, 2913-2917, Paris (France), October 2014.
- **P. Ruiz**, J. Mateos, R. Molina, and A.K. Katsaggelos, “A Bayesian Active Learning Framework for a Two-Class Classification Problem” in *MUSCLE International Workshop on Computational Intelligence for Multimedia Understanding*, edited by Emanuele Salerno, A. Enis Çetin and Ovidio Salvetti, vol. LNCS-7252, 42-53, Pisa (Italy), 2012.
- **P. Ruiz**, J. Mateos, G. Camps-Valls, R. Molina, and A.K. Katsaggelos, “Bayesian Active Remote Sensing Image Classification”, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 4, 2186-2196, April 2014.
- **P. Ruiz**, N. Pérez de la Blanca, R. Molina, and A.K. Katsaggelos, “Bayesian Classification and Active Learning Using ℓ_p -Priors. Application to Image Segmentation” in *22th European Signal Processing Conference (EUSIPCO 2014)*, 1183-1187, Lisbon (Portugal), September 2014.

Block III: Other Related Problems (Light Field Acquisition and Video Retrieval)

For the light field acquisition problem the following contributions were presented:

- S.D. Babacan, R. Ansorge, M. Luessi, **P. Ruiz**, R. Molina, and A. K. Katsaggelos, “Compressive Light Field Sensing”, *IEEE Transaction on Image Processing*, vol. 21, no. 12, 4746-4757, December 2012.
- **P. Ruiz**, J. Mateos, C. Cárdenas, S. Nakajima, R. Molina, and A.K. Katsaggelos, “Light Field Acquisition from Blurred Observations Using a Programmable Coded Aperture Camera” in *21th European Signal Processing Conference (EUSIPCO 2013)*, 1569743131, Marrakech (Morocco), September 2013.

For the video retrieval problem the following contributions were presented:

- **P. Ruiz**, S.D. Babacan, L. Gao, Z. Li, R. Molina, and A.K. Katsaggelos, “Video Retrieval Using Sparse Bayesian Reconstruction” in *IEEE International*

Conference on Multimedia and Expo (ICME2011), 1-6, Barcelona (Spain), July 2011.

- **P. Ruiz**, S.D. Babacan, R. Molina, and A.K. Katsaggelos, “Retrieval of Video Clips with Missing Frames using Sparse Bayesian Reconstruction” in *7th International Symposium on Image and Signal Processing and Analysis (ISPA 2011)*, 443-448, Dubrovnik (Croatia), September 2011.

Conclusions

In this Dissertation we have applied Bayesian Modeling and Inference to Image Recovery and Classification Problems. We have shown that the image restoration, blind image deconvolution, multispectral image classification, pansharpening, active learning, light field acquisition and video retrieval problems can be modeled using the Bayesian framework, and Bayesian inference has allowed us to find solutions to these problems. In some cases, point estimates have been utilized to reduce the inference problems to an optimization one. In other, variational inference has allowed us to approximate the posterior distribution, and estimate the model parameters. In the performed experiments, the proposed methods have been shown to be very accurate and efficient and, in almost all problems, they have outperformed the state-of-the-art methods.

Below we detail the specific conclusions for each block.

Block I: Image Restoration and Blind Deconvolution

- First, we have presented a novel image restoration method that uses the Bayesian paradigm to combine two prior models: the total variation (TV) model that preserves edge structure while imposing smoothness on the solution and controlling noise, and the Poisson singular integral (PSI) model which is capable of preserving textures but cannot differentiate between highly detailed textures and noise. The final product is a restoration algorithm that combines the advantages of the two models. A study of TV and PSI models and the parameters that control their shape has been carried out. The work concludes that neither the TV nor the PSI image models alone can successfully recover textures and control noise. A set of experiments has been carried out, where the proposed method has been compared against both classical and state-of-the-art methods. The experimental results supported that for images with a combination of detailed and smooth regions, the proposed restoration method, which combines TV and PSI prior models, provides the best restorations.
- For the BID problem we have written a review of the recent literature on

Bayesian blind image deconvolution (BID) methods. We have stated that two events have marked the recent history of BID: the predominance of variational Bayes (VB) inference as a tool to solve BID problems and the increasing interest of the computer vision community in solving BID problems. We have shown that VB inference in combination with recent image models like the ones based on Super Gaussian (SG) and scale mixture of Gaussians (SMG) representations have led to the use of very general and powerful tools to provide clear images from blurry observations. In the provided review emphasis has been paid on VB inference and the use of SG and SMG models with coverage of recent advances in sampling methods. We have also provided examples of current state of the art BID methods and have discussed problems that very likely will mark the near future of BID.

Block II: Multispectral Image Classification Problems

- We have shown that pansharpening techniques can be used to increase the performance of classification methods when applied to multispectral images. We have addressed the problem of adaptively modifying the parameter of a pansharpening method in order to improve the precision and recall figures of merit of a classifier on a given class without deteriorating its performance over the other classes. The validity of the proposed technique has been demonstrated using a real Quickbird image.
- We have also presented a new method to jointly filter and classify a signal or an image. Using Bayesian modeling and variational inference we have developed an iterative procedure to jointly estimate the classifier parameters, the filter bank and the model parameters. We have experimentally shown that the estimated filters improve the classifier performance. The proposed method has been compared with other classification/filtering approaches, and experimental results have shown that the new method is both more accurate and more efficient.
- We have presented a non-parametric Bayesian learning approach based on kernels for remote sensing image classification. The Bayesian methodology efficiently tackles purely supervised and active learning approaches, and shows competitive performance when compared to support vector machines (SVMs) and recent active learning (AL) approaches. An incremental learning approach based on three different approaches was presented: maximum differential of entropies, minimum distance to decision boundary, and minimum normalized distance. Automatic parameter estimation is performed by using the evidence Bayesian approach, the kernel trick, and the marginal distribution of the observations instead of the posterior distribution of the adaptive parameters. The proposed approach was tested on several scenes dealing with urban monitoring

problems using multispectral and SAR data. We observed that, while similar results are obtained by SVMs in supervised mode, an improvement in accuracy and convergence is observed for the active learning scenario. Interestingly our methods do not only provide point-wise class predictions but confidence intervals.

- We have also developed a multiclass classification system using a prior on the adaptive coefficients based on the ℓ_p pseudo-norm. The contribution of the adaptive coefficients corresponding to no relevant data will be zero, which allows us to identify the irrelevant coefficients. Variational inference has been used to estimate all model parameters and connections with independent Gaussian priors was established. The predictive distribution of the classes has been calculated. This distribution has been used to define two active learning methods, named Minimum Probability Criteria and Maximum Entropy Criteria. Experimental results have shown that the use of ℓ_p -priors allows the classifier to select discriminative features and discard non-relevance components. The proposed approach has shown higher accuracy than SVM methods in both classification and AL problems.

Block III: Other Related Problems (Light Field Acquisition and Video Retrieval)

- We have developed a new programmable aperture camera prototype to capture the light field. The prototype was constructed in collaboration with the Instituto de Astrofísica de Andalucía (IAA). In [15] we developed a system which uses the compressive sensing theory to capture the light field by taking much fewer observations than views of the light field. In [16], we addressed the problem of recovering blurred light fields. We developed a method to deconvolve blurred light fields and experimentally showed that it is possible to obtain sharp images from blurred observations using both synthetic and real images.
- We have developed a robust and efficient system for video retrieval, based on the use of sparse representation, compressive sensing and Bayesian modeling of the video retrieval problem. Experimental results demonstrate that the proposed method performs better than existing state-of-the-art systems and also it is robust against noise. We have also shown that the new system is very effective and robust to noise and missing frames, and does not require sophisticated and data-dependent feature extraction methods.

Chapter 1

Introduction

1.1 Introduction

A good part of the research and applications on image recovery and image classification deal with inverse problems, that is, moving from known events back to their most probable causes. Solutions to these problems have been originally derived using numerous approaches. Spatial and spectral image structures lead to a high correlation between pixels, which can be explicitly exploited by probabilistic methods (see for instance [1, 2, 3, 4]).

Image recovery and classification problems like image restoration [5, 6, 7], blind image deconvolution [8, 9, 10], super-resolution [11, 12, 13], light field acquisition [14, 15, 16], pansharpening [17], multispectral image classification [18, 19, 20, 21], active learning [22, 23, 24], video retrieval [25, 26, 27], video compression [28], video surveillance [29], face recognition [30, 31, 32], image alignment [33] and medical imaging [34], among many others can be approached by using Bayesian modeling and inference.

A fundamental principle of the Bayesian philosophy is to regard all parameters and unobservable variables of a given problem as unknown stochastic quantities, assigning probability distributions based on beliefs. For instance, in an image recovery problem, the original image(s), the observation noise, and even the function(s) defining the acquisition process can all be treated as samples of random variables, with corresponding prior Probability Density Functions (PDFs) that model the available knowledge on the nature of images and the imaging process.

Within Bayesian inference, Variational Bayesian (VB) methods have attracted a lot of interest in Bayesian statistics, machine learning and related areas. A major disadvantage of traditional methods like ML or MAP, is that they do not make use of the information provided by the posterior. EM requires the complete knowledge of some posterior probabilities which cannot, in many cases, be calculated. Simulation methods aim at obtaining the true posterior however they usually are very time consuming. Variational Bayesian methods [35, 36, 37, 1, 38, 39, 40] overcome

limitations by approximating the unknown posterior distributions with simpler, analytically tractable distributions, and therefore extend the applicability of Bayesian inference to a much wider range of modeling options. For instance, the use of more complex priors (which are very often needed in image processing problems) to model the unknowns can be utilized with ease, resulting in improved estimation accuracy.

In this dissertation we explore the application of Bayesian modeling and inference to the following problems: image restoration, blind image deconvolution, multispectral image classification, active learning, light field acquisition and video retrieval. By doing so we prove the wide applicability of the Bayesian framework to solve a wide range of image processing and classification problems.

We believe that it is important to indicate that the dissertation contains very broad as well as very specific contributions. Its content includes a Bayesian review of the very interesting blind image deconvolution problem and also includes contributions on very specific problems. This wide range format of the dissertation makes it particularly useful to anyone wanting to learn on Bayesian modeling and inference. Furthermore, it provides its author with the knowledge of an extremely powerful tool which can be applied to many interesting problems.

Let us now briefly describe the image processing and classification problems we will contribute to in the thesis.

1.1.1 Image Restoration and Blind Image Deconvolution

As stated in [10], thousands of millions of pictures are taken everyday. If the claim in [41] is right, 880 billion photos were taken in 2014. Every minute, 27,800 pictures are uploaded to Instagram, 208,300 photos are uploaded to Facebook and more than one thousand to Flickr, and the trend, with a digital camera in every mobile phone, is probably exponentially increasing. Those pictures are intended to be a detailed representation of reality, but very often the captured image is degraded by blur and noise. Blur can occur, for instance, by movement during the capturing process or because the scene is out of focus. Furthermore, noise can be introduced, for instance, by sensor imperfections, poor illumination or communication errors [42].

Image restoration, also referred to as image deconvolution, is a mature topic that aims at recovering the underlying original image from its blurred and noisy observations. Sometimes, the blur is completely or partially known or can be estimated prior to the deconvolution process. For instance, in astronomical imaging, an accurate representation of the blur can be obtained by imaging a single star first before photographing the astronomical object of interest. In contrast, blind image deconvolution (BID) tackles the restoration problem without knowing the blur in advance, leading to one of the most challenging image processing problems, since many combinations of blur and “true” image can produce the observed image. To start with, image restoration with a known blur is an ill posed problem in the Hadamard sense [43], that is, small variations in the data result in large variations in the solution.

The problem is exacerbated in the BID problem, since in addition, small variations in the estimated blur can lead to large variations in the restored image. BID is an underdetermined nonlinear inverse problem, which requires the estimation of many more unknown variables than the available observed data. To find meaningful solutions, not only prior information about the unknowns is crucial, but also a good and sound estimation approach. Variational Bayes inference has emerged as a dominant approach for the solution of restoration and blind image deconvolution problems. VB inference in combination with recently introduced image models has led to the development of very general and powerful tools to obtain clear images from blurry observations which will be explored in this dissertation.

1.1.2 Multispectral Image Classification and Active Learning

Remote sensing images are of great interest in numerous applications. Map drawing, delimitation of parcels, studies on hydrology, forest or agriculture are just a few examples where these images are used [44, 45, 46]. Many of these applications involve the classification of pixels in an image into a number of classes. In supervised classification, the user provides the label of a set of samples to train the classifiers. Usually, the larger the training set, the better the classification performance but more expensive (in time or money) the construction of such a set is.

Due to physical and technological constraints, satellite images need to be processed before the classifier is trained. Many real classification tasks take into account the sequentiality (or vicinity) of the pixels, by first filtering the data and then performing classification. For instance, in [20] multispectral images are filtered before training the classifier, using a filter specially designed to improve the separation between classes. In [47], the authors use pansharpening methods, an image fusion approach that combines low resolution multispectral and panchromatic images to obtain an image with the spectral resolution of the multispectral image and the spatial resolution of the panchromatic image, followed by a classification method on the improved multispectral image.

However, as proved in [48] if processing is carried out before training, the classifier performance may not be optimal. To deal with this problem, an optimal filterbank is estimated during the training phase of a maximum margin classifier in [48]. The idea of jointly optimizing a filter and a classifier dates back to the 1990s within the field of artificial neural networks. It was, for instance, used in convolutional networks [49] or to define a neural model for temporal processing [50, 51].

While extracting the training pixels is normally straightforward and inexpensive, labeling each one of those pixels is a tedious and often expensive task. Active learning is a supervised learning technique that attempts to overcome the labeling bottleneck by asking queries in the form of unlabeled samples to be labeled by an oracle (e.g., a human annotator) [22]. An active learning procedure queries only the most informative samples from the whole set of unlabeled samples. The objective is

to obtain a high classification performance using as few labeled samples as possible, minimizing, this way, the cost of obtaining labeled data. In the literature there are active learning methods that use Support Vector Machines (SVM) as classifier [22, 24, 52], as well as Bayesian approaches [23, 21]. A survey of active learning algorithms for supervised remote sensing image classification can be found in [53]. The selection of the best sample in SVM active learning approaches is based on the distance of the samples to the boundary decision. In the Bayesian approach, however, the selection is based on the samples posterior probability of belonging to each class, thus taken into account the prediction uncertainty and providing a significant advantage during active learning tasks. Bayesian simultaneous filtering and classification, and active learning techniques will be contributed to in this dissertation.

1.1.3 Light Field Acquisition

Moving from analog to digital has been a major advance in the world of photography. Besides the cost reduction, digital images can be edited and post-processed in countless ways by using a computer. In computational photography (CP), the post-processing does most of the work, considering the image captured by the sensor as an intermediate data [54]. The light field of a given scene contains different angular views of the same scene. It is used to reconstruct a 3D model of the scene [55], in re-focusing problems [14] or to synthesize interpolated views [56].

To capture light fields different techniques have been proposed. Plenoptic cameras, like Lytro [57] or Raytrix [58], introduce an array of microlenses in front of the sensor. This allows the sensor to record different angular views of the scene. Depending on the number of microlenses used, the resolution of the captured images can be greatly reduced. That is, there is a trade-off between angular resolution and spatial resolution of the light field; the more angular views are generated, the smaller the spatial resolution of each view. To deal with this problem, systems using a coded aperture have been designed. In coded aperture acquisition systems, a pattern mask is introduced to modify the lens aperture and to capture images that, once processed, allow the reconstruction of the light field. Coded aperture started to be used for light field acquisition only a few years ago. In [14], the N angular views are obtained from N scrambled images captured with different masks and then solving a determined system of linear equations. The masks are loaded into a programmable LCD that is placed into the lens. The design in [59] uses Liquid Crystal on Silicon (LCoS) to create the masks. This reduces the loss of light and improves the brightness and contrast but makes the lens bulkier than the LCD design. None of the proposed models has dealt with the problem of defocused light fields. In spite of the small size of the individual blocks composing the coded aperture, the depth of field is limited and objects outside it will appear defocused in the reconstructed views. In this dissertation we will explore the use of Bayesian CS techniques to capture the light field as well as methods to deblur the obtained images.

1.1.4 Video Retrieval

With the rapidly increasing growth of digital video content, there is an equally growing need for efficient techniques to analyze, search and retrieve video content. Individuals may want to search for video content they are interested in from YouTube videos, media companies may want to locate video content that violates their copyright protection (fingerprint) and, security systems may want to detect suspicious events among surveillance videos. Video retrieval is a key step in many applications including copyright protection, multimedia content search, security and surveillance. Fast and accurate algorithms in all the above, and many other, cases are needed for efficient video retrieval.

A number of methods have been developed for video retrieval. Generally, methods identify features distinguishing video frames and employ classification, indexing and searching based on these features. Surveys and comparisons for feature identification can be found in [60, 61]. After identifying the distinguishing features, the second step in video retrieval is searching based on these features. Indexing and hashing are commonly used to improve the search efficiency. In [62] geometric hashing is used to build database indices, while [63, 64, 65] used tree-based indexing. A powerful data structure for indexing is the kd-trees. In [66], video trajectories over time are indexed using kd-trees with a dimensionality reduction using PCA. Random projections instead of PCA are utilized in [67], followed by several kdtrees for indexing. In this dissertation we explore the use of Bayesian techniques for video retrieval.

1.2 Outline

This dissertation is organized as follows:

- Chapter 2, brief introduction to Bayesian modeling and inference and objectives of the thesis.
- Chapter 3, contributions on image restoration [68, 69] and blind image deconvolution [10].
- Chapters 4, contributions on image processing for classification [70, 71].
- Chapter 5, contributions on active learning [72, 73, 74].
- Chapter 6, contributions on light field acquisition [15, 16], and video retrieval [75, 76].
- Chapter 7, conclusions and future works.

Chapter 2

Problem Formulation and Objectives of the Thesis

In this chapter we provide an introduction to Bayesian modeling and inference and present the thesis contributions.

2.1 Bayesian Modeling

All the problems described in the previous chapter, can be tackled using Bayesian modeling and inference. This section contains a basic introduction to the Bayesian framework. A complete description of Bayesian modeling and inference, which is beyond the scope of this thesis, can be found in [1, 2, 3, 4, 35]

Let $\mathbf{y} = \{\mathbf{y}_i\}_{i=1}^N$ be a set of N observed variables, where each of them is a vector of size n . For instance, in an image restoration problem, \mathbf{y} would denote the noisy and blurred image, ordered using the lexicographic order as a column vector, in a classification problem, \mathbf{y} would be a vector of labels, and in a light field acquisition system, it would be the set of captured images. $\mathbf{z} = \{\mathbf{z}_i\}_{i=1}^M$ denotes the unknown variables which have led to the observations \mathbf{y} , each one of size m_i , $i = 1 \dots, M$. For example, in a classification problem, \mathbf{z} would be an adaptive coefficient vector, while in a blind deconvolution problem $\mathbf{z} = \{\mathbf{x}, \mathbf{h}\}$, where \mathbf{x} represents the original image and \mathbf{h} the blur kernel. Bayesian methods start with a prior distribution, a probability distribution over unknown \mathbf{z} , $p(\mathbf{z}|\Omega)$, where Ω is the set of parameters. In the prior distribution is where the expected structure of \mathbf{z} is incorporated. Usually, a prior $p(\Omega)$ on the model parameters is also incorporated. Sometimes the prior on the model parameters is called hyperprior and the elements of Ω are called hyperparameters. It is also necessary to specify $p(\mathbf{y}|\mathbf{z}, \Omega)$, the probability distribution of observed variables \mathbf{y} if \mathbf{z} , Ω were known. We then finally have

$$p(\mathbf{y}, \mathbf{z}, \Omega) = p(\mathbf{y}|\mathbf{z}, \Omega)p(\mathbf{z}|\Omega)p(\Omega). \quad (2.1)$$

The objective of Bayesian analysis is to infer the unknown \mathbf{z} , Ω given the observed \mathbf{y} . Before describing how inference is performed, let us describe each of component in the above equation, that is, the prior, observation and hyperparameter models.

2.1.1 Prior models

The Bayesian paradigm starts by modeling the previous knowledge on the unknown \mathbf{z} . For instance, in a classification problem we may want to estimate a set of adaptive parameters \mathbf{w} which are assumed a priori to be close to zero. This information is introduced in the problem by considering that \mathbf{w} is a realization of the Gaussian distribution

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}), \quad (2.2)$$

where α is the precision parameter. Note that α here is used to model how close to zero we expect the values to be.

In an image restoration problem, the difference between one pixel of the original image \mathbf{x} and its neighbors is expected to be small, this is modelled using the distribution

$$p(\mathbf{x}|\alpha) = \mathcal{N}(\mathbf{x}|\mathbf{0}, \alpha^{-1}\mathbf{C}^{-1}), \quad (2.3)$$

where \mathbf{C} is the Laplacian operator, and α is again a precision parameter.

These probability distributions are called *priors* since they encapsulate previous knowledge on the unknowns.

More complex priors can be used depending on the application and the knowledge we have on the problem solutions. Together with the classical prior models such as Conditional Autoregression (CAR) or Simultaneous Autoregression (SAR) used by Molina *et al.* [77] to impose smoothness, or Total Variation proposed by Rudin, Osher and Fatemi [78] to impose piecewise smoothness, the ℓ_p prior has been used in a large number of works such [79, 80, 81, 82, 83, 84]. In these papers the prior distribution is based on the use of quasi-norms $\|\cdot\|_p^p$ with $0 < p < 1$ as energy functions. Levin *et al.* [81] suggest the use of p in the range $[0.6, 0.8]$ for natural images. This prior model is not only used in image recovery, and can be found in problems like classification [85, 86], or compressive sensing [87].

The Super-Gaussianity property of probability distributions presented by Palmer in [88], was used in Babacan et al. [9] as the building block to propose a general representation for sparse priors. Interestingly, almost all previous and very recently proposed prior models can be represented using SG. This representation is used in the same work [9] to introduce two new image priors log and exp. Recent models like the one proposed by Zhang and Wifp [89], or the Student-t prior recently proposed by Mohammad-Djafari [90] are particular cases of SG distributions.

In general, previous knowledge on the unknowns is modeled using the probability distribution $p(\mathbf{z}|\Omega)$. However, the unknowns cannot be directly estimated from the priors, this task is carried out once the information provided by the observations is

incorporated.

2.1.2 Observation Model

The observation model describes how the observations \mathbf{y} are obtained from the hidden variables \mathbf{z} . In the Bayesian formulation it is represented using the conditional probability distribution of \mathbf{y} given \mathbf{z} , also known as likelihood,

$$p(\mathbf{y}|\mathbf{z}, \boldsymbol{\Omega}). \quad (2.4)$$

For instance, in image restoration it is usual to model the degradation process as

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \boldsymbol{\epsilon}, \quad (2.5)$$

where \mathbf{H} is a convolution matrix representing the blur, assumed to be known, and $\boldsymbol{\epsilon}$ is a noise vector. Assuming that $\boldsymbol{\epsilon}$ is a Gaussian noise with inverse of variance β we have

$$p(\mathbf{y}|\mathbf{x}, \beta) = \mathcal{N}(\mathbf{y}|\mathbf{H}\mathbf{x}, \beta^{-1}\mathbf{I}). \quad (2.6)$$

where for this case $\boldsymbol{\Omega} = \{\beta\}$.

In a two-class classification problem the relation between the observed labels \mathbf{y} and the hidden variables \mathbf{w} can be modeled using the distribution

$$p(\mathbf{y}|\mathbf{w}) = \prod_{i=1}^N \left(\frac{1}{1 + \exp(-\mathbf{w}^T \mathbf{x}_i)} \right)^{y_i} \left(\frac{\exp(-\mathbf{w}^T \mathbf{x}_i)}{1 + \exp(-\mathbf{w}^T \mathbf{x}_i)} \right)^{1-y_i}. \quad (2.7)$$

where $\{\mathbf{x}_i\}_{i=1}^N$ is the training set.

2.1.3 Hyperprior Models

Observation and prior models depend on a set of parameters $\boldsymbol{\Omega}$ that needs to be estimated along with the unknown variables \mathbf{z} . Bayesian modeling and inference provides different ways to estimate the system parameters from the input data. Thus we can develop algorithms capable of working automatically, which help non-expert users to obtain satisfactory problem solutions.

In a hierarchical Bayesian framework, parameter modeling is carried out in a second stage (the first stage models the degradation and prior on the data), where a probability distribution is assigned to each parameter. These distributions, allow to incorporate knowledge on the hyperparameters into the model.

A large part of the Bayesian literature is devoted to finding hyperprior distributions $p(\boldsymbol{\Omega})$ for which $p(\mathbf{z}, \boldsymbol{\Omega}|\mathbf{y})$ can be calculated in a straightforward way or at least be approximated. These are the so called conjugate priors [91], which were developed extensively in Raiffa and Schlaifer [92].

Besides providing for easy calculation or approximations of $p(\mathbf{z}, \boldsymbol{\Omega}|\mathbf{y})$, conjugate priors have the intuitive feature of allowing one to begin with a certain functional form for the prior and end up with a posterior of the same functional form, but with the parameters updated by the sample information.

The a priori models for the hyperparameters depend on the type of the unknown parameters, and different models proposed in the literature. For parameters corresponding to inverses of variances, the gamma distribution is normally used

$$p(\omega) = \Gamma(\omega|a, b), \quad (2.8)$$

which has mean and variance

$$\mathbb{E}[\omega] = \frac{a}{b}, \quad \text{Var}[\omega] = \frac{a}{b^2}, \quad (2.9)$$

where $\omega \in \boldsymbol{\Omega}$ and a and b are the hyperparameters which must be set in advance.

Other methods proposed in the literature use the uninformative prior model on a proper scale

$$p(\boldsymbol{\Omega}) = \text{constant}, \quad (2.10)$$

which is appropriate when no knowledge on the prior value of the parameters is available.

2.2 Bayesian Inference

Once the observation and prior models have been described, in other words, once the elements of the joint probability model in (2.1) have been specified, the goal now becomes the inference on the unknown variables $\Theta = \{\mathbf{z}, \boldsymbol{\Omega}\}$ given the observations.

In the Bayesian framework, Θ is inferred calculating (or approximating) the posterior distribution $p(\Theta|\mathbf{y})$, expressed using the Bayes' rule as

$$p(\Theta|\mathbf{y}) = \frac{p(\Theta, \mathbf{y})}{p(\mathbf{y})} = \frac{p(\mathbf{y}|\Theta)p(\mathbf{z}|\boldsymbol{\Omega})p(\boldsymbol{\Omega})}{p(\mathbf{y})}. \quad (2.11)$$

Unfortunately, since the integral $p(\mathbf{y}) = \int p(\Theta, \mathbf{y})d\Theta$ is not tractable in most real applications, the above posterior cannot be analytically calculated. Different estimation methods have been proposed to address this problem and we will now review them.

Probably the most widely used method in the literature is Maximum a Posteriori (MAP). Since $p(\Theta|\mathbf{y}) \propto p(\Theta, \mathbf{y})$ the maximum of the posterior distribution can be obtained by maximizing the joint distribution $p(\Theta, \mathbf{y})$ with respect to Θ . However, as pointed out in the landmark papers by Levin *et al.* [93, 94], MAP is not a suitable estimation procedure in many problems.

Table 2.1: Comparison of inference methods

	MAP	VB	MCMC
Has full posterior	no	yes	yes
Has point estimates	yes	yes	yes
Has uncertainty info	no	yes	yes
Allows hidden data	no	yes	yes
Complexity	low	medium	high

In contrast, variational Bayesian (VB) inference provides approaches for estimation of the posterior distributions of \mathbf{z} and $\boldsymbol{\Omega}$, and moreover generalizes MAP (see [35] for a proof).

Together with the well established use of VB inference, Markov Chain Monte Carlo (MCMC) methods are also popular. MCMC is the most general method used to approximate a posterior distribution, see [95, 96, 97] for details. The model in Eq. (2.1) is used to generate thousands of samples of $p(\mathbf{z}, \boldsymbol{\Omega} | \mathbf{y})$, which are used to infer the posterior distribution. In theory, sampling methods can find the exact form of the posterior distribution, but in practice they are computationally intensive (especially for multidimensional signals such as images) and their convergence is hard to establish.

In computationally cost terms, VB is much more efficient than MCMC, and more expensive than MAP. The features of each method are summarized in Table 2.1.

We now describe the application of VB to solve inverse problems.

2.2.1 Variational Inference

Variational Bayes inference is a powerful alternative to MAP, MCMC and many other inference methods. It provides more accurate approximations to the posterior distribution than point estimation methods, and it is computationally much more efficient than sampling approaches.

VB methods provide analytically tractable approximations $q(\boldsymbol{\Theta})$ to the true *posterior* $p(\boldsymbol{\Theta} | \mathbf{y})$ by assuming that $q(\boldsymbol{\Theta})$ has specific parametric or factorized forms. The distribution $q(\boldsymbol{\Theta})$ is found as the distribution that minimizes the Kullback-Leibler (KL) divergence

$$\text{KL}(q(\boldsymbol{\Theta}) || p(\boldsymbol{\Theta} | \mathbf{y})) = \int_{\boldsymbol{\Theta}} q(\boldsymbol{\Theta}) \log \frac{q(\boldsymbol{\Theta})}{p(\boldsymbol{\Theta} | \mathbf{y})} d\boldsymbol{\Theta} = \int_{\boldsymbol{\Theta}} q(\boldsymbol{\Theta}) \log \frac{q(\boldsymbol{\Theta})}{p(\mathbf{y}, \boldsymbol{\Theta})} d\boldsymbol{\Theta} + \text{constant}, \quad (2.12)$$

which is always non-negative and 0 only when $q(\boldsymbol{\Theta})$ and $p(\boldsymbol{\Theta} | \mathbf{y})$ coincide. (See [3] for a proof).

To minimize the Kullback-Leibler divergence it is necessary to assume some constraints on the distribution $q(\boldsymbol{\Theta})$. The first options is to consider that $q(\boldsymbol{\Theta})$

belongs to a parametric family of distributions. Another (very commonly used) assumption is to consider that $q(\Theta)$ factorizes as M disjoint groups, i.e.,

$$q(\Theta) = \prod_{i=1}^M q_i(\Theta_i), \quad (2.13)$$

where each factor q_i depends on a subset of unknown variables $\Theta_i \subseteq \Theta$. This method is known as Mean Field in Physics [98]. Thus, the KL divergence can be minimized with respect to each factor q_i separately, while the remaining factors are fixed.

The solution to this problem is given in [1]

$$q_i(\Theta_i) = Z_i \exp \left\{ \mathbb{E}_{\Theta \setminus \Theta_i} [\log p(\Theta, \mathbf{y})] \right\}, \quad (2.14)$$

where Z_i is a normalization constant.

Note that Eq. (2.14) defines a system of nonlinear equations in $\{\Theta_i\}_{i=1}^M$. One way to solve this system of equations is via an alternating optimization procedure, where the distribution of each factor is iteratively updated using the most recent distributions of all the other factors. This update process is cyclic and is repeated until convergence. Since the KL divergence Eq. (2.12) is convex with respect to $q_i(\Theta_i)$ [99], the convergence is guaranteed.

2.3 Objectives of the Ph.D. Thesis

Having described the tool we will use in this thesis to solve image filtering and classification problems we now state the thesis objectives.

The main goal of this Ph.D. Thesis is to find solutions to the problems described in section 1.1 using Bayesian modeling and inference. These problems can be grouped in three blocks: image restoration and blind deconvolution, multispectral image classification and other related problems (light field acquisition and video retrieval). We detail next the specific research objectives for each block.

2.3.1 Image Restoration and Blind Deconvolution

Image restoration and blind image deconvolution problems have been addressed using different approaches, being those based on Bayesian paradigm one of the most successful. In this area the thesis objectives are:

- To provide a comprehensive review of Bayesian modeling and inference techniques that have been used in BID problems.

- To examine the use of image priors which combine smoothness with texture preservation in image restoration problems.

2.3.2 Multispectral Image Classification and Active Learning

On this area of research the thesis objectives are:

- To use Bayesian modeling and inference to develop new classification techniques for multispectral images.
- To propose new image joint image filtering and classification techniques aimed at improving classifier performance.
- To develop new Bayesian active learning techniques outperforming the existing ones.

2.3.3 Other Related problems (Light Field Acquisition and Video Retrieval)

For the light field acquisition problem, the thesis objectives are:

- To create a new coded aperture camera prototype implementing CS concept.
- To develop a new image acquisition algorithm for the new prototype.
- To develop techniques to recover the light field from a set of blurred multiplexed observations captured with the coded aperture camera.

In video retrieval the thesis objectives are:

- To develop new techniques of video retrieval which use Bayesian modeling and inference. Those techniques should to retrieve sequences even if a large number of frames is missing in the query clip.

Chapter 3

Image Restoration and Blind Deconvolution

3.1 Image Restoration

3.1.1 Image Deblurring Combining Poisson Singular Integral and Total Variation Prior Models

- H. Madero-Orozco, **P. Ruiz**, J. Mateos, R. Molina, and A.K. Katsaggelos, “Image deblurring combining Poisson Singular Integral and Total Variation prior models” in *21th European Signal Processing Conference (EUSIPCO 2013)*, 1569744251, Marrakech (Morocco), September 2013.

- Status: Published
- Indexed in CORE Conference Ranking as CORE B ¹
- H index: 9 (Q3: 755/1201) ²

¹2014 CORE Conference Ranking <http://www.core.edu.au/index.php/conference-rankings>

²H index obtained from A. Martín-Martín, E. Orduña-Malea, J. M. Ayllón, E. Delgado López-Cózar, “Proceedings Scholar Metrics: H Index of proceedings on Computer Science, Electrical & Electronic Engineering, and Communications according to Google Scholar Metrics (2009-2013)”, <http://arxiv-web3.library.cornell.edu/abs/1412.7633>

IMAGE DEBLURRING COMBINING POISSON SINGULAR INTEGRAL AND TOTAL VARIATION PRIOR MODELS

Hiram Madero Orozco¹, Pablo Ruiz², Javier Mateos², Rafael Molina², Aggelos K. Katsaggelos³

¹Departamento de Ingeniería
Eléctrica y Computación
Universidad Autónoma de Ciudad Juárez,
Chihuahua, México
all16148@alumnos.uacj.mx

²Departamento de Ciencias
de la Computación e I.A.
Universidad de Granada,
Granada, Spain
{mataran, jmd, rms}@decsai.ugr.es

³Electrical Engineering and Computer
Science Department
Northwestern University
Evanston, IL, USA
aggk@eecs.northwestern.edu

ABSTRACT

In this paper a new combination of image priors is introduced and applied to Bayesian image restoration. Total Variation (TV) image prior preserves edge structure while imposing smoothness on the solutions. However, it does not perform well in textured areas. To alleviate this problem we propose to combine TV with the Poisson Singular Integral (PSI) image prior, which is able to preserve image textures. The proposed method utilizes a bound for the TV image model based on the majorization-minimization principle, and performs maximum a posteriori Bayesian inference. In the experimental section the proposed approach is tested on synthetically degraded images with different levels of spatial activity and areas with different types of texture. Since the proposed method depends on a set of parameters, an analysis, about their impact on the final restorations, is carried out.

Index Terms— Deblurring, Bayesian image restoration, Total Variation, Poisson Singular Integral

1. INTRODUCTION

When we take a picture, we want a detailed representation of the scene, but very often the observed image is degraded. The degradation is usually caused by movement during the recording process or because the scene is out of focus. Image deconvolution is an important task in image processing. Its goal is to recover or estimate the original image \mathbf{x} from a blurred and noisy observation \mathbf{y} . The image degradation model is a convolution between the original image and the known blurring operator \mathbf{H} . It can be expressed as

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}, \quad (1)$$

where \mathbf{n} is Gaussian additive white noise with zero mean and precision β .

Nowadays, different approaches try to solve this inverse problem. Methods based on Wavelets and Curvelets, capture and preserve sharp features in the image, and combined with threshold or shrinkage rules provide good results [1, 2].

Models based on the Bayesian paradigm provide solution to problems like blind deconvolution [3], space-variant deblurring [4], camera shake [5] and light field sensing [6]. Many of the proposed methods utilize a Total Variation (TV) image prior [7].

Total Variation preserves object boundaries (edges) but often eliminates image texture, because TV restricts the space of solutions to the space $BV(R^2)$ of functions of bounded variation; however, most natural images do not exactly belong to this space [8]. The texture in an image plays an important role in visual quality and it is not well modeled in such a space.

Carasso in [8] formulates the image restoration problem in Lipschitz spaces where a broader class of images can be accommodated. He proposes a new approach to recover the texture in images. The central idea is the implementation of the Poisson Singular Integral (PSI), which recovers the texture where the TV fails. PSI is also utilized in [9], where its authors propose a model which combines PSI and curvelet-type decomposition space semi-norm as regularizer.

The work presented by Chen *et al.* [10] proposes the use of texture-preserving image deblurring method. The authors adopt a two-step non-iterative processing procedure which first uses regularization in the frequency domain to remove the noise, and then utilizes a modified non-local means filter to reduce the leaked colored noise in order to obtain a good texture-preserving deblurred image.

In this paper, we propose a novel algorithm for image deconvolution, using a prior model combination (TV and PSI) in order to impose different properties on the restored image. The method produces restorations with edges and textures preserved, high PSNR and good visual quality. The paper is organized as follows. In section 2, the Bayesian modeling of the problem is presented. Section 3 discusses the inference procedure and proposes an algorithm to restore the images.

¹Special thanks to CONACYT.

²This research was supported by the Spanish Ministry of Economy and Competitiveness under project TIN2010-15137, the European Regional Development Fund (FEDER) and, in part by the US Department of Energy grant DE-NA0000457.

Section 4 contains the experimental section and, finally, section 5 concludes the paper.

2. BAYESIAN MODELING

The Bayesian paradigm is one of the most popular tool in image restoration (see [11] and references therein). The observation \mathbf{y} and the original image \mathbf{x} are treated as stochastic variables, and an inference process using Bayes' rule allows to obtain the restored image.

2.1. Observation Model

The degradation model in Eq. (1) provides the conditional probability distribution:

$$p(\mathbf{y}|\mathbf{x}, \beta) \propto \exp\left(-\frac{\beta}{2}\|\mathbf{y} - \mathbf{H}\mathbf{x}\|^2\right). \quad (2)$$

2.2. Image Model

In this paper we use a prior model combination, in order to ensure different properties of the restored image are presented [12]. The TV [11] prior has the advantage of preserving the edge structure while imposing smoothness on the solution. It is defined as

$$p_1(\mathbf{x}|\alpha_1) \propto \exp(-\alpha_1 \text{TV}(\mathbf{x})), \quad (3)$$

where $\text{TV}(\mathbf{x}) = \sum_{i=1}^P \sqrt{\Delta_i^h(\mathbf{x})^2 + \Delta_i^v(\mathbf{x})^2}$ with the operators $\Delta_i^h(\mathbf{x})$ and $\Delta_i^v(\mathbf{x})$ corresponding to the horizontal and vertical first order differences at pixel i , respectively, and P is the image size. However, this model does not work well in textured areas. To alleviate this problem, we combine TV with the Poisson Singular Integral (PSI) [8] filter which preserves textures. The PSI filter is defined in the Fourier domain for each $t > 0$ as

$$\mathbf{z}(\xi, \nu, t) = \left(t + \frac{4e^{-t\rho} - e^{-2t\rho} - 3}{2\rho}\right)^{1/2}, \quad (4)$$

where ξ, ν are the coordinates in Fourier domain and $\rho = \sqrt{\xi^2 + \nu^2}$. We denote by \mathbf{Z} the convolution matrix associated to filter \mathbf{z} in the spatial domain, and then define the second prior model as

$$p_2(\mathbf{x}|\alpha_2) \propto \exp\left(-\frac{\alpha_2}{2}\|\mathbf{Z}\mathbf{x}\|^2\right). \quad (5)$$

Figure 1 shows a set of realizations of the PSI prior model with variance 1, for different t values. As it can be observed t controls the smoothness of the texture. As t changes so does the texture granularity (notice the log scale).

Combining both models in Eq. (3) and (5), the prior distribution is given by

$$p(\mathbf{x}|\alpha_1, \alpha_2) \propto \exp\left(-\alpha_1 \text{TV}(\mathbf{x}) - \frac{\alpha_2}{2}\|\mathbf{Z}\mathbf{x}\|^2\right). \quad (6)$$

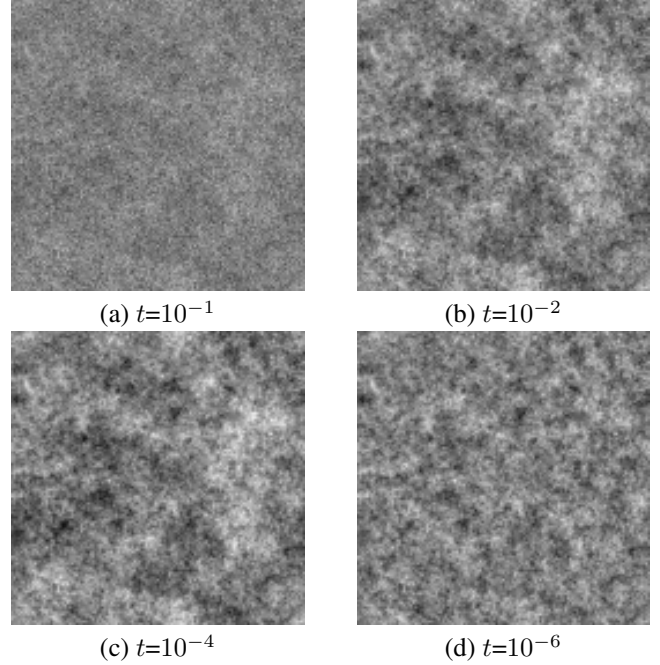


Fig. 1. Realizations of the prior model in Eq. (5) for different values of t .

3. BAYESIAN INFERENCE

The restored image sought after is the Maximum a Posteriori (MAP)

$$\begin{aligned} \hat{\mathbf{x}} &= \arg \max_{\mathbf{x}} p(\mathbf{x}|\mathbf{y}, \beta, \alpha_1, \alpha_2) \\ &= \arg \max_{\mathbf{x}} p(\mathbf{y}|\mathbf{x}, \beta)p(\mathbf{x}|\alpha_1, \alpha_2), \end{aligned} \quad (7)$$

which is obtained by minimizing

$$\mathcal{L}(\mathbf{x}) = \frac{\beta}{2}\|\mathbf{y} - \mathbf{H}\mathbf{x}\|^2 + \alpha_1 \text{TV}(\mathbf{x}) + \frac{\alpha_2}{2}\|\mathbf{Z}\mathbf{x}\|^2. \quad (8)$$

Due to use of the TV image prior, we need to utilize a majorization-minimization procedure [13]. Based on the average inequality [11], we have

$$\text{TV}(\mathbf{x}) \leq \frac{1}{2} \sum_{i=1}^P \frac{\Delta_i^h(\mathbf{x})^2 + \Delta_i^v(\mathbf{x})^2 + u_i}{\sqrt{u_i}} = \frac{1}{2}\mathbf{M}(\mathbf{x}, \mathbf{u}). \quad (9)$$

We then minimize

$$\bar{\mathcal{L}}(\mathbf{x}) = \frac{\beta}{2}\|\mathbf{y} - \mathbf{H}\mathbf{x}\|^2 + \frac{\alpha_1}{2}\mathbf{M}(\mathbf{x}, \mathbf{u}) + \frac{\alpha_2}{2}\|\mathbf{Z}\mathbf{x}\|^2. \quad (10)$$

This procedure introduces an additional parameter set $\mathbf{u} = (u_1, u_2, \dots, u_P)$, calculated as (see [11] for details)

$$u_i = \Delta_i^h(\mathbf{x})^2 + \Delta_i^v(\mathbf{x})^2. \quad (11)$$

Then the MAP estimator, $\hat{\mathbf{x}}$, is obtained as the solution of the linear equation system

$$\mathbf{A}\mathbf{x} = \beta\mathbf{H}^T\mathbf{y}, \quad (12)$$

where

$$\mathbf{A} = \beta \mathbf{H}^T \mathbf{H} + \alpha_1 ((\Delta^h)^T \mathbf{W} \Delta^h + (\Delta^v)^T \mathbf{W} \Delta^v) + \alpha_2 \mathbf{Z}^T \mathbf{Z}, \quad (13)$$

and Δ^h and Δ^v are the convolution matrices associated with horizontal and vertical gradients, respectively, and $\mathbf{W} = \text{diag}(\frac{1}{\sqrt{u_i}})$. We solve this system utilizing a conjugate gradient method. Since the estimation of \mathbf{x} and \mathbf{u} are coupled, we have the following iterative algorithm that alternatively estimates \mathbf{x} and \mathbf{u} until convergence.

Algorithm 1 Proposed Restoration Algorithm

Require: An initial estimate of the original image, \mathbf{x}^0

Set $k = 0$

repeat

1. Set $u_i^k = \Delta_i^h(\mathbf{x}^k)^2 + \Delta_i^v(\mathbf{x}^k)^2$ for $i = 1, \dots, P$.
2. Compute \mathbf{A}^k using the $\{u_i^k\}_{i=1, \dots, P}$ in Eq. (13).
3. Set \mathbf{x}^{k+1} as the solution of $\mathbf{A}^k \mathbf{x} = \beta \mathbf{H}^T \mathbf{y}$.
4. Set $k = k + 1$.

until $\|\mathbf{x}^k - \mathbf{x}^{k-1}\|^2 / \|\mathbf{x}^{k-1}\|^2 < \text{tol}$

4. EXPERIMENTS AND RESULTS

We tested the proposed algorithm on three different images, *cameraman*, *barbara*, and *baboon*. We chose these test images because they have different levels of spatial activity and areas with different types of texture. The images were synthetically degraded following the observation model in Eq. (2) by normalization to $[0, 1]$ interval, blurring each original image with a Gaussian blur with support 21×21 and standard deviation 1.5. Zero mean Gaussian noise with variance $\sigma_1^2 = 0.0001$ and $\sigma_2^2 = 0.001$ was added to blurred images to obtain two set of degraded images with a PSNR of about 34 dB and 24 dB, respectively.

To obtain the restored images, we run Algorithm 1 starting from the degraded image as initial estimate of the original image, that is, $\mathbf{x}^0 = \mathbf{y}$ and using $\text{tol} = 10^{-4}$ in the stopping criterion. The proposed method depends on a set of parameters, which need to be set to obtain the best performance. The PSI prior in Eq. (5) depends on the parameter t that controls the texture preservation. We run experiments to test the influence of this parameter on the restored images. We changed the parameter t in the range $-6 \leq \log t \leq -1$, following [8], and found that the difference on PSNR obtained with different values for the parameter t was low. This was a surprising result since the value of t conditions the shape of the prior model and it was supposed to preserve different textures on the image. Using a single value of t for the whole image is very likely not optimal and changing it locally will better adapt the algorithm to the different textures of the image. In this paper, however, we fixed $t = 0.1$ as suggested in [8].

We searched a set of values for the parameters that control the prior and degradation models as follows. First, notice that

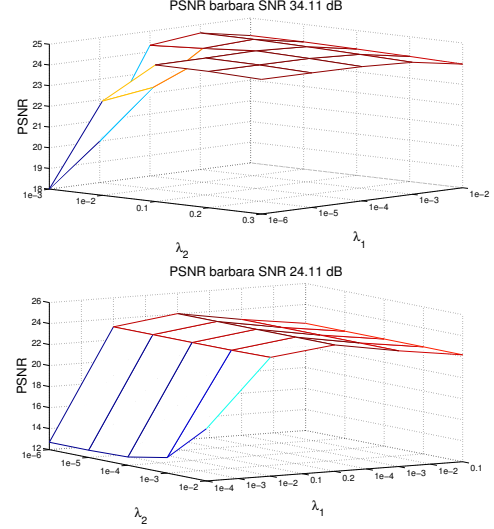


Fig. 2. PSNR evolution with different values of λ_1 and λ_2 for the *barbara* image. (a) Degradation with a SNR of 34 dB, (b) Degradation with a SNR of 24 dB.

Eq. (10) can be written as

$$\bar{\mathcal{L}}(\mathbf{x}) = \lambda \|\mathbf{y} - \mathbf{H}\mathbf{x}\|^2 + \lambda_1 \mathbf{M}(\mathbf{x}, \mathbf{u}) + \lambda_2 \|\mathbf{Z}\mathbf{x}\|^2, \quad (14)$$

with $\lambda = (1 - \lambda_1 - \lambda_2)$,

$$\lambda_1 = \frac{\alpha_1}{\beta + \alpha_1 + \alpha_2} \quad \text{and} \quad \lambda_2 = \frac{\alpha_2}{\beta + \alpha_1 + \alpha_2}. \quad (15)$$

In these equations, λ , λ_1 and λ_2 take values in the interval $[0, 1)$ and satisfy $\lambda + \lambda_1 + \lambda_2 = 1$. Thus, λ , λ_1 and λ_2 represent the influence on the restored image of the observed data, the TV, and the PSI models, respectively. Notice that selecting λ_1 and λ_2 in Eq. (14) is easier and more intuitive than selecting β , α_1 and α_2 in Eq. (10). We performed a search on this range by moving λ_1 and λ_2 in the set of values $[0, 10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 0.1, 0.2, 0.3, 0.4]$, and as we indicated above setting λ to $1 - \lambda_1 - \lambda_2$. We note that for values of λ_1 or λ_2 larger than 0.4 the quality of the restored image reduces drastically so we did not consider them in our experiments. In Fig. 2 we present the evolution of the PSNR as a function of λ_1 and λ_2 for the *barbara* image for the two noise degradations considered. In both cases, the shape of the curve is similar. We note that the value of the PSNR is quite similar around this maximum, which means that the method is not very sensitive to different values of the parameters λ_1 and λ_2 . This behavior was also observed on the rest of the test images so it confirms that it is not needed to select the parameters with a high precision to obtain good restorations.

However, we found significant differences on the values of the parameters that achieve the maximum PSNR for the different images. The values of the parameters as well as the

Table 1. Numerical results for the test images.

Cameraman								
SNR=24.44 dB				SNR=34.44 dB				
PSNR		Param		PSNR		Param		
Obs	Rest	λ_1	λ_2	Obs	Rest	λ_1	λ_2	
22.81	24.93	10^{-2}	10^{-7}	23.62	26.61	10^{-3}	10^{-7}	
	24.93	10^{-2}	0		26.61	10^{-3}	0	
	24.20	0	10^{-1}		25.90	0	10^{-2}	
Barbara								
SNR=24.11 dB				SNR=34.11 dB				
PSNR		Param		PSNR		Param		
Obs	Rest	λ_1	λ_2	Obs	Rest	λ_1	λ_2	
23.01	24.03	10^{-2}	10^{-4}	23.88	24.65	10^{-4}	10^{-1}	
	24.03	10^{-2}	0		24.59	10^{-3}	0	
	23.88	0	10^{-1}		24.60	0	10^{-2}	
Baboon								
SNR=24.56 dB				SNR=34.56 dB				
PSNR		Param		PSNR		Param		
Obs	Rest	λ_1	λ_2	Obs	Rest	λ_1	λ_2	
21.25	22.27	10^{-3}	0.2	21.80	23.33	10^{-7}	10^{-1}	
	21.88	10^{-3}	0		23.11	10^{-4}	0	
	22.02	0	10^{-1}		23.33	0	10^{-2}	

value of the PSNR for the observed and restored images are summarized in Table 1. We can extract some conclusions from those values. First, as the noise increases, higher reliance on prior information is needed and, hence, the values of the parameters λ_1 and λ_2 increase. Second, the relation of the importance of the PSI and TV models highly depend on the contents of the image. So, if the image presents a low level of detail, as it happens in the *cameraman* image, the restoration method prefers smooth restorations and the maximum value for the PSNR is obtained when λ_2 is equal to zero, giving control of the smoothness of the solution to the TV prior model. However, if the image contains a very high level of detail, as is the case with the *baboon* image, better results are obtained if the TV prior influence is almost neglected by setting the value of λ_1 very close to zero and leaving the control of the noise and texture preservation to the PSI prior model. This is expected since the TV prior tends to smooth out the small details in the image. Note however that, as the noise increases, including a small contribution by the TV prior provides better results since the PSI prior cannot differentiate between highly detailed textures and noise [9]. In images with a combination of detailed and smooth regions, a combination of both prior models provides the best result. This is the case with the *barbara* image that reaches its maximum PSNR when λ_1 and λ_2 are both greater than zero.

In Fig. 3 we present the original, observed and restored images for different noise degradations and different values for the parameters λ_1 and λ_2 . Although all restored images



(a) Original image.



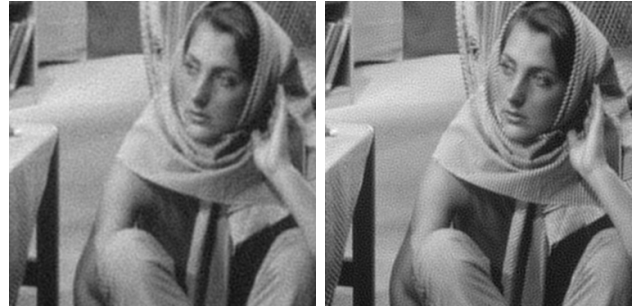
(b) Observation with a SNR of 24 dB.

(c) Observation with a SNR of 34 dB.



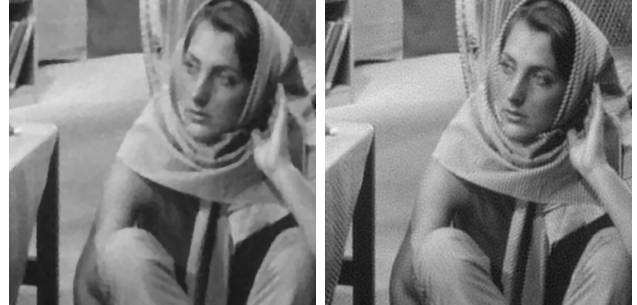
(d) Restoration of (b): only TV model.

(e) Restoration of (c): only TV model.



(f) Restoration of (b): only PSI model.

(g) Restoration of (c): only PSI model.



(h) Restoration of (b): both models.

(i) Restoration of (c): both models.

Fig. 3. Experimental results for the *Barbara* image.

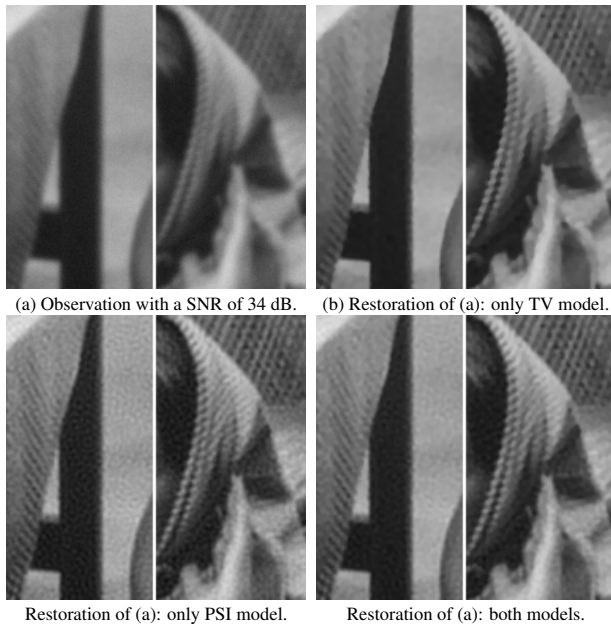


Fig. 4. Details of the restorations in Fig. 3.

present a high quality, the images obtained using a combination of the TV and the PSI models (Figs. 3h and 3i) show a higher visual quality and better preserve textures in areas as the handkerchief and the tablecloth while controlling noise, as can be seen in the details in Fig. 4. The images obtained using only the TV prior, that is, using $\lambda_2 = 0$ (Figs. 3d and 3e) look flat and most of the texture has been lost while the images using only the PSI prior ($\lambda_1 = 0$), depicted in Figs. 3f and 3g, are noisy. This agrees with the numerical results in Table 1. Notice that when the noise is higher, more contribution of the TV prior was needed in order to eliminate noise and, thus, texture in the restored image, as the handkerchief and the trousers, could not be successfully recovered.

5. CONCLUSIONS

In this paper we present a novel methodology to restore blurred images with noise. The combination of TV and PSI prior models provides better visual quality and PSNR than utilizing both models alone. The model recovers fine-scale details (texture) in cases where TV completely fails and our experimental results confirm this. The proposed method shows good performance with images with a combination of detailed and smooth regions, and textured images with high noise where the combination of TV and PSI controls the noise while preserving the details.

6. REFERENCES

- [1] J.-L. Stark, F. Murtagh, and J. M. Fadili, *Sparse image and signal processing*, Cambridge University, 2010.

- [2] E. Shaked and O. V. Michailovich, “Deconvolution of Poissonian images via iterative shrinkage,” *IEEE Int. Symp. on Biomedical Imaging*, pp. 1309–1312, 2010.
- [3] S.D. Babacan, R. Molina, M.N. Do, and A.K. Katsaggelos, “Blind deconvolution with general sparse image priors,” in *European Conference on Computer Vision (ECCV)*, September 2012, pp. 341–355.
- [4] M. Tallón, J. Mateos, S. D. Babacan, R. Molina, and A. K. Katsaggelos, “Space-variant kernel deconvolution and denoising in dual exposure problem,” *Information Fusion*, 2012.
- [5] M. Hirsch, C. J. Schuler, S. Harmeling, and B. Schölkopf, “Fast removal of non-uniform camera shake,” *ICCV*, 2011.
- [6] S. D. Babacan, R. Ansorge, M. Luessi, P. Ruiz, R. Molina, and A. K. Katsaggelos, “Compressive light field sensing,” *IEEE Trans. on Image Processing*, vol. 21, no. 12, pp. 4746–4757, December 2012.
- [7] L. I. Rudin, S. Osher, and E. Fatemi, “Nonlinear total variation based noise removal algorithms,” *Physica D*, pp. 259–268, 1992.
- [8] A. S. Carasso, “Singular integrals, image smoothness, and the recovery of texture in image deblurring,” *SIAM Journal on Applied Mathematics*, vol. 64, no. 5, pp. 1749–1774, June-July 2004.
- [9] L. Huang, L. Xiao, Z. Wei, and Z. Zhang, “Variational image restoration based on Poisson singular integral and curvelet-type decomposition space regularization,” in *IEEE International Conference on Image Processing*, 2011, vol. 18, pp. 685–688.
- [10] F. Chen, X. Huang, and W. Chen, “Texture-preserving image deblurring,” *IEEE Signal processing letters*, vol. 17, no. 12, pp. 1018–1021, December 2010.
- [11] S. D. Babacan, R. Molina, and A. K. Katsaggelos, “Variational Bayesian blind deconvolution using a total variation prior,” *IEEE Transactions on Image Processing*, vol. 18, no. 1, pp. 12–26, January 2009.
- [12] M. Vega, J. Mateos, R. Molina, and A. Katsaggelos, “Astronomical image restoration using variational methods and model combination,” *Statistical Methodology*, vol. 9, no. 1-2, pp. 19–31, January 2012.
- [13] S.D. Babacan, R. Molina, and A.K. Katsaggelos, “Parameter estimation in TV image restoration using variational distribution approximation,” *IEEE Transactions on image processing*, vol. 17, no. 3, pp. 326–339, 2008.

3.1.2 Combining Poisson Singular Integral and Total Variation prior models in Image Restoration

- **P. Ruiz**, H. Madero-Orozco, J. Mateos, O.O. Vergara-Villegas, R. Molina, and A.K. Katsaggelos, “Combining Poisson Singular Integral and Total Variation prior models in Image Restoration”, *Signal processing*, vol. 103, 296-308, October 2014.

- Status: Published
- Impact Factor (JCR 2013): 2.238
- Subject Category: Engineering, Electrical & Electronic (Q1: 51/248)

Combining Poisson Singular Integral and Total Variation prior models in Image Restoration

Hiram Madero-Orozco^a, Pablo Ruiz^b, Javier Mateos^b, Osslan Osiris Vergara-Villegas^c, Rafael Molina^b, Aggelos K. Katsaggelos^d

^a*Departamento de Ingeniería Eléctrica y Computación, Universidad Autónoma de Ciudad Juárez, Chihuahua, México*

^b*Departamento de Ciencias de la Computación e I.A., Universidad de Granada, Spain*

^c*Departamento de Ingeniería Industrial y Manufactura, Universidad Autónoma de Ciudad Juárez, Chihuahua, México*

^d*Electrical Engineering and Computer Science Department, Northwestern University, Evanston, IL, USA*

Abstract

In this paper, a novel Bayesian image restoration method based on a combination of priors is presented. It is well known that the Total Variation (TV) image prior preserves edge structures while imposing smoothness on the solutions. However, it tends to oversmooth textured areas. To alleviate this problem we propose to combine the TV and the Poisson Singular Integral (PSI) models, which, as we will show, preserves the image textures. The PSI prior depends on a parameter that controls the shape of the filter. A study on the behavior of the filter as a function of this parameter is presented. Our restoration model utilizes a bound for the TV image model based on

*This research was supported by CONACYT, by the Spanish Ministry of Economy and Competitiveness under project TIN2010-15137, the European Regional Development Fund (FEDER) and, in part by the US Department of Energy grant DE-NA0000457.

Email addresses: a1116148@alumnos.uacj.mx (Hiram Madero-Orozco), mataran@decsai.ugr.es (Pablo Ruiz), jmd@decsai.ugr.es (Javier Mateos), overgara@uacj.mx (Osslan Osiris Vergara-Villegas), rms@decsai.ugr.es (Rafael Molina), aggk@eecs.northwestern.edu (Aggelos K. Katsaggelos)

the majorization-minimization principle, and performs maximum *a posteriori* Bayesian inference. In order to assess the performance of the proposed approach, in the experimental section we compare it with other restoration methods.

Keywords: Deblurring, Denoising, Bayesian image restoration, Total Variation, Poisson Singular Integral.

1. Introduction

In the digital age we live in, millions of pictures are taken every day with digital cameras or mobile devices like cell-phones, tablets, etc. Those pictures are intended to be a detailed representation of the reality, but very often the captured image is degraded by blur and noise. Blur can occur, for instance, by movement during the capturing process or because the scene is out of focus. Furthermore, noise can be introduced, for instance, by sensor imperfections, poor illumination or communication errors [1]. When such problems occur, the usual solution is to take another picture, but sometimes it is not possible to retake the same picture and the moment is lost. Image restoration can help in those situations by estimating the original image from its blurred and noisy observation.

The image restoration problem has been addressed successfully using different approaches (see [2] for a detailed review of classical models and [1] for references of recent restoration models). When only noise is present, that is, the image is crisp but noisy, denoising algorithms such as [3, 4, 5] can be used. However, if the image is also blurred, image restoration methods, that handles both blurring and noise, are needed. Many restoration methods

utilize a Total Variation (TV) image prior or regularizer [6, 7, 8, 9]. TV is well known for preserving object boundaries (edges) and removing noise, but it often eliminates image texture, which plays an important role in visual quality.

Different methods have been developed to preserve image textures. Chen *et al.* [10] adopt a two-step non-iterative processing procedure which first employs a simplified Wiener filter to obtain a distortion free but noisy estimate, and then utilizes a modified non-local means filter to reduce the leaked colored noise in order to obtain a good texture-preserving restoration.

Within the Bayesian paradigm, constraints on the characteristics of the resulting image are formulated as prior distributions. Two of the most classical prior distributions are conditional and simultaneous autoregressive (CAR and SAR) models [11]. They impose smoothness constraints on the original image and are able to preserve image textures better than TV, unfortunately, they oversmooth edge regions.

Carasso proposes in [12] a new approach to preserve image textures. He formulates the image restoration problem in a Lipschitz space where a broader set of images can be accommodated. The central idea is the introduction of the Poisson singular integral (PSI), which recovers the image texture in cases where TV fails. PSI is also utilized in [13], where it is combined with a curvelet-type decomposition to preserve textures while controlling the noise. Other methods based on wavelets and curvelets have also been proposed in combination with shrinkage-threshold rules [1, 14] to capture and preserve sharp features in the image.

Wang *et al.* [15] proposes to combine a weighted anisotropic TV (WATV)

[16] and tetrolet shrinkage [17]. WATV can recover sharp and clear edges along four directions, but this approach also eliminates image textures. To alleviate this problem, the tetrolet transform is used in combination with a TV regularizer.

Recently, a new approach that combines different priors has been used to solve super resolution [18], blind deconvolution [19, 20], astronomical and natural image [21] restoration problems. The idea behind the combination of priors is that using priors that preserve edges jointly with priors that preserve textures, can achieve better reconstructions than simply using one image prior. Notice that this idea is also related to the model in [19].

Using this approach, Vega *et al.* [21] tackle image restoration in Astronomy by combining a prior based on the the ℓ_1 norm of the horizontal and vertical first order differences which preserve edges and a simultaneous autoregression (SAR) prior model which preserves image texture. A similar approach was used by Villena *et al.* in super resolution problems [18]. The problem of blind deconvolution is addressed in [19] using Bayesian inference with super-Gaussian sparse image priors. This methodology can be used in blind and non-blind image deconvolution problems with the only knowledge of the noise variance.

Based on these recent developments, in this paper we propose a novel Bayesian image restoration algorithm that uses a combination of the TV and PSI prior models in order to preserve different properties on the restored image. This combination takes advantage of each prior: the TV prior preserves edge structure and removes the noise while the PSI prior preserves the image textures.

The rest of the paper is organized as follows. In Section 2, the TV and PSI models are presented within the Bayesian framework, their relations are established and an analysis of the PSI is presented. Section 3 discusses the inference procedure and proposes our algorithm to restore the images. Section 4 contains the experimental results and Section 5 concludes the paper.

2. Bayesian Modeling

The Bayesian paradigm is one of the most popular tools in image restoration (see [8] and references therein). The use of prior distributions that impose constraints on the estimates and act as regularizers, allows the introduction of additional information in the restoration process. In this section, we first model the image acquisition process to obtain the observed image from the original one and the blur and then introduce the proposed combination of priors models we will use.

2.1. Observation Model

It is usual to model the degradation process as a convolution between the original image and a known blurring operator that is expressed in vector-matrix notation as

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}, \tag{1}$$

where \mathbf{x} and \mathbf{y} are column vectors of size $P = m \times n$ obtained by lexicographically ordering the pixels in the original and observed image, respectively, \mathbf{H} is a known blurring matrix of size $P \times P$, and \mathbf{n} is Gaussian additive white noise with zero mean and precision β . From this degradation model, the conditional probability distribution of the observed image \mathbf{y} given the original

image \mathbf{x} and the noise precision parameter, β , is given by

$$p(\mathbf{y}|\mathbf{x}, \beta) \propto \exp\left(-\frac{\beta}{2}\|\mathbf{y} - \mathbf{H}\mathbf{x}\|^2\right). \quad (2)$$

2.2. Image Model

Digital images are discrete representations of continuous bidimensional signals, i.e., each image \mathbf{x} is assumed to have been obtained by discretizing a continuous bidimensional signal \mathbf{f} that belongs to the space of signals with bounded p -norm ($L^p(\mathbb{R}^2)$). In this space the continuous Total Variation (TV_c) semi-norm is defined as

$$TV_c(\mathbf{f}) = \int_{\mathbb{R}^2} \|\nabla \mathbf{f}(\mathbf{s})\|^2 ds. \quad (3)$$

Notice that for constant signal $\mathbf{f} \neq 0$, $TV_c(\mathbf{f}) = 0$ and, therefore, TV_c is not a norm. The equivalent semi-norm in discrete case is the Total Variation function, that is defined as

$$TV(\mathbf{x}) = \sum_{i=1}^P \sqrt{\Delta_i^h(\mathbf{x})^2 + \Delta_i^v(\mathbf{x})^2}, \quad (4)$$

where the operators $\Delta_i^h(\mathbf{x})$ and $\Delta_i^v(\mathbf{x})$ correspond to the horizontal and vertical first order differences at pixel i , respectively. In other words, $\Delta_i^h(\mathbf{x}) = x_i - x_{l(i)}$ and $\Delta_i^v(\mathbf{x}) = x_i - x_{a(i)}$ with $l(i)$ and $a(i)$ denoting the nearest neighbors to the left and above of pixel i , respectively. Using this energy function we obtain the so called TV [8] prior defined as

$$p_1(\mathbf{x}|\alpha_1) \propto \exp(-\alpha_1 TV(\mathbf{x})). \quad (5)$$

The TV prior has the advantage of preserving the edge structure while imposing smoothness on the solution.

This model implicitly imposes that the continuous total variation is bounded. However, it is demonstrated in [22] that the continuous signals corresponding to images with high texture have an unbounded total variation, and for this reason, the TV model fails on highly textured images.

Following [12], the space of bounded total variation ($BV(\mathbb{R}^2)$) is composed of all signals $\mathbf{f} \in L^p(\mathbb{R}^2)$ satisfying the constrain

$$\int_{\mathbb{R}^2} \|\mathbf{f}(\mathbf{s} + \mathbf{d}) - \mathbf{f}(\mathbf{s})\| ds \leq Const \|\mathbf{d}\|. \quad (6)$$

To preserve textures, Carasso [12] proposes to work in the Lipschitz (Besov) space $\Lambda(\alpha, 2, \infty)$, where the weaker constraint

$$\left\{ \int_{\mathbb{R}^2} \|\mathbf{f}(\mathbf{s} + \mathbf{d}) - \mathbf{f}(\mathbf{s})\|^2 ds \right\}^{1/2} \leq Const \|\mathbf{d}\|^\alpha, \quad 0 < \alpha < 1, \quad (7)$$

must be satisfied.

In [23] it is shown that \mathbf{f} belongs to the Lipschitz space $\Lambda(\alpha, 2, \infty)$ if, and only if

$$\sup_{t>0} t^{-\alpha} \|\mathbf{U}^t \mathbf{f} - \mathbf{f}\|_2 < \infty, \quad (8)$$

where \mathbf{U}^t is the Poisson integral operator defined as

$$\mathbf{U}^t \mathbf{f} = \int_{\mathbb{R}^2} \phi(x, y, t) \mathbf{f}(x - u, y - v) dudv, \quad (9)$$

and ϕ is the Poisson kernel in \mathbb{R}^2

$$\phi(x, y, t) = \frac{t}{2\pi(x^2 + y^2 + t^2)^{3/2}}. \quad (10)$$

Carasso [12] shows that this space contains a rich and significant class of images, and propose a restoration method for them. To force the signal \mathbf{f} to

be in $\Lambda(\alpha, 2, \infty)$, in [12] it is imposed that $\int_0^t \|\mathbf{U}^s \mathbf{f} - \mathbf{f}\|^2 ds$ is bounded, and it is demonstrated that

$$\int_0^t \|\mathbf{U}^s \mathbf{f} - \mathbf{f}\|^2 ds = \|\mathbf{Z}\mathbf{f}\|_2^2, \quad (11)$$

where \mathbf{Z} is the continuous convolution operator of the filter \mathbf{z} , which is defined in the Fourier domain as follows

$$\mathbf{z}(\xi, \nu, t) = \left(t + \frac{4e^{-t\rho} - e^{-2t\rho} - 3}{2\rho} \right)^{1/2}, \quad (12)$$

where ξ, ν are coordinates in the Fourier domain, $\rho = \sqrt{\xi^2 + \nu^2}$. By continuity, we have $\mathbf{z}(0, 0, t) = 0$. The obtained filter is normalized so that its squared components add to 1.

Using the discrete version of the convolution operator of the filter \mathbf{z} , \mathbf{Z} , obtained by sampling the filter at the resolution of the image, we can define the PSI based prior model

$$p_2(\mathbf{x}|\alpha_2) \propto \exp\left(-\frac{\alpha_2}{2}\|\mathbf{Z}\mathbf{x}\|^2\right), \quad (13)$$

where α_2 is the prior precision parameter.

The filter \mathbf{z} in Eq. (12), depends on a parameter t . To illustrate the effect of this parameter on the prior, Fig. 1 shows a set of realizations of the PSI prior model in Eq. (13) with precision $\alpha_2 = 1$, for four different values of t . As it can be observed t controls the smoothness of the texture. As t changes so does the texture granularity (notice the log scale).

Figure 2 shows the Fourier spectrum of the PSI filter. Notice that as t decreases, the radius from the center of non preserved frequencies increases.

To see this more clearly, Figure 3 depicts a transversal section of the PSI filter in the Fourier domain for different values of t . Here, we appreciate in

detail the described behavior. As the value of t decreases passing frequencies will diminish. Furthermore we can observe that the passing high frequency will be amplified. Notice that for values of $t < 10^{-3}$, the shape of the PSI filter almost does not vary, and therefore we can define the range of useful values of t in the interval $[10^{-3}, 1]$. In [12] it was found that the useful range of values was in the interval $(0, 1]$. Note, however, that in [12] the filter was not normalized so that its squared components add to 1, and, hence, t produces large variations that had to be compensated with large variations on the regularization parameter. For comparison purposes, Figure 3 also shows a transversal section of the SAR filter. The SAR filter presents a frequency response similar to the PSI when the value of t is close to 0, specially in the middle frequencies, but it attenuates more the low frequencies and does not amplify very high frequencies.

As it is well known, the high frequencies in an image are associated with abrupt changes, fine details, edges, and unfortunately also to noise in the spatial domain. In Fig. 4(b) the original *Barbara* image in Fig. 4(a) was filtered using the PSI filter with a value of $t = 1$, which produces an image very similar to the original where only a narrow band of low frequencies are eliminated. In Figure 4(c-d) the filtered images with a parameter $t = 0.1$ and $t = 0.03$ respectively are presented. In these figures the eliminated frequencies becomes more evident, smooth areas are lost, and the edges and fine details become sharper since the high frequencies are amplified. This effect is more notorious in Fig. 4(e) where the image was filtered using a value of $t = 0.001$, and all but the textures are removed, and edges and fine details like the trousers, scarf and tablecloth are highlighted. Hence, as

the value of t decreases, only high frequencies will be preserved and smooth regions will be removed. For comparison, Fig. 4(f) also shows the application of the SAR filter to the same image.

In this section, we have presented two prior models: the TV prior that preserves the edges but smooths textures and the PSI model that preserves the textures. To take advantage of the characteristics of both models, we combine them and define the following new prior

$$p(\mathbf{x}|\alpha_1, \alpha_2) \propto \exp(-\alpha_1 \text{TV}(\mathbf{x}) - \frac{\alpha_2}{2} \|\mathbf{Z}\mathbf{x}\|^2). \quad (14)$$

Once the prior and degradation models are defined, we perform inference to estimate the original image.

3. Bayesian Inference

In the inference stage we use the observation and prior models presented in the previous section, to obtain a maximum *a posteriori* (MAP) estimation of the restored image.

The MAP, $\hat{\mathbf{x}}$, satisfies

$$\hat{\mathbf{x}} = \arg \min_{\mathbf{x}} \left\{ \frac{\beta}{2} \|\mathbf{y} - \mathbf{H}\mathbf{x}\|^2 + \alpha_1 \text{TV}(\mathbf{x}) + \frac{\alpha_2}{2} \|\mathbf{Z}\mathbf{x}\|^2 \right\}. \quad (15)$$

Due to the form of the TV prior, the above objective function is difficult to evaluate. However, we can majorize the TV prior by a function which renders the function easier to calculate. Based on the average inequality [8], we utilize the following upper bound of the TV function

$$\text{TV}(\mathbf{x}) \leq \frac{1}{2} \sum_{i=1}^P \frac{\Delta_i^h(\mathbf{x})^2 + \Delta_i^v(\mathbf{x})^2 + u_i}{\sqrt{u_i}} = \frac{1}{2} \mathbf{M}(\mathbf{x}, \mathbf{u}), \quad (16)$$

where $\mathbf{u} \in (\mathbb{R}^+)^P$, is a P -dimensional vector with components u_1, u_2, \dots, u_P , that needs to be computed along with the image and has, as will be shown later, an intuitive interpretation related to the unknown image \mathbf{x} .

We then minimize

$$\bar{\mathcal{L}}(\mathbf{x}, \mathbf{u}) = \frac{\beta}{2} \|\mathbf{y} - \mathbf{H}\mathbf{x}\|^2 + \frac{\alpha_1}{2} \mathbf{M}(\mathbf{x}, \mathbf{u}) + \frac{\alpha_2}{2} \|\mathbf{Z}\mathbf{x}\|^2. \quad (17)$$

By alternating between the minimization of \mathbf{x} and \mathbf{u} .

For a given \mathbf{x} , we calculate \mathbf{u} as

$$\mathbf{u} = \arg \min_{\mathbf{u}} \sum_{i=1}^P \frac{\Delta_i^h(\mathbf{x})^2 + \Delta_i^v(\mathbf{x})^2 + u_i}{\sqrt{u_i}} \quad (18)$$

and, obtain

$$u_i = \Delta_i^h(\mathbf{x})^2 + \Delta_i^v(\mathbf{x})^2. \quad (19)$$

Note that that vector \mathbf{u} is a function of the spatial first order differences of the unknown image \mathbf{x} and represents its local spatial activity.

For a given \mathbf{u} , to obtain the estimation of the image, first notice that Eq. (17) can be rewritten as

$$\bar{\mathcal{L}}(\mathbf{x}, \mathbf{u}) = \lambda \|\mathbf{y} - \mathbf{H}\mathbf{x}\|^2 + \lambda_1 \mathbf{M}(\mathbf{x}, \mathbf{u}) + \lambda_2 \|\mathbf{Z}\mathbf{x}\|^2, \quad (20)$$

with $\lambda = (1 - \lambda_1 - \lambda_2)$,

$$\lambda_1 = \frac{\alpha_1}{\beta + \alpha_1 + \alpha_2} \quad \text{and} \quad \lambda_2 = \frac{\alpha_2}{\beta + \alpha_1 + \alpha_2}, \quad (21)$$

take values in the interval $[0, 1)$ and satisfy $\lambda + \lambda_1 + \lambda_2 = 1$. Thus, λ, λ_1 and λ_2 represent the relative influence on the restored image of the fidelity to the observed data and the combination of priors. Notice that selecting λ_1

and λ_2 in Eq. (20) is easier and more intuitive than selecting β , α_1 and α_2 in Eq. (17).

Then the MAP estimator, $\hat{\mathbf{x}}$, is obtained as the solution, utilizing for instance a conjugate gradient method, of the linear equation system

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{H}^T\mathbf{y}, \quad (22)$$

where

$$\mathbf{A} = \lambda\mathbf{H}^T\mathbf{H} + \lambda_1((\Delta^h)^T\mathbf{W}\Delta^h + (\Delta^v)^T\mathbf{W}\Delta^v) + \lambda_2\mathbf{Z}^T\mathbf{Z}, \quad (23)$$

Δ^h and Δ^v are the convolution matrices associated with horizontal and vertical gradients, respectively, and $\mathbf{W} = \text{diag}(1/\sqrt{u_i})$. This matrix controls the smoothness applied at each pixel of the image. So, for pixels in areas with a low spatial activity, the value of \mathbf{W} will be large, thus enforcing smoothness. In those areas, the PSI will be responsible for the texture preservation. However, for pixels in high spatial activity areas \mathbf{W} will be very small which means that no smoothness is enforced, thus preserving the edges and other features of the image.

The proposed restoration method is summarized in the Algorithm 1.

Notice that if we fix λ_2 to zero in Eq. (23) we have a Bayesian formulation of the TV model and, when $\lambda_1 = 0$ we use only the PSI model.

4. Experimental results

Before comparing the proposed method with other restoration approaches, we assess the influence of the parameter t . We blurred the region of the the original *Barbara* image depicted in Fig. 5(a) scaled to the range $[0, 1]$, using a Gaussian kernel of size 21×21 and standard deviation 1, and then added

Algorithm 1 Proposed Restoration Algorithm

Require: An initial estimate of the original image, \mathbf{x}^0

Set $k = 0$

repeat

1. Set $\mathbf{u}^k = \arg \min_{\mathbf{u}} \bar{\mathcal{L}}(\mathbf{x}^k, \mathbf{u})$.
2. Set $\mathbf{x}^{k+1} = \arg \min_{\mathbf{x}} \bar{\mathcal{L}}(\mathbf{x}, \mathbf{u}^k)$.
3. Set $k = k + 1$.

until $\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 / \|\mathbf{x}^k\|^2 < tol$

white Gaussian noise of variance 10^{-3} . It produced the observation shown in Fig. 5(b), whose peak signal-to-noise ratio is PSNR = 24.68 dB.

Since we want to assess the influence of t on the final restorations, we fixed $\lambda_1 = 0$, to see how the PSI restoration method works alone. Then we obtained the restoration for $\lambda_2 \in \{0, 10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 0.1, 0.2, 0.3, 0.4\}$ and $t \in \{0.001, 0.01, 0.03, 0.1, 1\}$.

In Figure 6(a), we plot the PSNR evolution for all considered couples of values (λ_2, t) . For a small values of λ_2 the PSI model has not much influence on the restorations, that it is not enough regularized and then low PSNR values are obtained. When λ_2 increases, such influence is higher and the obtained restorations achieve up to appoint better PSNR values.

The maximum is reached at the point $(\lambda_2, t) = (0.1, 0.01)$, and for values of λ_2 higher than 0.1 the PSNR slightly decreases. Figure 6(b) shows the PSNR values for fixed $\lambda_2 = 0.1$ and the different values of t , which highlights the influence of t on the final restorations.

To visually observe this behavior, Figures 5(d)–(f) depict the obtained restorations for $\lambda_2 = 0.1$ and $t = 0.001$, $t = 0.01$ and $t = 1$. When $t = 0.001$

we obtain the noisiest restoration (Fig. 5(d)), but if we look at the scarf and the chair behind *Barbara* we can see that the textures are more pronounced than in the other two restorations. On the other extreme, when $t = 1$ (Fig. 5(f)) we obtain the smoothest and less noisy restoration, but textures are less marked. Finally, when $t = 0.01$ the restored image (Fig. 5(e)) has an acceptable level of noise, while the textures remain quite marked. Hence, there exists a trade-off between the restored textures and level of noise in the image, which can be tuned by modifying the value of the parameter t . As we mentioned before, this result is also numerically supported, since the PSNR values for the restoration of the image in Fig. 5(b) with $t = 0.001$ and $t = 1$ are 26.02 dB and 25.65 dB, respectively, while the best PSNR, 26.05 dB, is obtained for $t = 0.01$. For comparison purposes, we included in Fig. 5(c) the restoration with the TV model, obtained with Alg. 1 by setting $\lambda_1 = 0.01$ and $\lambda_2 = 0$, which has a PSNR of 25.02 dB. This is an almost noise free image but the textures in some parts of the scarf are lost.

From this experiment we can conclude that (a) neither the TV nor the PSI image models alone are able to successfully recover the textures and control the noise and (b) that the parameter t of the PSI model will control the amount of texture in the image. In the following experiments we will show that a sensible combination of the TV and PSI models produces better results than using just one model alone.

To obtain the restored images using the TV, PSI, and TV+PSI priors, we run Alg. 1 starting from the degraded image as initial estimate of the original image, that is, $\mathbf{x}^0 = \mathbf{y}$, and used $tol = 10^{-4}$ for the stopping criterion in Alg. 1. To select the value of the parameters governing the weight of the

TV and PSI prior models on the final restoration, λ_1 and λ_2 , we performed a search in the set of values $\{0, 10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 0.1, 0.2, 0.3, 0.4\}$ and, as we have already indicated, we calculated the weight of the fidelity to the data term, λ , as $1 - \lambda_1 - \lambda_2$. Note that if $\lambda_2 = 0$, the TV model is used alone and that for $\lambda_1 = 0$ the PSI model is selected. We have experimentally observed that values of λ_1 or λ_2 larger than 0.4 reduce drastically the quality of the restored image so we did not consider them in our reported experiments. Additionally, the PSI prior in Eq. (13) depends on the parameter t . As we explained in section 2.2, the range of useful values for this parameter is $\{10^{-3}, 1\}$ so, we explored the range $t \in \{0.001, 0.01, 0.03, 0.1, 1\}$.

We compared the performance of the proposed method with several classical and state-of-the art methods. First, we used the classical method in [11], that uses a simultaneous autoregressive (SAR) prior model obtains a MAP estimate of the image and, simultaneously, estimates the model parameters by maximum likelihood.

Also, we compared with the method in [21] that proposes a combination of ℓ_1 and SAR prior models. The ℓ_1 prior model is similar to the TV prior but considers a different parameter for the horizontal and vertical first order differences. Following [21], we assumed an exact knowledge of the noise variance and let the method estimate only the ℓ_1 and SAR prior parameters. The combination parameter that controls the relative importance of the ℓ_1 and SAR restoration methods is selected by exploring the interval $[0, 1]$ in steps of 0.1 and selecting the one resulting in a better PSNR. Finally, we compared with the recently proposed log model in [19], named *General Sparse Prior*

(GSP). We want to note that, although the method in [19] was originally formulated as a blind deconvolution method, in this paper, we assume that the blur is known. We supplied the method with the real value for the noise variance.

We run all the restoration methods on four classic images in image processing: *Barbara*, *Cameraman*, *Baboon* and *Lena*. These test images have different levels of spatial activity and areas with different types of texture. The original images were synthetically degraded following the observation model in Eq. (1) after been scaled to the interval $[0, 1]$. Each image was blurred with a Gaussian blur with support 21×21 and standard deviation 1. Zero mean Gaussian noise with variance $\sigma_1^2 = 10^{-5}$, $\sigma_2^2 = 10^{-4}$ and $\sigma_3^2 = 10^{-3}$, was added to the blurred images to obtain three set of degraded images with SNR of about 50 db, 40 dB and 30 dB, respectively. We repeated each experiment 3 times to decrease the dependence on a given realization of the noise and report the mean value of the results for all experiments.

We present detailed results on two representative images and noise combinations and, finally, we summarized and extract conclusions for the complete set of experiments.

Figure 7(a) shows to the original *Baboon* image. To better appreciate the details in the images we show results in a small region of interest, marked with a square, of size 200×200 . In Fig. 7(b) the region of interest of the original image is depicted. It contains different features that allow us to evaluate how the method works. We can distinguish high frequency information as the hair or the details around the eye, as well as smoother zones as the nose. In Fig. 7(c) we shown the degraded image for a noise variance $\sigma_2^2 = 10^{-4}$.

The PSNR for the whole image is 22.99 dB. Using the SAR model restoration is depicted in Fig. 7(d). Notice that this model restores the details in the image; however it also amplifies the noise, as can be seen in the nose. In fact, the PSNR for this restoration is smaller than the one of the observation, 22.85 dB. On the other side, TV and GSP models, whose restoration are shown in Fig. 7(f) and 7(g), respectively, obtain the smoothest restorations and similar PSNR values (24.54 dB for the TV model and 24.52 dB for the GSP model). The zone of the nose is almost noise-free, but the details around the eye and in the hair are smoothed out. The PSI, ℓ_2 +PSI, ℓ_1 +SAR and the proposed method, shown in Fig. 7(e) and 7(g), 7(i) and 7(j), respectively, achieve a good balance between noise and texture, however, if we compare the zone of the nose, we observe that the proposed method better eliminates the noise, while preserving a similar quality in textures. The numerical results also support this fact. PSI model alone obtains PSNR = 24.71 dB, ℓ_2 +PSI model obtains PSNR = 24.74, and the ℓ_1 +SAR gets PSNR = 24.27 dB, while the proposed method obtain PSNR = 24.86 dB.

For the image of *Barbara*, shown in Fig. 8(a), the same behavior is repeated. In this case, we have selected an area of interest of size 256×256 , which is marked by the square, and in Fig. 8(b). The degraded image was generated with noise variance $\sigma_2^2 = 10^{-4}$, Fig. 8(c), has PSNR = 25.32 dB. In this case, the noise amplification and edge smoothness produced by the SAR model is much more evident (see Fig. 8(d)), getting PSNR = 24.14 dB. The TV and GSP models in Fig. 8(e) and 8(g), manage to eliminate the noise almost completely, obtaining PSNR = 26.75 dB and PSNR = 26.73 dB, respectively. However, it can be appreciated that some textures in the

scarf are lost. The PSI model (Fig. 8(e)), ℓ_2 +PSI model (Fig. 8(f)), ℓ_1 +SAR model (Fig. 8(h)) and TV+PSI model (Fig. 8(i)), are capable to restore the texture in the scarf and obtain PSNR = 27.26 dB, PSNR = 27.32, PSNR = 25.93 dB and PSNR = 27.55 dB, respectively. If we look at top and left corner of the images, we can see that the proposed model again removes the noise better than the PSI, and ℓ_1 +SAR models.

To summarize the experiments, in Tables 1 and 2 we report the mean PSNR values obtained in the experiments for the four images. We can see that the proposed method obtains the highest PSNR for all the restorations, except for the *Barbara* and *Cameraman* images when the noise variance is $\sigma_3^2 = 10^{-3}$. In these cases GSP obtain better results since it better controls high noise. Note that, in these cases, the differences between TV, GSP and TV+PSI methods is small for all the images. However, as the noise level is reduced the proposed TV+PSI method produces better results, especially in highly textured images. We want to note that the combination of PSI and TV models clearly improves over a single model, PSI or TV. In all the experiments the value of λ_1 and λ_2 was greater than zero, meaning that our method always included information from both models. Although we evaluated different values for the parameter t , it is worth mentioning that in most cases $t = 0.001$ produces good results. Also the proposed method is competitive with the two state-of-the-art methods we compared with, ℓ_1 +SAR model combination and GSP.

5. Conclusions

In this paper, we have presented a novel image restoration method that uses the Bayesian paradigm to combine two prior models: the TV model that preserves the edge structure while imposes smoothness on the solution and, the PSI model which is capable to preserve the textures. The final product is a restoration algorithm that combines the advantages of the two models. An study of PSI model and the parameter that controls its shape has been carried out, and concluded that neither the TV nor the PSI image models alone successfully recover the textures and control the noise. Finally a set of experiments has been carried out, where the proposed method has been compared against both classical and state of art methods. The experimental results supported the proposed model and demonstrated that TV + PSI obtains high-quality restorations.

Future work will adapt the model to the local image characteristics and perform automatic parameter estimation within the Bayesian framework.

- [1] M. Elad, Sparse and redundant representations, Springer, 2010.
- [2] R. Molina, J. Núñez, F. J. Cortijo, J. Mateos, Image restoration in astronomy: A Bayesian perspective, *IEEE Signal processing magazine* 18 (2) (2001) 11–29.
- [3] A. Buades, B. Coll, J. M. Morel, A review of image denoising algorithms, with a new one, *Multiscale Model and Simulation* 4 (2005) 490–530.
- [4] K. Dabov, A. Foi, V. Katkovnik, K. Egiazarian, Image denoising by sparse 3-D transform-domain collaborative filtering, *IEEE Transactions on Image Processing* 16 (8) (2007) 2080–2095.

- [5] R. Yan, L. Shao, Y. Liu, Nonlocal hierarchical dictionary learning using wavelets for image denoising, *IEEE Transactions on Image Processing* (2013) accepted for publication.
- [6] L. I. Rudin, S. Osher, E. Fatemi, Nonlinear total variation based noise removal algorithms, *Physica D* (1992) 259–268.
- [7] S. D. Babacan, R. Molina, A. K. Katsaggelos, Parameter estimation in TV image restoration using variational distribution approximation, *IEEE Transactions on image processing* 17 (3) (2008) 326–339.
- [8] S. D. Babacan, R. Molina, A. K. Katsaggelos, Variational Bayesian blind deconvolution using a total variation prior, *IEEE Transactions on Image Processing* 18 (1) (2009) 12–26.
- [9] Z. Dogan, S. Lefkimmiatias, A. Bourquard, M. Unser, A second-order extension of TV regularization for image deblurring, in: *Proceedings of the 18th IEEE International Conference on Image Processing (ICIP)*, 2011, pp. 713–716.
- [10] F. Chen, X. Huang, W. Chen, Texture-preserving image deblurring, *IEEE Signal processing letters* 17 (12) (2010) 1018–1021.
- [11] R. Molina, On the hierarchical Bayesian approach to image restoration. Applications to astronomical images, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 16 (11) (1994) 1122–1128.
- [12] A. S. Carasso, Singular integrals, image smoothness, and the recovery of texture in image deblurring, *SIAM Journal on Applied Mathematics* 64 (5) (2004) 1749–1774.

- [13] L. Huang, L. Xiao, Z. Wei, Z. Zhang, Variational image restoration based on Poisson singular integral and curvelet-type decomposition space regularization, in: Proceedings of the 18th IEEE International Conference on Image Processing (ICIP), 2011, pp. 685–688.
- [14] J.-L. Stark, F. Murtagh, J. M. Fadili, Sparse image and signal processing, Cambridge University, 2010.
- [15] L. Wang, L. Xiao, J. Zhang, Z. Wei, New image restoration method associated with tetrolets shrinkage and weighted anisotropic total variation, Signal processing 93 (4) (2013) 661–770.
- [16] X. Shu, N. Ahuja, Hybrid compressive sampling via new total variation TVL1, in: European Conference on Computer Vision (ECCV), 2010, pp. 393–404.
- [17] J. Krommweh, J. Ma, Tetrolet shrinkage with anisotropic total variation minimization for image approximation, Signal processing 90 (8) (2010) 2529–2539.
- [18] S. Villena, M. Vega, D. S. Babacan, R. Molina, A. K. Katsaggelos, Bayesian combination of sparse and non-sparse priors in image super resolution, Digital Signal Processing 23 (2) (2013) 530–541.
- [19] S. D. Babacan, R. Molina, M. Do, A. K. Katsaggelos, Blind deconvolution with general sparse image priors, in: European Conference on Computer Vision (ECCV), 2012, pp. 341–355.
- [20] E. Vera, M. Vega, R. Molina, A. K. Katsaggelos, Iterative image restoration using nonstationary priors, Applied Optics 52 (10) (2013) 102–110.

- [21] M. Vega, J. Mateos, R. Molina, A. K. Katsaggelos, Astronomical image restoration using variational methods and model combination, *Statistical Methodology* 9 (1-2) (2012) 19–31.
- [22] Y. Gousseau, J.-M. Morel, Are natural images of bounded variation?, *SIAM Journal on Mathematical Analysis* (33) (2001) 634–648.
- [23] M. Taibleson, On the theory of Lipschitz spaces of distributions on Euclidean n -space, *J. Math. Mechanics* 13 (1964) 407–478.

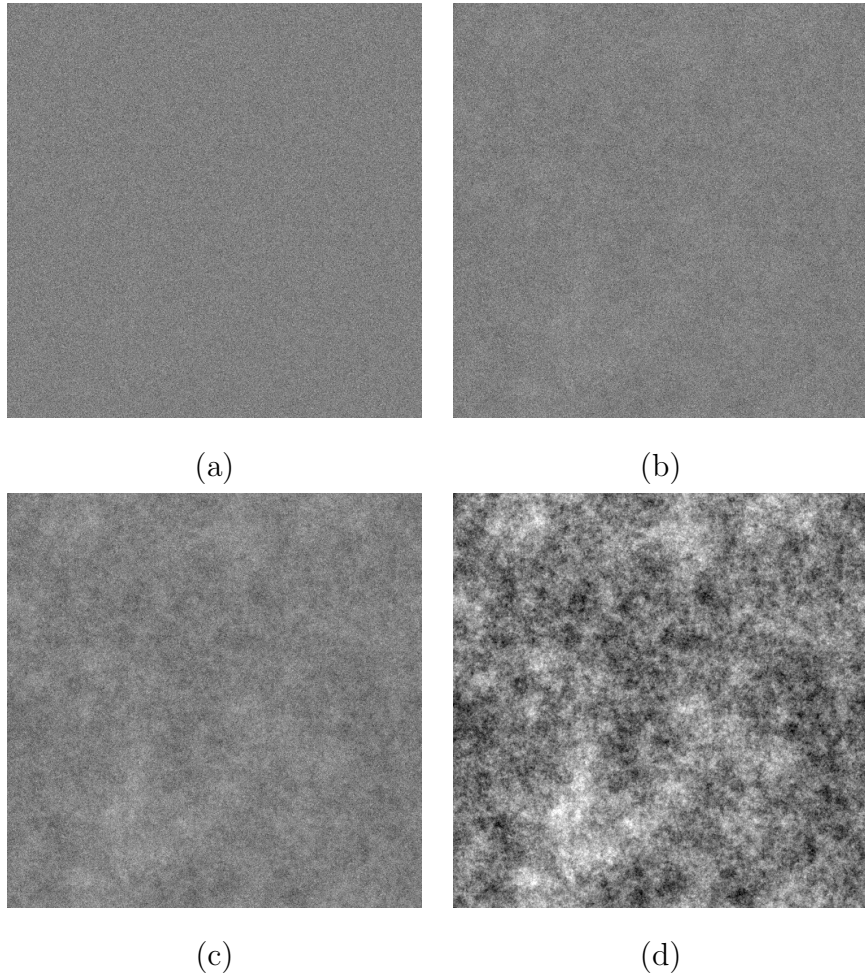


Figure 1: Realizations of the PSI prior model for different values of t , (a) $t = 1$, (b) $t = 0.1$, (c) $t = 0.03$ and (d) $t = 0.001$

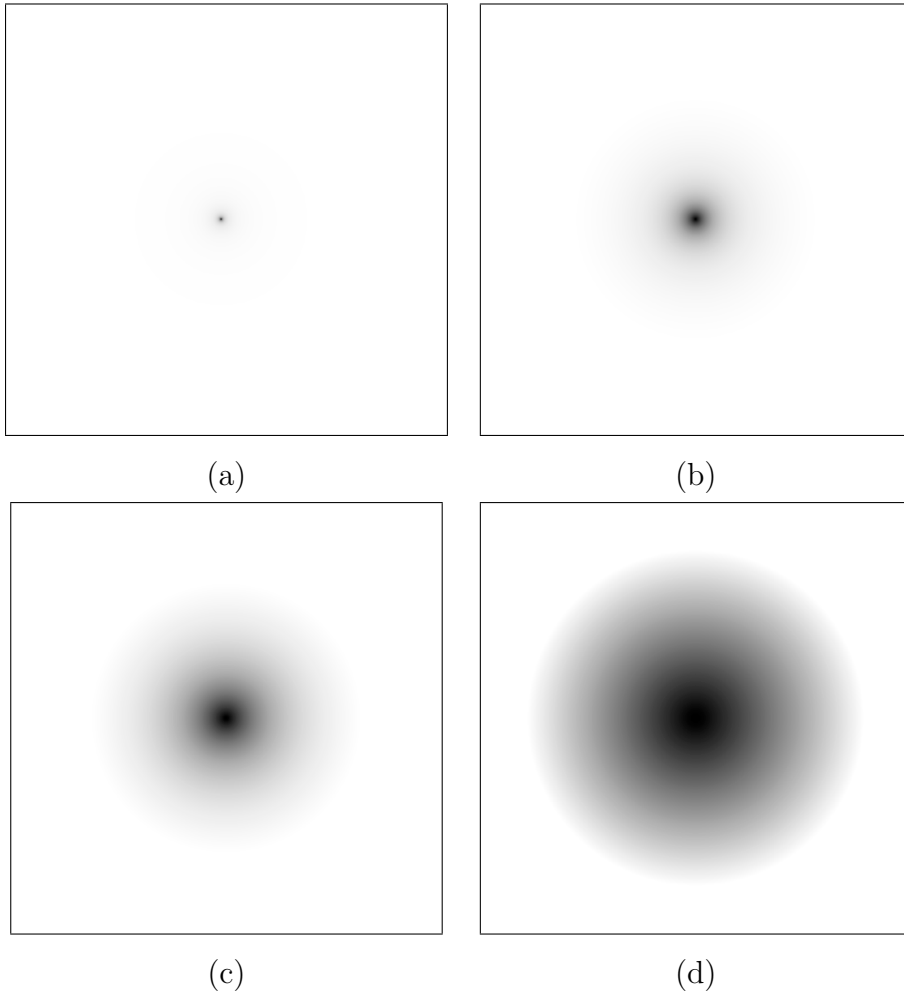


Figure 2: Fourier spectrum of the PSI filter for (a) $t = 1$, (b) $t = 0.1$, (c) $t = 0.03$, (d) $t = 0.001$.

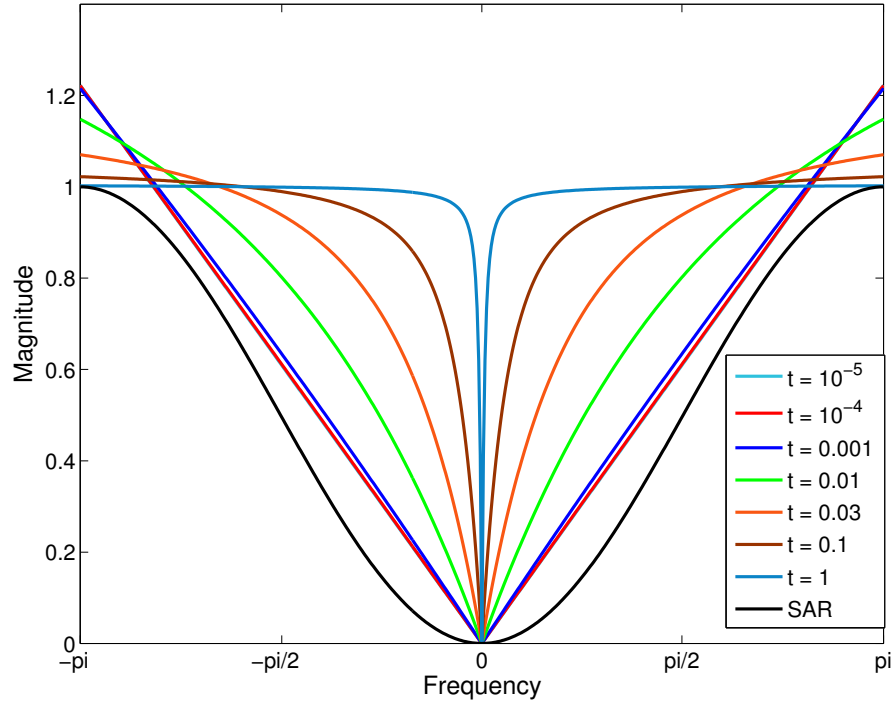


Figure 3: Transversal cut of SAR and PSI filters in Fourier domain for different values of t .



(a)

(b)



(c)

(d)



(e)

(f)

Figure 4: (a) Original *Barbara* image, filtered images with (b) $t = 1$, (c) $t = 0.1$, (d) $t = 0.03$, (e) $t = 0.001$ and (f) SAR.



(a)



(d)



(b)



(e)

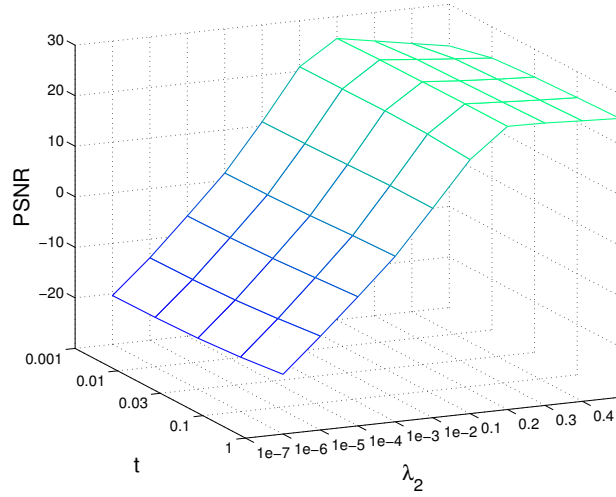


(c)

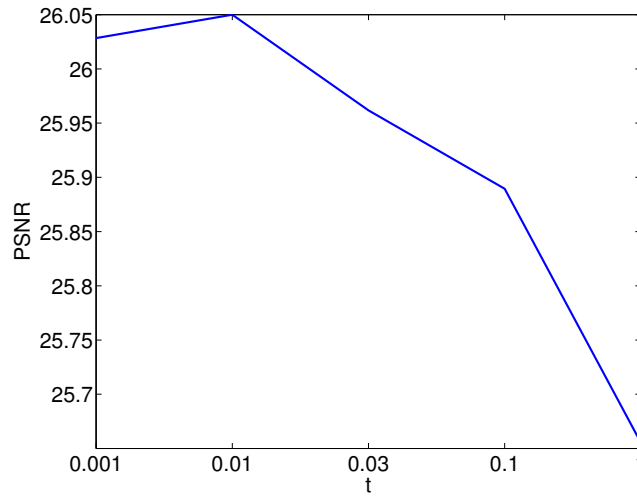


(f)

Figure 5: (a) Original *Barbara* image, (b) degraded observation, (c) restoration with the TV model. Restorations with PSI method for (d) $t = 0.001$, (e) $t = 0.01$ and (f) $t = 1$.



(a)



(b)

Figure 6: (a) PSNR values for the restorations using the PSI model varying parameters λ_2 and t . (b) PSNR values for the restorations using the PSI model fixing $\lambda_2 = 0.1$ and varying parameter t .

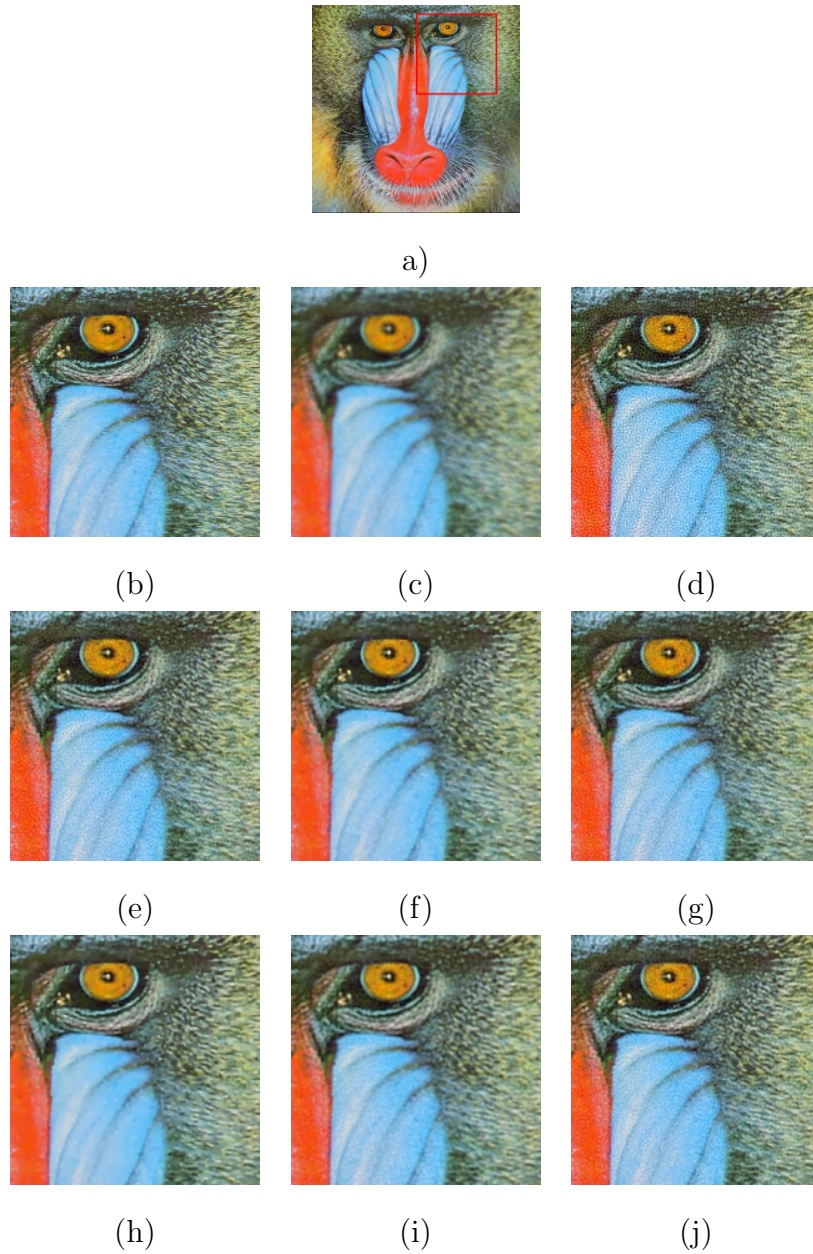


Figure 7: (a) Original image with the area of interest marked, (b) Area of interest of the original image, (c) Degraded image, (d) Restored image with the SAR model, (e) Restored image with the PSI model, (f) Restored image with the TV model, (g) Restored image with the $\ell_2 + PSI$ model, (h) Restored image with the GSP model, (i) Restored image with the $\ell_1 + SAR$ model, (j) Restored image with the proposed model.



a)



(b)



(c)



(d)



(e)



(f)



(g)



(h)



(i)



(j)

Figure 8: (a) Original image with the area of interest marked, (b) Area of interest of the original image, (c) Degraded image, (d) Restored image with the SAR model, (e) Restored image with the PSI model, (f) Restored image with the TV model, (g) Restored image with the $\ell_2 + PSI$ model, (h) Restored image with the GSP model, (i) Restored image with the $\ell_1 + SAR$ model, (j) Restored image with the proposed model.

Table 1: Peak signal to noise ratio results for *Baboon* and *Barbara* with different levels of noise.

Image	Method	PSNR	PSNR	PSNR	
		(SNR = 30 dB)	(SNR = 40 dB)	(SNR = 50 dB)	
<i>Baboon</i>	Observation	22.27	22.99	23.06	
	SAR	14.77	22.85	26.28	
	PSI	23.10	24.71	26.06	
	TV	22.65	24.54	26.19	
	$\ell_2 + PSI$	23.10	24.74	26.28	
	<i>GSP</i>	23.03	24.52	25.81	
	$\ell_1 + SAR$	22.62	24.27	25.74	
	TV + PSI	23.18	24.86	26.34	
	Optimum parameters		$\lambda_1 = 10^{-3},$	$\lambda_1 = 10^{-4},$	$\lambda_1 = 10^{-5},$
			$\lambda_2 = 10^{-1},$	$\lambda_2 = 10^{-2},$	$\lambda_2 = 10^{-3},$
		$t = 0.001$	$t = 0.001$	$t = 0.03$	
<i>Barbara</i>	Observation	24.16	25.32	25.46	
	SAR	14.45	24.14	30.74	
	PSI	24.87	27.26	29.91	
	TV	24.70	26.75	30.59	
	$\ell_2 + PSI$	24.87	27.32	30.79	
	<i>GSP</i>	25.02	26.73	30.13	
	$\ell_1 + SAR$	24.54	25.93	29.29	
	TV + PSI	24.93	27.55	30.93	
	Optimum parameters		$\lambda_1 = 10^{-3},$	$\lambda_1 = 10^{-4},$	$\lambda_1 = 10^{-5},$
			$\lambda_2 = 10^{-1},$	$\lambda_2 = 10^{-2},$	$\lambda_2 = 10^{-3},$
		$t = 0.01$	$t = 0.001$	$t = 0.001$	

Table 2: Peak signal to noise ratio results for *Lena* and *Cameraman* with different levels of noise.

Image	Method	PSNR	PSNR	PSNR	
		(SNR = 30 dB)	(SNR = 40 dB)	(SNR = 50 dB)	
<i>Lena</i>	Observation	27.91	31.43	32.01	
	SAR	15.09	24.67	33.33	
	PSI	30.69	33.02	34.51	
	TV	31.62	33.53	34.85	
	$\ell_2 + PSI$	30.69	33.02	34.58	
	<i>GSP</i>	31.52	32.94	33.45	
	$\ell_1 + SAR$	30.79	32.81	34.43	
	TV + PSI	31.62	33.61	34.91	
	Optimum parameters		$\lambda_1 = 10^{-2},$	$\lambda_1 = 10^{-3},$	$\lambda_1 = 10^{-4},$
			$\lambda_2 = 10^{-2},$	$\lambda_2 = 10^{-2},$	$\lambda_2 = 10^{-2},$
		$t = 0.001$	$t = 0.001$	$t = 0.001$	
<i>Cameraman</i>	Observation	24.51	25.82	25.97	
	SAR	14.38	24.15	30.67	
	PSI	25.95	27.85	30.38	
	TV	26.93	29.83	32.46	
	$\ell_2 + PSI$	25.95	28.24	30.55	
	<i>GSP</i>	27.41	29.32	30.29	
	$\ell_1 + SAR$	25.71	28.02	30.13	
	TV + PSI	26.94	29.84	32.47	
	Optimum parameters		$\lambda_1 = 10^{-2},$	$\lambda_1 = 10^{-3},$	$\lambda_1 = 10^{-4},$
			$\lambda_2 = 10^{-6},$	$\lambda_2 = 10^{-5},$	$\lambda_2 = 10^{-5},$
		$t = 0.03$	$t = 0.01$	$t = 1$	

3.2 Blind Deconvolution

- **P. Ruiz**, X. Zhou, J. Mateos, R. Molina, and A.K. Katsaggelos, “Variational Bayesian Blind Image Deconvolution: A Review”, *Digital Signal Processing*, 2015. doi:10.1016/j.dsp.2015.04.012
 - Status: Accepted for Publication. Available online 4 may 2015.
 - Impact Factor (JCR 2013): 1.495
 - Subject Category: Engineering, Electrical & Electronic (Q2: 98/248)

Variational Bayesian Blind Image Deconvolution: A Review

Pablo Ruiz, Xu Zhou, Javier Mateos, Rafael Molina and Aggelos K. Katsaggelos

Abstract

In this paper we provide a review of the recent literature on Bayesian Blind Image Deconvolution (BID) methods. We believe that two events have marked the recent history of BID: the predominance of Variational Bayes (VB) inference as a tool to solve BID problems and the increasing interest of the computer vision community in solving BID problems. VB inference in combination with recent image models like the ones based on Super Gaussian (SG) and Scale Mixture of Gaussians (SMG) representations have led to the use of very general and powerful tools to provide clear images from blurry observations. In the provided review emphasis is paid on VB inference and the use of SG and SMG models with coverage of recent advances in sampling methods. We also provide examples of current state of the art BID methods and discuss problems that very likely will mark the near future of BID.

I. INTRODUCTION

Thousands of millions of pictures are taken everyday. If the claim in [1] is right, 880 billion photos were taken in 2014. Every minute, 27,800 pictures are uploaded to Instagram, 208,300 photos are uploaded to Facebook and more than one thousand to Flickr, and the trend, with a digital camera in every mobile

P. Ruiz, J. Mateos, and R. Molina, are with the Departamento de Ciencias de la Computación e I. A. E.T.S. Ing. Informática y Telecomunicación. Universidad de Granada, 18071 Granada, Spain. (e-mail: {mataran,jmd,rms}@decsai.ugr.es).

X. Zhou is with the Image Processing Center, Beihang University, 100191 Beijing, China (e-mail: xuzhou@buaa.edu.cn).

A. K. Katsaggelos is with the Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL 60208 USA (e-mail: aggk@eecs.northwestern.edu).

Corresponding author: X. Zhou.

This paper has been partially supported by the Spanish Ministry of Economy and Competitiveness under project TIN2013-43880-R, the European Regional Development Fund (FEDER), the CEI BioTic at the Universidad de Granada, the National Natural Science Foundation of China (61233005) and the Department of Energy (DE-NA0002520).

phone, is probably exponentially increasing. The quality of these pictures varies widely from professional to amateur, in which case in many instances the images are taken under adverse conditions, such as low lighting or with motion between the camera and the scene, thus resulting in blurred images. While in some cases the introduction of blur in photography is intentional, being a powerful element of visual aesthetics, in most cases it is an undesirable effect degrading the quality of the image. Examples of the intentional introduction of blur includes the silky water effect obtained by using a long exposure when photographing a water flow (Fig. 1a), the bokeh effect obtained in parts of the scene lying outside the depth of field (Fig. 1b) and used to focus the attention of the viewer on a specific subject, or the motion blur effect (Fig. 1c) used to provide a sense of speed. Unintentional blur is caused by a number of causes, the most important ones being: camera or subject motion while the shutter is open (Fig. 1d) which leads to motion blur, out-of-focus (Fig. 1e) that blurs the whole the image or relevant parts of it or, simply, the presence of the atmosphere (Fig. 1f) as is the case with astrophotography.

Not only commercial photography is affected by blur. Modern science makes an intensive use of images in areas such as astronomy, remote sensing, medical imaging and microscopy and, in all of them, imperfections and characteristics of the capture system lead to images degraded during the observation process by blur, noise, and other degradations that diminish the quality and, hence, the value of the captured images.

Image deconvolution is a mature topic that aims at recovering the underlying original image from its blurred and noisy observations. Sometimes, the blur is completely or partially known or can be estimated prior to the deconvolution process. For instance, in astronomical imaging, an accurate representation of the blur can be obtained by imaging a single star first before photographing the astronomical object of interest. In contrast, blind image deconvolution (BID) tackles the restoration problem without knowing the blur in advance, leading to one of the most challenging image processing problems, since many combinations of blur and “true” image can produce the observed image. To start with, deconvolution is an ill posed problem in the Hadamard sense [2], that is, small variations in the data result in large variations in the solution. The problem is exacerbated in the BID problem, since in addition, small variations in the estimated blur can lead to large variations in the restored image.

BID is an underdetermined nonlinear inverse problem, which requires the estimation of many more unknown variables than the available observed data. To find meaningful solutions, not only prior information about the unknowns is crucial, but also a good and sound estimation approach. In this paper, we provide a comprehensive survey of BID methods reported since the publication of the review [3], with a focus on Bayesian approaches. In our opinion, since the publication of [3], Variational Bayes



Fig. 1. Blurred pictures due to intentional blur: a) silky water effect by Geraint Rowland (<https://www.flickr.com/photos/geezaweezer/15327097294/>), b) bokeh by Rodrigo Gomez (<https://www.flickr.com/photos/rgomez74/2970906336/>), c) motion blur by Ernest (<https://www.flickr.com/photos/viernest/3380560365/>). Blurred pictures due to unintentional blur: d) camera motion by tunguska (<https://www.flickr.com/photos/tunguska/103472115/>), e) out of focus by Nacho (<https://www.flickr.com/photos/gonmi/8193430914/>), f) atmosphere by Mike Durkin (<https://www.flickr.com/photos/madmiked/43831827/>).

(VB) inference has emerged as a dominant approach for the solution of BID problems. VB inference in combination with recently introduced image models, like the ones based on Super Gaussian (SG) and Scale Mixture of Gaussian (SMG) representation, has led to the development of very general and powerful tools to obtain clear images from blurry observations. We review the recent BID literature with an emphasis on VB inference and the use of SG and SMG models but without ignoring recent advances in sampling methods. We also provide examples of current state of the art BID methods and discuss problems that very likely will mark the near future of BID. The paper is organized as follows. In Section II, we briefly introduce the BID problem as well as the prior models. Section III shows the variational Bayesian methodology and its advantages over other inference approaches. We also present two representation models for variational inference, followed by the final BID algorithm. Section IV discusses some important outstanding challenges regarding the applications of VB based BID methods and BID as a whole research field. Experimental results are presented in Section V.

II. BAYESIAN PROBLEM FORMULATION

A. Bayesian framework for BID

In BID the image formation model is usually assumed to be:

$$\mathbf{y} = \mathbf{x} \otimes \mathbf{h} + \mathbf{n} = \mathbf{H}\mathbf{x} + \mathbf{n}, \quad (1)$$

where $\mathbf{y} \in \mathbb{R}^N$ is the observed blurred image (a column vector of N pixels), \otimes represents the convolution operation, $\mathbf{x} \in \mathbb{R}^N$ is the unknown original image, $\mathbf{H} \in \mathbb{R}^{N \times N}$ is the convolution matrix obtained from the also unknown blur kernel $\mathbf{h} \in \mathbb{R}^K$ and $\mathbf{n} \in \mathbb{R}^N$ is a noise term which is assumed to be i.i.d. Gaussian with variance β^{-1} . As discussed in section IV-D. other degradation models than the Linear and Spatially Invariant model above are also utilized.

Notice that although the BID problem is defined here in the image domain, it can also be easily formulated in transformed domains, such as the derivative, wavelet, and curvelet domains. The use of the filter space has gained popularity recently, however, there are still some open questions which need to be addressed before deciding which one is the right domain to work on, see section IV-A.

From a Bayesian perspective, given the observed blurred image \mathbf{y} , the goal is to infer the latent (hidden) variables $\mathbf{z} = \{\mathbf{x}, \mathbf{h}\}$ and possibly the model parameters denoted by Ω . The image degradation model in Eq. (1) can be written as:

$$p(\mathbf{y}|\mathbf{z}, \beta) = \mathcal{N}(\mathbf{y}|\mathbf{H}\mathbf{x}, \beta^{-1}\mathbf{I}), \quad (2)$$

where β is the precision parameter of the observation model, and possibly one of the model parameters to be estimated.

It is well known that the inverse problem of Eq. (1) is ill-posed [3]. Therefore, additional information on the latent variables and model parameters must be provided. The Bayesian paradigm introduces this necessary information for the BID problem as a prior distribution $p(\mathbf{z}|\Omega)$, which models the information on \mathbf{z} , and a prior $p(\Omega)$ on the model parameters. Sometimes the prior on the model parameters is called hyperprior and the elements of Ω are called hyperparameters.

With these ingredients, the global modeling of the BID problem can be written as

$$p(\mathbf{z}, \Omega, \mathbf{y}) = p(\mathbf{y}|\mathbf{z}, \Omega)p(\mathbf{z}|\Omega)p(\Omega). \quad (3)$$

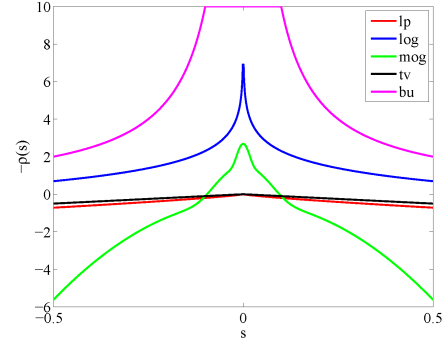
Before describing how inference is performed, we will now review the image, blur and hyperparameters priors proposed for the BID problem since the publication of [3].

Prior	$\rho(s)$
TV	s
ℓ_p	$\frac{1}{p} s ^p$
log	$\log s $
BU	$-s^{-1}$
MOG	$-\log \sum_j \frac{\pi_j}{\sqrt{2\pi}\sigma_j} \exp\left(-\frac{s^2}{2\sigma_j^2}\right)$

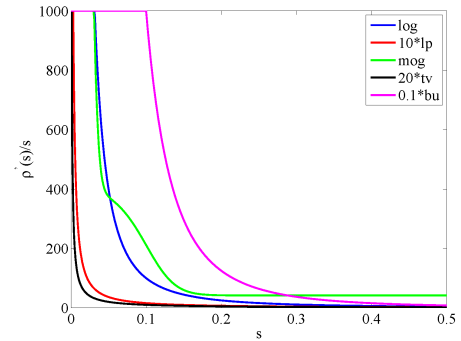
(a)

Prior	$\rho'(s)/s$
TV	s^{-1}
ℓ_p	$ s ^{p-2}$
log	$ s ^{-2}$
BU	s^{-3}
MOG	$\sum_j \frac{\frac{\pi_j}{\sigma_j} \exp\left(-\frac{s^2}{2\sigma_j^2}\right)}{\sigma_j^2 \sum_j \frac{\pi_j}{\sigma_j} \exp\left(-\frac{s^2}{2\sigma_j^2}\right)}$

(c)



(b)



(d)

Fig. 2. (a) and (b): examples of penalty functions $\rho(s)$, where the MOG is obtained from Levin *et al.* [4]. (c): their corresponding $\rho'(s)/s$. (d): plots of $\rho'(s)/s$, where an upper bounding is taken for visualization. Note that TV is replaced with anisotropic TV (ℓ_1 prior) since isotropic TV cannot be shown in 1-D function.

B. Image prior models

An unquestionable landmark on the recent history of BID is the paper by Miskin and MacKay [5]. In that work the authors propose the use of a mixture of Laplacians to restore cartoon images and utilize, for the first time in the BID literature, VB inference (to be described later) to restore the observed image. Later, Likas and Galatsanos [6] proposed a Gaussian prior to impose smoothness on the image and blur, see also [7], and Fergus *et al.* [8] proposed a mixture-of-Gaussians (MOG) to impose sparsity.

The 2007 Bishop *et al.* [3] review on BID describes, among others, classical prior models such as Conditional Autoregression (CAR) or Simultaneous Autoregression (SAR) used by Molina *et al.* [9] to impose smoothness, or Total Variation proposed by Rudin, Osher and Fatemi [10] to impose piecewise-smoothness. The TV prior model has been frequently used in BID, see for instance [11]–[15], see also [16] and [17]. Fergus *et al.* [8] represents the first publication on the use of filtered versions of the original

image to estimate image and blur. The use in [12] of majorization methods with variational inference and diagonal covariance approximation led to a new way to approach BID in image processing (not widely acknowledged in the computer vision community). As we will see in the following, the TV prior used in [12] is a particular case of the use of Super Gaussian Distributions in BID.

Since the influential work of Fergus *et al.* [8], sparse prior models have attracted the attention of BID researchers and are, in our opinion, rightly considered to be the state of the art representation in filtered domains. It is a well known fact that when high-pass filters are applied to natural images, the resulting coefficients are sparse; i.e., most of the coefficients are zero or very small while only a small number of them are large (e.g., at the edges). This is a very important characteristic that should be taken into account when restoring natural images.

The ℓ_p prior has been used in a large number of works such [12], [13], [18]–[21]. They use a prior distribution based on the minimization of quasi-norms $\|\cdot\|_p^p$ with $0 < p \leq 1$. Levin *et al.* [18] suggest the use of p in the range [0.6, 0.8] for natural images.

The Super-Gaussianity property presented by Palmer in [22], was used in Babacan *et al.* [23] as the building block to propose a general representation for sparse priors. As we will see, almost all previous and very recently proposed prior models can be represented using SG. This representation is used in the same work [23] to introduce two new image priors *log* and *exp*. Recent models like the one proposed by Zhang and Wipf [24], or the Student-t prior recently proposed by Mohammad-Djafari [25] are particular cases of SG distributions.

1) *Sparse General Representation*: A probability distribution is considered to be sparse when it is Super Gaussian (SG) [22], i.e., compared to the Gaussian distribution, it has heavier tails, it is more peaked, and has a positive excess kurtosis. These distributions are referred to as sparse since most of the distribution mass is located around zero (hence strongly favoring zero values), but the probability of occurrence of large signal values is higher compared to the Gaussian distribution.

Babacan *et al.* [23] propose the use of the following general framework to define the prior model either in the image or the filter space. First they consider L high-pass filters $\{f_\gamma\}_{\gamma=1}^L$ (such as derivatives, wavelets, curvelets, etc.) and define

$$x_\gamma = f_\gamma \otimes \mathbf{x}, \quad \gamma = 1, \dots, L. \quad (4)$$

Using these filters on the real underlying image the following prior model in the image space can be defined

$$p(\mathbf{x}) \propto \prod_{\gamma=1}^L \prod_{i=1}^N \exp(-\alpha_\gamma \rho(x_\gamma(i))), \quad (5)$$

where $\rho(\cdot)$ is an energy function symmetric around zero with $\rho(\sqrt{s})$ increasing and concave for $s \in (0, \infty)$ [22] and α_γ a scale parameter.

Alternatively, the filtered original images can be assumed to be independent. The following set of independent priors is then considered

$$p(x_\gamma) \propto \prod_{i=1}^N \exp(-\rho(x_\gamma(i))), \gamma = 1, \dots, L. \quad (6)$$

In this case, a set of blurred and noisy observations can be defined, associated with the filtered original images

$$y_\gamma = f_\gamma \otimes \mathbf{y} = \mathbf{h} \otimes f_\gamma \otimes \mathbf{x} + f_\gamma \otimes \mathbf{n} = \mathbf{h} \otimes x_\gamma + n_\gamma, \quad (7)$$

where n_γ is assumed to be Gaussian independent noise with precision β . It is important to note that the observations y_γ , $\gamma = 1, \dots, L$, are assumed to be independent and they provide information on the blur but not exactly on \mathbf{x} but on its filtered versions.

Notice that the most popular recent prior models, such as TV, ℓ_p , or MOG are Super Gaussian distributions (see Fig. 2 for some examples), and therefore can be represented using Eq. (5). Notice also in Fig. 2, that \log enforces sparsity very strongly due to its infinite peak at zero and heavy tails.

A sub-class of Super Gaussian distributions is the so called Scale Mixture of Gaussians (SMG), proposed by Andrews and Mallows [26] and used as a general framework for BID in Babacan *et al.* [23]. Here, associated with each filter γ and each pixel i we have

$$p(x_\gamma(i)) = \int p(x_\gamma(i)|\xi_\gamma(i))p(\xi_\gamma(i))d\xi_\gamma(i), \quad (8)$$

where $p(x_\gamma(i)|\xi_\gamma(i))$ is a Gaussian distribution with precision $\xi_\gamma(i)$. This model can also benefit from the introduction of a global scale parameter α_γ in Eq. (5).

SMG requires complete monotonicity of $p(\sqrt{s})$, i.e., $(-1)^n p(\sqrt{s})^n \geq 0$ must be satisfied for all $n = 0, 1, 2, \dots$. As can be seen in [22], this representation is more strict, in the sense that fewer classes of sparse priors can be represented with it than using Eq. (5). Finding $p(\xi_\gamma(i))$ is in general a difficult task; however, as we will see in Section III its full knowledge is not needed for our purposes. One example of SMG is the Student-t prior proposed by Mohammad-Djafari [25]. It is clear that an MOG is an SMG model and that spike and slab distributions on $z \in \mathbb{R}$, $p(z) = \lambda\delta(z) + (1 - \lambda)\mathcal{N}(z|0, \sigma^2)$ [27] are the limit of MOG models with two components, one of them with very small variance. Inference with these models is complicated due to the image size; however variational inference can still be carried out, see [28]. Notice that sparse promoting spike and slab and Bernoulli-Gaussian [27], [29] priors will

very likely receive more attention by the BID community especially when estimating the blur in the filter space.

C. Blur models

Although the above described prior models were proposed for the image, all of them can also be used for the blur as well. The BID literature also contains specific blur models which we now describe. Molina *et al.* [9] propose a Dirichlet prior for kernel modeling. Since the curvelet representation can take into account both the continuity and sparsity of the motion blur kernel, Cai *et al.* [30] suggest the use of this representation for this type of blur. Oh and Kim [31] propose a piecewise-linear model for motion blur in order to reduce the dimensionality of the solution space and make the kernel estimation process more robust.

Based on the assumption that the power spectrum of natural images drops quadratically as the frequency increases Goldstein and Fattal [32] introduce a power spectrum prior on the blur kernel. Recently, a novel convex blur regularizer based on the spectral properties of the convolution operators can be found in [33]. Since the spectral properties used are based on a linear and shift-invariant model without considering the noise, these methods do not work well for spatially varying blurs and noisy observations.

Let us consider the observation model in Eq. (1) and assume that the original image \mathbf{x} is known. In this case we have N observations and aim at estimating K coefficients, where K is the size of the blur. Since the image size is usually much larger than the blur size, N observations should be sufficient to obtain a good blur estimate, even more so if L filtered observations are used. Based on the fact that usually $K \ll N$, many authors [15], [23], [34], [35] have recently advocated the use of flat priors on the blur, enforcing only its nonnegativity and normalization constraints.

D. Hyperparameters models

So far we have studied the distributions $p(\mathbf{z}|\Omega)$, $p(\mathbf{y}|\mathbf{z}, \Omega)$ that appear in the Bayesian modeling of the BID problem in Eq. (3). We complete this modeling by studying now the distribution $p(\Omega)$.

An important problem is the estimation of the vector of parameters Ω when they are unknown. To deal with this estimation problem, the hierarchical Bayesian paradigm introduces a second stage, where the hyperprior $p(\Omega)$ is also formulated.

For parameters, ω , corresponding to inverses of variances, the gamma distribution is used. It is defined by:

$$p(\omega) = \Gamma(\omega|a_\omega, b_\omega) = \frac{(b_\omega)^{a_\omega}}{\Gamma(a_\omega)} \omega^{a_\omega-1} \exp[-b_\omega \omega], \quad (9)$$

where $\omega > 0$ denotes a hyperparameter, $b_\omega > 0$ is the rate parameter, and $a_\omega > 0$ is the shape parameter. These parameters are assumed known. The gamma distribution has the following mean, variance, and mode:

$$\mathbb{E}(\omega) = \frac{a_\omega}{b_\omega}, \text{Var}(\omega) = \frac{a_\omega}{b_\omega^2}, \text{Mode}(\omega) = \frac{a_\omega - 1}{b_\omega}. \quad (10)$$

Note that the mode does not exist when $a_\omega \leq 1$ and that mean and mode do not coincide. The literature also reports the use of non-informative prior models, $p(\Omega) \propto \text{constant}$, which can be considered as the limits of the above described hyperpriors.

Finally, we would like to mention here that the SG and SMG formulations turn the parameter estimation into a difficult task, especially when several filtered images are considered, since their partition functions can not usually be calculated.

III. BAYESIAN INFERENCE

Once the observation and prior models have been described, in other words, once the elements of the joint probability model in (3) have been specified, the goal now becomes the drawing of inference of the unknown variables $\Theta = \{\mathbf{z}, \Omega\}$ given the observations.

In the Bayesian framework Θ is inferred calculating (or approximating) the posterior distribution $p(\Theta|\mathbf{y})$, expressed using the Bayes' rule as

$$p(\Theta|\mathbf{y}) = \frac{p(\Theta, \mathbf{y})}{p(\mathbf{y})} = \frac{p(\mathbf{y}|\Theta)p(\mathbf{z}|\Omega)p(\Omega)}{p(\mathbf{y})}. \quad (11)$$

Unfortunately, since the integral $p(\mathbf{y}) = \int p(\Theta, \mathbf{y})d\Theta$ is not tractable, the above posterior cannot be analytically calculated. Different estimation methods have been proposed to address this problem in the BID context and we will now review them.

Probably the most widely used method in the literature is Maximum a Posteriori (MAP). Since $p(\Theta|\mathbf{y}) \propto p(\Theta, \mathbf{y})$ the maximum of the posterior distribution can be obtained by maximizing the joint distribution $p(\Theta, \mathbf{y})$ with respect to Θ . However, as pointed out in the landmark papers by Levin *et al.* [4], [34], MAP is not a suitable estimation procedure in BID problems, because the associated cost function favors flat images for many sparse priors and leads to a delta blur estimate. To avoid the delta blur solution, Perrone and Favaro [14] show that a delayed normalization [11] should be used while other authors [36]–[38] suggest using non-dimensional sparsity measures.

Another very popular inference method is MAP_h [39] [34] [40] [24]. Unlike MAP, this method integrates the joint distribution with respect to \mathbf{x} before estimating \mathbf{h} and Ω , that is, blur and parameters

TABLE I
COMPARISON OF INFERENCE METHODS

	MAP	MAP _h	VB	MCMC
Has full posterior	no	partial	yes	yes
Has point estimates	yes	yes	yes	yes
Has uncertainty info	no	partial	yes	yes
Allows hidden data	no	yes	yes	yes
Complexity	low	low	medium	high

are estimated by maximizing the evidence [41]. The restored image is finally calculated by maximizing the joint distribution, using the estimate values of \mathbf{h} and Ω through the above integration on the image.

Variational Bayesian inference has been widely used in BID (see [4]–[8], [13], [20], [23], [24], [40]). VB generalizes MAP and MAP_h (see [42] for a proof) providing approaches for estimation of the posterior distributions of \mathbf{x} , \mathbf{h} and Ω .

Together with the well established use of VB inference in BID, Markov Chain Monte Carlo (MCMC) methods are also gaining popularity. MCMC is the most general method used to approximate a posterior distribution, see [43] [44] [45] for details. The model in Eq. (3) is used to generate thousands of samples of $p(\mathbf{z}, \Omega | \mathbf{y})$, which are used to infer the posterior distribution. In theory, sampling methods can find the exact form of the posterior distribution, but in practice they are computationally intensive (especially for multidimensional signals such as images) and their convergence is hard to establish.

In computationally cost terms, VB is much more efficient than MCMC, and more expensive than MAP or MAP_h. The features of each method are summarized in Table I.

We now describe the application of VB and MCMC to BID.

A. Variational Inference in the image space for Super Gaussian priors

Since SG distributions are flexible enough to represent most of the image models used in the BID literature, we restrict, without loss of generality, the VB description to this representation. Furthermore we will formulate the inference in the image space; a detailed account of the use of SMG for the filter representation can be found in [23].

As it has already been explained above, the posterior distribution cannot be calculated analytically using the Bayes' rule in Eq. (11). To approximate $p(\mathbf{z}, \Omega | \mathbf{y})$, VB minimizes the following Kullback-Leibler

divergence

$$\text{KL}(q(\Theta)||p(\Theta|\mathbf{y})) = \int q(\Theta) \log \frac{q(\Theta)}{p(\Theta|\mathbf{y})} d\Theta = \int \frac{q(\Theta)}{p(\Theta, \mathbf{y})} d\Theta + \text{const}. \quad (12)$$

The Kullback-Leibler divergence is always non negative and is zero if and only if $q(\Theta) = p(\Theta|\mathbf{y})$. Since the minimizer $q(\Theta) = p(\Theta|\mathbf{y})$ cannot be calculated, some assumptions on $q(\Theta)$ have to be made. One possible assumption is that $q(\Theta)$ has a specific parametric form, e.g., a Gaussian distribution. Another widely used assumption is that $q(\Theta)$ factorizes into disjoint groups, i.e.,

$$q(\Theta) = q(\mathbf{x})q(\mathbf{h})q(\Omega). \quad (13)$$

This factorized form of variational inference is called mean field theory in physics [46].

Using Eq. (13), the KL divergence can be minimized with respect to each factor while holding the other factors fixed. The optimal solution for each factor is then [47]

$$\log q(\theta) = \mathbb{E} [\ln p(\Theta, \mathbf{y})]_{q(\bar{\Theta})} + \text{const}, \quad (14)$$

where $\bar{\Theta} = \Theta \setminus \theta$ is the set of unknowns excluding θ and $\mathbb{E} [\ln p(\Theta, \mathbf{y})]_{q(\bar{\Theta})}$ denotes the expectation taken with respect to all the approximating factors $\bar{\Theta}$. This system of equations is solved by an alternating minimization procedure, where each distribution $q(\theta)$ is iteratively updated using the latest distributions of all the other factors. Since the KL divergence (12) is convex with respect to $q(\theta)$ [48], the convergence of this alternating minimization procedure is guaranteed.

The penalty function $\rho(\cdot)$ defined in (5) can be represented as (see [49])

$$\rho(s) = \inf_{\xi > 0} \frac{1}{2} \xi s^2 - \rho^* \left(\frac{1}{2} \xi \right), \quad (15)$$

where $\rho^*(\xi/2)$ is the concave conjugate function

$$\rho^* \left(\frac{1}{2} \xi \right) = \inf_s \frac{1}{2} \xi s^2 - \rho(s). \quad (16)$$

Furthermore, the infimum in (15) is achieved at $\xi = \rho'(s)/s$, as shown in [23]. Directly applying VB inference using $p(\mathbf{x}, \mathbf{h}, \mathbf{y})$ is unfeasible, since the expectation of the logarithm of the joint distribution with respect to $q(\mathbf{x})$ is intractable.

Since $\rho(s)$ is the penalty associated to a SG distribution we can write

$$p(\mathbf{x}) \geq Z \prod_{\gamma=1}^L \prod_{i=1}^N \exp(-\alpha_{\gamma} (\frac{\xi_{\gamma}(i)}{2} x_{\gamma}^2(i) - \rho^*(\frac{1}{2} \xi_{\gamma}(i)))) , \forall \xi_{\gamma}(i) > 0, \quad (17)$$

where Z is a constant. This Gaussian like lower bound will allow the expectation of the joint distribution to be calculated analytically. We have

$$\begin{aligned}
p(\mathbf{x}, \mathbf{h}, \mathbf{y}) &\geq p(\mathbf{y}|\mathbf{x}, \mathbf{h})p(\mathbf{h})Z \prod_{\gamma=1}^L \prod_{i=1}^N \exp(-\alpha_{\gamma}(\frac{\xi_{\gamma}(i)}{2}x_{\gamma}^2(i) - \rho^*(\frac{1}{2}\xi_{\gamma}(i)))) \\
&= M(\mathbf{y}, \mathbf{x}, \mathbf{h}, \boldsymbol{\xi}),
\end{aligned} \tag{18}$$

where $\boldsymbol{\xi} = \{\xi_{\gamma}(i), \gamma = 1, \dots, L, i = 1, \dots, N\}$ with all component positive.

We then have

$$\begin{aligned}
&\int \int q(\mathbf{x})q(\mathbf{h}) \log \frac{q(\mathbf{x})q(\mathbf{h})}{p(\mathbf{x}, \mathbf{h}, \mathbf{y})} d\mathbf{x}d\mathbf{h} \\
&\leq \int \int q(\mathbf{x})q(\mathbf{h}) \log \frac{q(\mathbf{x})q(\mathbf{h})}{M(\mathbf{y}, \mathbf{x}, \mathbf{h}, \boldsymbol{\xi})} d\mathbf{x}d\mathbf{h} \\
&= \int \int \int q(\boldsymbol{\xi})q(\mathbf{x})q(\mathbf{h}) \log \frac{q(\boldsymbol{\xi})q(\mathbf{x})q(\mathbf{h})}{M(\mathbf{y}, \mathbf{x}, \mathbf{h}, \boldsymbol{\xi})} d\mathbf{x}d\mathbf{h}d\boldsymbol{\xi},
\end{aligned} \tag{19}$$

where $q(\boldsymbol{\xi})$ is a degenerate distribution on $\boldsymbol{\xi}$.

We then minimize the above integral on $q(\boldsymbol{\xi}), q(\mathbf{x})$, and $q(\mathbf{h})$ assuming that $q(\mathbf{h})$ is degenerate. According to (14), we obtain

$$\hat{\mathbf{h}} = \arg \max_{\mathbf{h}} \mathbb{E}[\log(M(\mathbf{y}, \mathbf{x}, \mathbf{h}, \hat{\boldsymbol{\xi}}))]_{\hat{q}(\mathbf{x})}, \tag{20}$$

$$\hat{q}(\mathbf{x}) \propto M(\mathbf{y}, \mathbf{x}, \hat{\mathbf{h}}, \hat{\boldsymbol{\xi}}), \tag{21}$$

$$\hat{\boldsymbol{\xi}} = \arg \max_{\boldsymbol{\xi} > 0} \mathbb{E}[\log(M(\mathbf{y}, \mathbf{x}, \hat{\mathbf{h}}, \boldsymbol{\xi}))]_{\hat{q}(\mathbf{x})}. \tag{22}$$

B. Estimation of blur, image, and variational parameters

For the latent image, we obtain from (21),

$$\log \hat{q}(\mathbf{x}) = -\frac{\beta}{2} \|\hat{\mathbf{H}}\mathbf{x} - \mathbf{y}\|_2^2 - \frac{1}{2} \sum_{\gamma=1}^L \alpha_{\gamma} x_{\gamma}^T \text{diag}(\hat{\xi}_{\gamma}) x_{\gamma}, \tag{23}$$

which is a multivariate Gaussian with precision matrix

$$\mathbf{C}_{\mathbf{x}}^{-1} = \beta \hat{\mathbf{H}}^T \hat{\mathbf{H}} + \sum_{\gamma=1}^L \alpha_{\gamma} \mathbf{F}_{\gamma}^T \text{diag}(\hat{\xi}_{\gamma}) \mathbf{F}_{\gamma}, \tag{24}$$

where \mathbf{F}_{γ} is an $N \times N$ convolution matrix formed by the filter f_{γ} , and $\hat{\mathbf{H}}$ is an $N \times N$ convolution matrix obtained from $\hat{\mathbf{h}}$. The mean value $\hat{\mathbf{x}}$ is used as the estimate for \mathbf{x} , which is obtained by solving the following system of linear equations

$$\mathbf{C}_{\mathbf{x}}^{-1} \hat{\mathbf{x}} = \beta \hat{\mathbf{H}}^T \mathbf{y}. \tag{25}$$

For the variational parameter $\hat{\xi}$, we obtain from (22)

$$\hat{\xi}_\gamma(i) = \frac{\rho'(\nu_\gamma(i))}{\nu_\gamma(i)}, \quad (26)$$

where $\nu_\gamma(i) = \sqrt{\mathbb{E}[x_\gamma^2(i)]}$, $1 \leq i \leq N$, with the expected value calculated using the distribution $\hat{q}(\mathbf{x})$.

To estimate the blur we have

$$\hat{\mathbf{h}} = \arg \min_{\mathbf{h}} \|\mathbf{H}\hat{\mathbf{x}} - \mathbf{y}\|_2^2 + \mathbf{h}^T \mathbf{D}_x \mathbf{h}, \quad (27)$$

$$\text{subject to } \mathbf{h}(i) \geq 0, \sum_{i=1}^K \mathbf{h}(i) = 1, \quad (28)$$

where \mathbf{D}_x is a $K \times K$ matrix given by

$$\mathbf{D}_x(m, n) = \sum_{j=1}^N \mathbf{C}_x(m+j, n+j). \quad (29)$$

To estimate the variational parameters in Eq. (26) and the blur in Eq. (27), the matrix \mathbf{C}_x is required. This means that the $N \times N$ matrix \mathbf{C}_x^{-1} has to be inverted, a very time and memory demanding task. Following [4] and [23], we approximate \mathbf{C}_x as the inverse of the diagonal of \mathbf{C}_x^{-1} . This inverse approximation is commented on in the open issues section IV-C.

The prior image model can also be made dependent on a global parameter, which models the general scale behavior of the prior model. Its estimation is a very hard problem which can be approached under some assumptions on the prior model, see [20].

C. Algorithm

The VB based blind deconvolution algorithm is presented in Alg. 1. However, as pointed out by Fergus *et al.* [8], directly applying it to estimate the blur may end up in the local minima, especially when the kernel support is large. To handle the large blur support problem, they suggest using a multiscale approach, namely building an image pyramid and then applying the BID method at each scale, which has proved to be very effective in BID problems. The rationale is that at the coarsest level, the blur is reduced significantly, so that it is easy to estimate a kernel from the downsampled image. At the next finer level, this kernel estimate is upsampled and can be used as a good initial guess for the single scale BID. Repeating this process until the finest level, we can obtain a better kernel estimate. After kernel estimation, we reconstruct the final sharp image using a non-blind deconvolution method (e.g., [50], [18], [51]).

The computation in Alg. 1 is dominated by the solution of Eqs. (25) and (27). Since the most time consuming part when solving these two equations is the 2-D convolution, the computational complexity

Algorithm 1 Single Scale Bayesian Blind Deconvolution Using Super Gaussian Priors

Require: Observation \mathbf{y} , noise level β , penalty $\rho(s)$, prior weight α .

- 1: Initialization $\mathbf{x} = \mathbf{y}$, $\mathbf{C}_x = 0$,
 - 2: **repeat**
 - 3: Initialize ξ
 - 4: **while** not converge **do**
 - 5: Update \mathbf{x} by solving the linear system 25
 - 6: Update ξ using Eq. (26)
 - 7: Approximate $\mathbf{C}_x(i, i)$ with $1/\mathbf{C}_x^{-1}(i, i)$
 - 8: **end while**
 - 9: Update \mathbf{h} by solving the quadratic programming problem in Eq.(27)
 - 10: **until** Convergence
-

is $O(NK)$ or $O(N \log(N))$, depending on the usage of spatial convolution or FFT, respectively. We should mention that the computational complexity increases to an extremely large number, $O(N^3)$, if \mathbf{C}_x^{-1} is inverted exactly. The number of iterations required for convergence depends on the image priors. For example, the use of log prior leads to faster convergence than the use of $\ell_{0.8}$ prior, because the log prior is more edge preserving and sparsity promoting than the $\ell_{0.8}$ prior. It is also shown in [38] that, for the same optimization method and parameter settings, the use of the normalized ℓ_1 prior [38] results in fewer iterations for convergence than the ℓ_1/ℓ_2 prior [36] in the kernel estimation step. Due to the use of the covariance matrix, the VB BID method is slower than the MAP method [36] [37] [38], and much slower than the edge prediction based methods [52] [53].

D. Sampling methods

Since the posterior distribution is not analytically available, sampling methods can be used to draw a large number of samples from it, and Monte Carlo integration techniques provide tools which allow performing inference on this dataset.

To simulate the posterior distribution Markov chains are used to develop different sampling methods. Perhaps the most widely used sampling method is Gibbs sampling described by the Geman and Geman in [54]. More recent methods are the Metropolis adjusted Langevin algorithms [55] and Hamiltonian Monte Carlo [56].

To better understand the sampling methods let us see an example of Gibbs sampling. If we can write down analytic expressions for the conditional distributions of all the unknowns we wish to estimate, given the others, we simply draw samples from each of the distributions in turn, conditioned on the most recently generated samples values for the other parameters. In our case we want to simulate $p(\mathbf{x}, \mathbf{h}, \Omega | \mathbf{y})$; the iterations would proceed as follows:

$$\begin{aligned}
&\text{First iteration: } \mathbf{x}^{(1)} \leftarrow p(\mathbf{x} | \mathbf{h}^{(0)}, \Omega^{(0)}, \mathbf{y}) \\
&\hspace{10em} \mathbf{h}^{(1)} \leftarrow p(\mathbf{h} | \mathbf{x}^{(1)}, \Omega^{(0)}, \mathbf{y}) \\
&\hspace{10em} \Omega^{(1)} \leftarrow p(\Omega | \mathbf{x}^{(1)}, \mathbf{h}^{(1)}, \mathbf{y}) \\
&\text{Second iteration: } \mathbf{x}^{(2)} \leftarrow p(\mathbf{x} | \mathbf{h}^{(1)}, \Omega^{(1)}, \mathbf{y}) \\
&\hspace{10em} \mathbf{h}^{(2)} \leftarrow p(\mathbf{h} | \mathbf{x}^{(2)}, \Omega^{(1)}, \mathbf{y}) \\
&\hspace{10em} \Omega^{(2)} \leftarrow p(\Omega | \mathbf{x}^{(2)}, \mathbf{h}^{(2)}, \mathbf{y}) \\
&\hspace{10em} \vdots
\end{aligned} \tag{30}$$

where the symbol \leftarrow means that the values are drawn from the distribution on the right. Once enough samples have been collected, point estimates and other statistics of the distribution may be found using Monte Carlo integration, for example the Minimum Squared Error estimator of the mean can be obtained as $\hat{\mathbf{x}} = \frac{1}{J} \sum_{j=1}^J \mathbf{x}^{(j)}$, where J is the number of drawn samples.

Due to the expensive computational cost (which is even worse when it is applied to high-dimensional data, such as images), the use of sampling methods in BID is not very extended. The works in this field are focused on developing more efficient algorithms. Ge *et al.* [57] or Kail *et al.* [58] propose modified versions of the Gibbs sampling, and Pereyra [35] uses the Langevin algorithm which uses convex analysis to simulate efficiently the distributions.

IV. OPEN ISSUES

Before presenting some BID examples we would like to comment here on some open problems, either on the Variational Bayesian BID (VBBID) or BID itself, that we believe will very likely be explored in the near future:

A. Image space versus filter space

VB methods can be formulated in either the image or the filter space. Levin *et al.* [4] state that the filter space has better performance for the MOG prior. Xu *et al.* [37] indicate that using the image space

formulation for latent image estimation and filter space formulation for kernel estimation is better than using the same spaces. In our opinion additional work is needed to establish the best spaces for image and kernel estimation. The image space appears to be less sensitive to noise since the noise is amplified in the filter space. Furthermore, the filter space is probably more computationally expensive than the image space. On one hand, utilizing the same number of iterations and L derivative filters, the total computation time in the filter space is about L times that of the image space. On the other hand, it is shown in Cho and Lee [52] that the kernel estimation in the image space requires more iterations to converge than in the filter space, since the symmetric matrix $\hat{\mathbf{X}}^T \hat{\mathbf{X}}$ of the quadratic program in Eq. (27), is not as diagonally dominant as $\sum_{\gamma} \hat{X}_{\gamma}^T \hat{X}_{\gamma}$, where $\hat{\mathbf{X}}$ and \hat{X}_{γ} denote the convolution matrices formed by $\hat{\mathbf{x}}$ and \hat{x}_{γ} , respectively. Based on the above two factors, it is conceivable that the image space is computationally less expensive than the filter space, when $L \geq 2$. Additionally, filter space methods have access to more ‘‘observations’’ to estimate the blur, although with an unrealistic independence assumption on them. The pros and cons of both approaches should be carefully analyzed.

B. Bottom-up approach

The bottom-up approach, first proposed by Babacan *et al.* [23], refers to formulating a weight update scheme $\phi(\nu) = \rho'(\nu)/\nu$ for the Gaussian prior approximation (without knowing explicitly the penalty function) provided that $\phi(\nu)$ is decreasing on $(0, +\infty)$. A crucial and very challenging question is how to choose a good penalty function ρ or ϕ for VB blind deconvolution.

Wipf and Zhang [59] state that the preferred distribution is not the one reflecting the accurate statistics of the latent image, but the one that is most likely to guide VB iterations to high quality global solutions by strongly differentiating between blurry and sharp images. This implies choosing a ϕ that strongly discriminates sharp and blurry images. To that end, ϕ should be strongly sparsity promoting and also very edge preserving, such as $\phi(\nu) = \nu^{-p}$ with $(p \geq 2)$.

We believe that a trade-off between preserving edges and promoting sparsity should be achieved when dealing with noisy images. If noise is high, a very edge preserving $\phi(\nu)$ cannot suppress it. Babacan *et al.* [23] also suggest a more general form $\phi(\nu) = (F\nu)^{-p}$, where F is a linear operator (e.g., nonlocal mean filter [60]). A variety of heuristics can easily be embedded through F to combat noise and increase robustness. Finally, we emphasize that given a ϕ , the value of α_{γ} should be chosen properly, as we will show in the experimental section.

C. Covariance approximation and general optimization issues

The covariance matrix \mathbf{C}_x plays a very important role not only in the image estimation step but also in the kernel estimation step. This matrix makes VB methods different from MAP methods. Intuitively, the introduction of \mathbf{C}_x in the estimation of the weights ξ makes their estimated values slightly smaller than when the covariance is not considered. As a result, the edges are better preserved. Besides, in the kernel estimation step, it provides an adaptive smoothness promoting regularization term which helps avoid the delta kernel estimates.

Unfortunately, due to the high computational cost, \mathbf{C}_x is approximated by the inverse of the diagonal of the weighted deconvolution matrix \mathbf{C}_x^{-1} . Since \mathbf{C}_x^{-1} is not diagonal, the diagonal approximation definitely introduces an error. The diagonal approximation is only reliable when β^{-1} and ξ are relatively large. If both β^{-1} and ξ are small, this approximation is not that reliable. Another alternative is the mean value approximation proposed by Babacan *et al.* [61], which replaces the weights ξ_γ with the average $\sum_{i=1}^N \xi_\gamma(i)/N$, so that \mathbf{C}_x is a circulant matrix associated with the kernel $\mathbf{h}_{\mathbf{C}_x} = \mathcal{F}^{-1}\Lambda_{\mathbf{h}}^{-1}$, where \mathcal{F} denotes the 2-D DFT and $\Lambda_{\mathbf{h}}$ is a column vector formed by the eigenvalues of \mathbf{C}_x^{-1} . $\mathbf{h}_{\mathbf{C}_x}$ has a large but finite support thanks to regularization and can be computed efficiently with the use of an FFT. This approximation takes the non-diagonal elements information into account, but the important information on the spatially variant weights is lost. The consequences of the use of the diagonal and mean value approximation remain an open question. Better but also feasible approximations to \mathbf{C}_x should also be explored.

Notice that the image estimation step involves solving a nonconvex problem when the penalty function is nonconvex, e.g., $\rho(s) = s^p/p$ ($0 < p < 1$). Assuming that the covariance term in ξ is ignored, it has been shown in [21] that the IRLS method which alternatively solves the linear equations in (25) and updates the weights by (26), definitely converges to a stationary point. Since the problem is nonconvex, the initial weights can make a difference in the final result, especially for the extremely nonconvex functions like log. A typical choice for the initial weights is the use of a large constant (e.g., 10^4 , see [4] [23]). It is conceivable that a ξ whose large values are located at the blur region may lead to a good stationary point, as the blur will be removed accurately. Since it is hard to know the blur region, finding such a good initial weights is not a trivial task in BID. Finally, we would like to mention that the linear equations (25) can be efficiently solved by ADMM [21] (Alternating Direction Method of Multipliers, see [62] for a comprehensive review), provided that the blur is spatially invariant.

Together with the IRLS method [21], other nonconvex optimization methods including variable splitting

and look-up-table based method [63], ℓ_1 -decomposition based method [64], and recently the smoothing trust region methods [65] [66] have also been applied to image deconvolution.

D. Spatially varying blur and other modeling problems

In this paper we have assumed that the blur is the same across the image. However, as shown in [67], even the camera optical system generates a considerable amount of spatially varying (SV) blur. In general, spatially varying degradation can be modeled as

$$y(s) = \sum_u h(s, s - u)x(u) + n(s), \quad (31)$$

where $y(s)$ is the value of the observed image at position s , $x(u)$ is the value of the unknown ideal image at position u , $h(s, s - u)$ is the blur affecting the image, that depends on each image pixel position, and $n(s)$ is the noise. When SV BID is addressed, some restrictions are applied to the way the blur varies in Eq. (31) in order to make the problem feasible. Such restrictions include the assumption that the blur is piecewise-invariant or piecewise smooth spatially varying, that is, the blur varies smoothly in the image, or that the blur is piecewise constant and location dependent, that is, different regions in the image have different blurs but the blur is spatially-invariant in each region. Another typical restriction is to assume that the type of the blur is known, for example, it is due to camera shake, or to consider images of a certain type, such as images with text [68] or star fields [69].

One approach to SV BID is to divide the image into non-overlapping patches where the blur is assumed to be stationary, apply a BID method on each patch independently, and merge the restored patches to obtain the final image. If the patches are not predefined, this approach casts the SV BID into a segmentation problem [70] in which the simpler case is to consider just two regions, a focused foreground and a out-of-focus background [71]. If the patches overlap and the blur varies smoothly on the image, the degradation model in Eq. (31) can be approximated as

$$y(s) = \sum_r \sum_u h_r(s - u)w_r(u)x_r(u) + n(s), \quad (32)$$

where $w_r(u) \geq 0$, $\sum_r w_r(u) = 1$, are weights allowing the smooth blending of the overlapping patches [72]. The advantage of this model is that it allows for an efficiently implementation using the Efficient Flow Filter (EFF) method [73] and also for different types of blur. On the other hand, its accuracy depends on the accuracy of the estimated kernel and may produce large errors if the kernels are not precisely estimated. In [74] the EFF method is extended to handle TV priors and a method to detect and replace erroneous blurs is proposed making it more robust. Another method to estimate smoothly

varying blurs, with applications to star field images, is proposed in [69]. The method estimates the blur at certain image positions and uses SVD to remove outliers and estimate a smooth PSF field from the individual PSFs.

If only camera-shake blur is considered, the Projective Motion Path approach [75] models the SV degradation as the average of multiple sharp images, each one corresponding to one of all possible camera poses, that is,

$$y(s) = \sum_i x(H_i u) + n(s), \quad (33)$$

where H_i are homographies, that is, combinations of rotations and translations, that project the sharp image given a camera orientation. The homographies can be obtained from auxiliary sensors attached to the camera, such as gyroscopes (see [76] and section IV-E), and high speed low resolution cameras [77], or they are estimated with the image [78].

A similar approximation is considered in [79] where the SV degradation is modeled as a weighted sum of sharp images obtained at all possible camera poses, that is,

$$\mathbf{y} = \sum_i w(i) \mathbf{C}_i \mathbf{x} + \mathbf{n}, \quad (34)$$

where \mathbf{C}_i is the matrix that applies the homography H_i to the image \mathbf{x} and $w(i)$ weights the i -th projection depending on the time spent by the camera at the i -th pose during the capture time. In [80], VB is used to estimate both the image \mathbf{x} and the weights $w(\cdot)$. The drawback of this approach is that is resource demanding since it has to compute and store all the projections. To alleviate this problem, [81] proposes an iterative method that, at each iteration, restricts the solution space to a small set of camera poses which the camera motion trajectory is most likely to belong to.

Despite all these advances, more research is still needed to solve the general SV BID problem as described by Eq. (31).

Even without mentioning the spatially variant nature of the blur, the linear model in Eq. (1), utilized by most BID methods, is not a realistic one for real-life images. Common violations include the presence of defective sensor pixels, saturated pixels [82], a nonlinear camera response curve [83], or non additive white Gaussian noise [84], [85] which, if not properly handled, may generate ringing artifacts when restoring the image even if the blur is accurately estimated [86]. We believe that developing methods that explicitly handle such model violations will improve the applicability of BID to real problems.

These modeling problems are alleviated with the use of more than one images. Considering color and, in general, multichannel images, remedies, to a great extent, the ill-posed nature of blind deconvolution [87]. Using image pairs with different properties facilitates blur estimation and helps handle saturated

pixels and other camera imperfections. For instance, in [88] a near-infrared image is captured together with a visible blurred image and, in [89], [90] a low exposure sharp but noisy image is used to improve the restoration results. If video is available, techniques can take into account the motion between frames [91]–[93] to tackle the deblurring problem. Of interest is also the approach in [94] where a single high-quality image is obtained from a sequence of images distorted by atmospheric turbulence. Having several images also allows blind image deconvolution to be addressed simultaneously with other problems, such as, super-resolution (see, for instance, [95] or [96]) or high dynamic range (HDR) imaging [97].

Finally, to conclude this section on modeling, we would like to mention the need to model what a good restoration is. We believe that more BID software applications will be developed if the quality of a restored image can be assessed, without human intervention, before presenting it to the user.

E. Deconvolution in mobile devices

The ubiquity of mobile devices, such as smartphones and tablets, and the not-so-high quality of their cameras make the restoration of images taken with those devices a succulent market. Running the deconvolution process on mobile devices is, nevertheless, difficult given their limited computational power. Some commercial applications that claim to remove blur from images are available for the different platforms (see *DeblurIt Pro* or *Photo Fix de Blur* for Android or *Photo Doctor* for iOS). However they seem to deal only with out-of-focus blur with a manually selected radius and implement simple deconvolution algorithms.

Smartphones and tablets are more than simple cameras. They usually have other built-in sensors to capture the device position and trajectory and the capability of processing images. Hence, some methods are being proposed to perform deconvolution on those devices. In an effort to take advantage of the sensors present on the mobile phones, Šindelář and Šroubek [98] used the information provided by the gyroscope to keep track of the device motion while taking the picture and, hence, obtain an estimation of the blur by rendering the camera trajectory on the image plane. This blur estimate is used to deconvolve the image by a simple Wiener filter. An extension considering spatially variant blur and rolling shutter compensation is presented in [99].

A similar approach was used in [100] where the blur is obtained from a combination of the kernel estimated from the fusion of gyroscope, magnetometer and accelerometer measurements and a Gaussian kernel with small variance to take into account the out-of-focus blur due to the motion of the camera from the finger movement on pressing on the screen. Additionally, to minimize artifacts on faces, a face detection algorithm is applied to the image and an SVM classifier is trained and used to select between

deblurring followed by denoising or sharpening the image, depending on the face characteristics.

Using a developer tablet modified by attaching a USB connected external gyroscope, a multi-image deconvolution which captures and combines multiple frames in order to make deblurring more robust and tractable is proposed in [101]. Blur is obtained from the gyroscope data and multi-image deconvolution is performed by minimizing

$$\sum_{i=1}^n \|\mathbf{y}_i - \mathbf{H}_i \mathbf{x}\|^2 + \lambda \|\Delta \mathbf{x}\|^p, \quad (35)$$

where λ is a regularization parameter and Δ is the gradient operator. The authors conclude that this deconvolution procedure outperforms, in most situations, the align-and-average strategy, that is, averaging multiple noisy images captured using a short exposure time, and hence blur-free, aligned using the gyroscope data. The optimization problem in Eq. (35) was first utilized in the work by Katsaggelos [102].

F. Implementation issues

Since BID methods need to estimate both image and blur, they usually take a significant amount of time. Apart from developing mathematically efficient methods to compute blur and image estimates, efficient implementations are needed to speed up the algorithms. Nowadays, most computers are equipped with graphical processing units (GPUs) that have several GFLOPS of computing power. Massive computing using these GPU or hybrid CPU+GPU computing can dramatically improve the speed of the algorithms. Most of the deconvolution implementations using GPUs are based on their capability to accelerate an FFT, with the CUDA framework and the CUFFT library [103] being the most popular implementation.

Some BID methods have been implemented using GPUs with great success as proved in [104] where the time needed to blindly deconvolve an 8 MPixel image using the method in [105] is reduced from 55.6s to 13.8s. The EFF spatially variant blind image deconvolution method in [73] runs about ten times faster using GPU than using only CPU.

Several efforts have also been carried out to use GPU computing in non-blind image deconvolution. For instance, Zhang *et al.* [106] performed real-time high definition 720p video processing with a Wiener filter using an NVIDIA GeForce GTX 460 GPU and Holder *et al.* [107] obtained an acceleration of 1:5 compared to CPU of the Richardson-Lucy algorithm on an NVIDIA Geforce GT640M. The GPU implementation of the non-blind Krishnan-Fergus [63] algorithm presented in [108] runs at 15 frames per second on 710×470 pixels color images on an NVIDIA GeForce GTX 260.

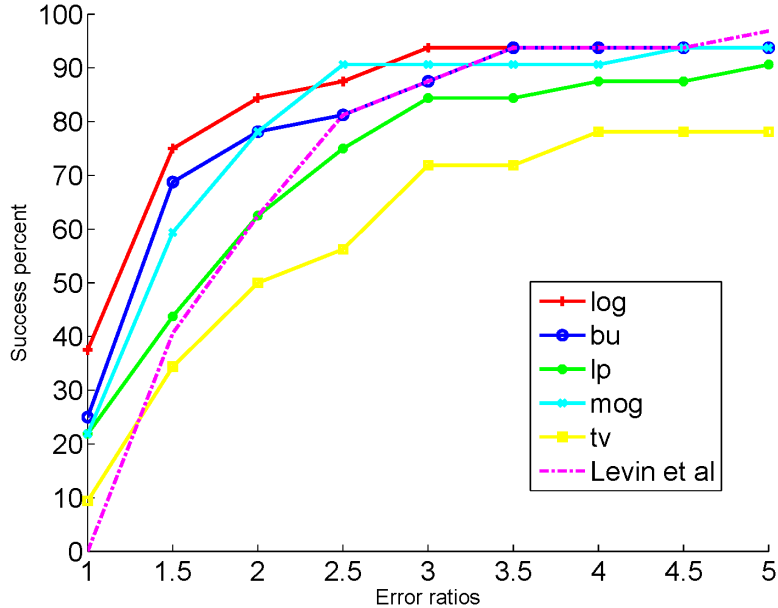


Fig. 3. Cumulative histograms of the error ratios across the dataset [39]

V. EXPERIMENTS

We test the performance of 5 image priors, including log, $\ell_{0.8}$, MOG, TV and bu3, where the parameters for MOG are borrowed from Levin *et al.* [4] and bu3 is referred to the bottom-up approach [20] [23] with $\phi(\nu) = \nu^{-3}$ (corresponding to $\rho(x) = -x^{-1}$). We choose the widely used dataset [39] which consists of 32 images generated by 4 groundtruth images with 8 motion blurs. For the priors log, $\ell_{0.8}$, MOG, TV and bu3, we set α_γ to 1, 10, 1, 20 and 0.1 respectively. After obtaining the kernels, we use the non-blind deconvolution method [18] with the same parameters used in [4] to reconstruct the final image.

Fig. 3 presents the success percent of 6 methods (ours with 5 different priors and Levin *et al.* [4]) in the sense of error ratio metric (ratio between sum of squared difference errors of the restoration with the estimated kernel and the restoration with the groundtruth kernel, see [39] for more details). As we can see, the log prior has the best performance, with over 80% good restorations (error ratio ≤ 2) and 90% successful restorations (error ratio ≤ 3 is regarded as successful restoration, according to Levin *et al.* [4]), followed by bu and MOG. $\ell_{0.8}$ and TV also have good performance with about 80% and 70% successful restorations. It should be emphasized that, a suitable α_γ is crucial for the different priors to work well. Fig. 4 shows some selected results for visual evaluation.

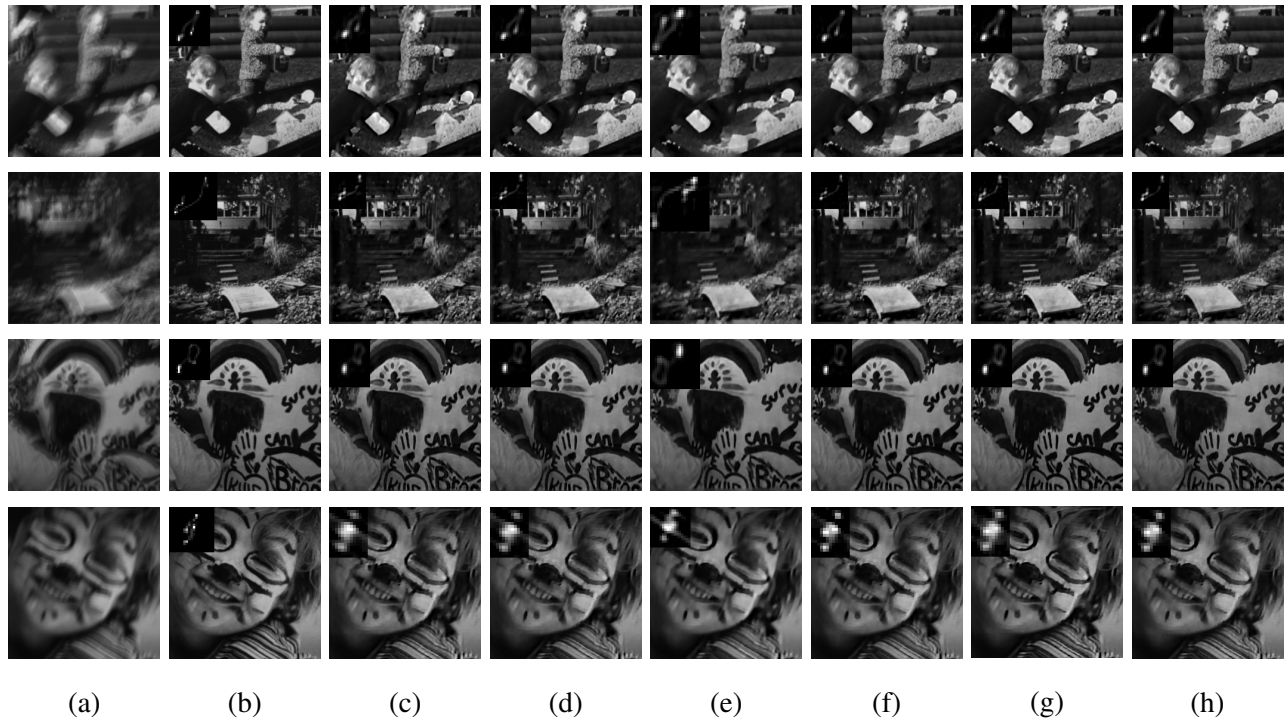


Fig. 4. Selected results on dataset [39] for visual comparison. (a) Blurred. (b) Groundtruth. (c) TV. (d) $\ell_{0,s}$. (e) Levin *et al.* [4]. (f) MOG. (g) bu3. (h) log.

REFERENCES

- [1] S. Horaczek, “How many photos are uploaded to the internet every minute?” May 2013. [Online]. Available: <http://www.popphoto.com/news/2013/05/how-many-photos-are-uploaded-to-internet-every-minute>
- [2] J. Hadamard, *Lectures on Cauchy’s Problem in Linear Partial Differential Equations*. Yale University Press, New Haven CT, 1923.
- [3] T. E. Bishop, S. D. Babacan, B. Amizic, A. K. Katsaggelos, T. Chan, and R. Molina, *Blind image deconvolution: problem formulation and existing approaches*. CRC press, 2007.
- [4] A. Levin, Y. Weiss, F. Durand, and W. T. Freeman, “Efficient marginal likelihood optimization in blind deconvolution,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 2657–2664.
- [5] J. Miskin and D. MacKay, *Ensemble learning for blind image separation and deconvolution*. Springer, 2000.
- [6] C. L. Likas and N. P. Galatsanos, “A variational approach for Bayesian blind image deconvolution,” *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2222–2233, Aug. 2004.
- [7] R. Molina, J. Mateos, and A. Katsaggelos, “Blind deconvolution using a variational approach to parameter, image, and blur estimation,” *IEEE Trans. Image Process.*, vol. 15, no. 12, pp. 3715–3727, Dec 2006.
- [8] R. Fergus, B. Singh, A. Hertzmann, S. T. Roweis, and W. T. Freeman, “Removing camera shake from a single photograph,” *ACM Transactions on Graphics (TOG) - Proceedings of ACM SIGGRAPH 2006*, vol. 25, no. 3, pp. 787–794, 2006.
- [9] R. Molina, A. Katsaggelos, J. Abad, and J. Mateos, “A Bayesian approach to blind deconvolution based on Dirichlet

- distributions,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, 1997, pp. 2809–2812.
- [10] L. Rudin, S. Osher, and E. Fatemi, “Nonlinear total variation based noise removal algorithms,” *Physica D*, vol. 60, pp. 259–268, 1992.
- [11] T. F. Chan and C.-K. Wong, “Total variation blind deconvolution,” *IEEE Trans. Image Process.*, vol. 7, no. 3, pp. 370–375, 1998.
- [12] S. Babacan, R. Molina, and A. Katsaggelos, “Variational Bayesian blind deconvolution using a total variation prior,” *IEEE Trans. Image Process.*, vol. 18, no. 1, pp. 12–26, Jan. 2009.
- [13] B. Amizic, R. Molina, and A. K. Katsaggelos, “Sparse Bayesian blind image deconvolution with parameter estimation,” *Eurasip Journal on Image and Video Processing*, vol. 2012, no. 1, Nov. 2012.
- [14] D. Perrone and P. Favaro, “Total variation blind deconvolution: The devil is in the details,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 2909–2916.
- [15] —, “A clearer picture of blind deconvolution,” *arXiv:1412.0251 [cs]*, Nov. 2014, arXiv: 1412.0251. [Online]. Available: <http://arxiv.org/abs/1412.0251>
- [16] G. Chantas, N. P. Galatsanos, A. Likas, and M. Saunders, “Variational Bayesian image restoration based on a product of t-distributions image prior,” *IEEE Trans. Image Process.*, vol. 17, no. 10, pp. 1795–1805, Oct. 2008.
- [17] G. Chantas, N. Galatsanos, R. Molina, and A. Katsaggelos, “Variational Bayesian image restoration with a product of spatially weighted total variation image priors,” *IEEE Trans. Image Process.*, vol. 19, no. 2, pp. 351–362, Feb. 2010.
- [18] A. Levin, R. Fergus, F. Durand, and W. T. Freeman, “Image and depth from a conventional camera with a coded aperture,” *ACM Transactions on Graphics (TOG) - Proceedings of ACM SIGGRAPH 2007*, vol. 26, no. 3, p. 70, 2007.
- [19] Q. Shan, J. Jia, and A. Agarwala, “High-quality motion deblurring from a single image,” *ACM Transactions on Graphics (TOG) - Proceedings of ACM SIGGRAPH 2008*, vol. 27, no. 3, p. 73, 2008.
- [20] M. Vega, R. Molina, and A. Katsaggelos, “Parameter estimation in Bayesian blind deconvolution with super Gaussian image priors,” in *European Signal Processing Conference (EUSIPCO)*, 2014, pp. 1632–1636.
- [21] X. Zhou, R. Molina, F. Zhou, and A. Katsaggelos, “Fast iteratively reweighted least squares for l_p regularized image deconvolution and reconstruction,” in *IEEE International Conference on Image Processing (ICIP)*, Oct. 2014, pp. 1783–1787.
- [22] J. A. Palmer, K. Kreutz-Delgado, and S. Makeig, “Strong sub- and super-Gaussianity,” in *Latent Variable Analysis and Signal Separation*, ser. Lecture Notes in Computer Science, V. Vigneron, V. Zarzoso, E. Moreau, R. Gribonval, and E. Vincent, Eds. Springer Berlin Heidelberg, 2010, vol. 6365, pp. 303–310.
- [23] S. D. Babacan, R. Molina, M. N. Do, and A. K. Katsaggelos, “Bayesian blind deconvolution with general sparse image priors,” in *European Conference on Computer Vision (ECCV)*, 2012, pp. 341–355.
- [24] H. Zhang and D. Wifp, “Non-uniform camera shake removal using a spatially-adaptive sparse penalty,” in *Advances in Neural Information Processing Systems*, 2013, pp. 1556–1564.
- [25] A. Mohammad-Djafari, “Bayesian blind deconvolution of images comparing JMAP, EM and BVA with a Student-t a priori model,” in *International Workshops on Electrical Computer Engineering Subfields*, 2014, pp. 98–103.
- [26] D. F. Andrews and C. L. Mallows, “Scale mixtures of normal distributions,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 36, no. 1, pp. 99–102, 1974.
- [27] K. P. Murphy, *Machine learning: a probabilistic perspective*, ser. Adaptive computation and machine learning series. Cambridge (Mass.): MIT Press, 2012.

- [28] M. Rattray, O. Stegle, K. Sharp, and J. Winn, “Inference algorithms and learning theory for Bayesian sparse factor analysis,” in *Journal of Physics: Conference Series*, vol. 197, 2009.
- [29] M. Lavielle, “Bayesian deconvolution of Bernoulli-Gaussian processes,” *Signal Processing*, vol. 33, no. 1, pp. 67–79, Jul. 1993.
- [30] J. Cai, H. Ji, C. Liu, and Z. Shen, “Blind motion deblurring from a single image using sparse approximation,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 104–111.
- [31] S. Oh and G. Kim, “Robust estimation of motion blur kernel using a piecewise-linear model,” *IEEE Trans. Image Process.*, vol. 23, no. 3, pp. 1394–1407, Mar. 2014.
- [32] A. Goldstein and R. Fattal, “Blur-kernel estimation from spectral irregularities,” in *European Conference on Computer Vision (ECCV)*, 2012, pp. 622–635.
- [33] G. Liu, S. Chang, and Y. Ma, “Blind image deblurring using spectral properties of convolution operators,” *IEEE Trans. Image Process.*, vol. 23, no. 12, pp. 5047–5056, Dec. 2014.
- [34] A. Levin, Y. Weiss, F. Durand, and W. Freeman, “Understanding blind deconvolution algorithms,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 12, pp. 2354–2367, Dec. 2011.
- [35] M. Pereyra, “Proximal Markov chain Monte Carlo algorithms,” *arXiv:1306.0187 [stat]*, Jun. 2013, arXiv: 1306.0187. [Online]. Available: <http://arxiv.org/abs/1306.0187>
- [36] D. Krishnan, T. Tay, and R. Fergus, “Blind deconvolution using a normalized sparsity measure,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 233–240.
- [37] L. Xu, S. Zheng, and J. Jia, “Unnatural L0 sparse representation for natural image deblurring,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 1107–1114.
- [38] X. Zhou, F. Zhou, and X. Bai, “Blind deconvolution using a nondimensional Gaussianity measure,” in *IEEE International Conference on Image Processing (ICIP)*, 2013, pp. 877–881.
- [39] A. Levin, Y. Weiss, F. Durand, and W. T. Freeman, “Understanding and evaluating blind deconvolution algorithms,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 1964–1971.
- [40] C. Wang, Y. Yue, F. Dong, Y. Tao, X. Ma, G. Clapworthy, and X. Ye, “Enhancing Bayesian estimators for removing camera shake,” *Computer Graphics Forum*, vol. 32, no. 6, pp. 113–125, Sep. 2013.
- [41] W. J. Fitzgerald, “The Bayesian approach to signal modelling,” in *IEE Colloquium on Non-Linear Signal and Image Processing (Ref. No. 1998/284)*, 1998, pp. 9/1–9/5.
- [42] Z. Chen, S. D. Babacan, R. Molina, and A. K. Katsaggelos, “Variational Bayesian methods for multimedia problems,” *IEEE Transaction on Multimedia*, vol. 16, no. 4, pp. 1000–1017, 2014.
- [43] J. J. K. O Ruanaidh and W. J. Fitzgerald, *Numerical Bayesian Methods Applied to Signal Processing*. Springer Verlag, 1996.
- [44] W. J. Fitzgerald, “Markov chain Monte Carlo methods with applications to signal processing,” *Signal Processing*, vol. 81, no. 1, pp. 3–18, 2001.
- [45] S. Gulam-Razul, W. J. Fitzgerald, and C. Andrieu, “Bayesian deconvolution in nuclear spectroscopy using RJMCMC,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, 2002, pp. 1309–1312.
- [46] G. Parisi, *Statistical Field Theory*. Addison-Wesley, 1988.
- [47] C. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.
- [48] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.
- [49] R. T. Rockafellar, *Convex Analysis*. Princeton University Press, 1996.

- [50] X. Zhou, F. Zhou, X. Bai, and B. Xue, "A boundary condition based deconvolution framework for image deblurring," *Journal of Computational and Applied Mathematics*, vol. 261, pp. 14–29, 2014.
- [51] D. Zoran and Y. Weiss, "From learning models of natural image patches to whole image restoration," in *IEEE International Conference on Computer Vision (ICCV)*, 2011, pp. 479–486.
- [52] S. Cho and S. Lee, "Fast motion deblurring," *ACM Transactions on Graphics (TOG) - Proceedings of ACM SIGGRAPH Asia 2009*, vol. 28, no. 5, p. 145, 2009.
- [53] L. Xu and J. Y. Jia, "Two-phase kernel estimation for robust motion deblurring," in *European Conference on Computer Vision (ECCV)*, 2010, pp. 157–170.
- [54] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-6, no. 6, pp. 721–741, Nov. 1984.
- [55] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*. Springer-Verlag, 2004.
- [56] R. M. Neal, "MCMC using Hamiltonian dynamics," *arXiv:1206.1901 [physics, stat]*, Jun. 2012, arXiv: 1206.1901. [Online]. Available: <http://arxiv.org/abs/1206.1901>
- [57] D. Ge, J. Idier, and E. Le Carpentier, "Enhanced sampling schemes for MCMC based blind Bernoulli-Gaussian deconvolution," *Signal Processing*, vol. 91, no. 4, pp. 759–772, 2011.
- [58] G. Kail, J. Y. Tourneret, F. Hlawatsch, and N. Dobigeon, "Blind deconvolution of sparse pulse sequences under a minimum distance constraint: A partially collapsed Gibbs sampler method," *IEEE Transactions on Signal Processing*, vol. 60, no. 6, pp. 2727–2743, Jun. 2012.
- [59] D. Wipf and H. Zhang, "Revisiting Bayesian blind deconvolution," *arXiv arXiv:1305.2362.*, 2013. [Online]. Available: <http://arxiv.org/abs/1305.2362>
- [60] A. Buades, B. Coll, and J. Morel, "A non-local algorithm for image denoising," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, 2005, pp. 60–65.
- [61] S. D. Babacan, R. Molina, and A. K. Katsaggelos, "Parameter estimation in TV image restoration using variational distribution approximation," *IEEE Trans. Image Process.*, vol. 17, no. 3, pp. 326–339, 2008.
- [62] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, pp. 1–122, 2011.
- [63] D. Krishnan and R. Fergus, "Fast image deconvolution using hyper-Laplacian priors," in *Advances in Neural Information Processing Systems (NIPS)*, 2009, pp. 1033–1041.
- [64] M. Nikolova, M. Ng, and C. P. Tam, "Fast nonconvex nonsmooth minimization methods for image restoration and reconstruction," *IEEE Trans. Image Process.*, vol. 19, no. 12, pp. 3073–3088, 2010.
- [65] M. Hintermüller and T. Wu, "Nonconvex TV^q -models in image restoration: analysis and a trust-region regularization based superlinearly convergent solver," *SIAM Journal on Imaging Sciences*, vol. 6, no. 3, pp. 1385–1415, 2013.
- [66] X. Chen, L. Niu, and Y. Yuan, "Optimality conditions and a smoothing trust region Newton method for nonLipschitz optimization," *SIAM Journal on Optimization*, vol. 23, no. 3, pp. 1528–1552, 2013.
- [67] E. Kee, S. Paris, S. Chen, and J. Wang, "Modeling and removing spatially-varying optical blur," in *IEEE International Conference on Computational Photography (ICCP)*, 2011, pp. 1–8.
- [68] X. Cao, W. Ren, W. Zuo, X. Guo, and H. Foroosh, "Scene text deblurring using text-specific multiscale dictionaries," *IEEE Trans. Image Process.*, vol. 24, no. 4, pp. 1302–1314, Apr. 2015.
- [69] D. Miraut, J. Ball, and J. Portilla, "Efficient shift-variant image restoration using deformable filtering (Part II): PSF field estimation," *EURASIP Journal on Advances in Signal Processing*, vol. 2012, no. 1, pp. 1–19, Aug. 2012.

- [70] F. Couzinie-Devy, J. Sun, K. Alahari, and J. Ponce, "Learning to estimate and remove non-uniform image blur," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 1075–1082.
- [71] S. Chan and T. Nguyen, "Single image spatially variant out-of-focus blur removal," in *IEEE International Conference on Image Processing (ICIP)*, 2011, pp. 677–680.
- [72] S. Harmeling, M. Hirsch, and B. Schölkopf, "Space-variant single-image blind deconvolution for removing camera shake," in *Advances in Neural Information Processing Systems 23*, 2010, pp. 829–837.
- [73] M. Hirsch, C. Schuler, S. Harmeling, and B. Schölkopf, "Fast removal of non-uniform camera shake," in *Proceedings of the IEEE International Conference on Computer Vision*, 2011, pp. 463–470.
- [74] X. Yu, F. Xu, S. Zhang, and L. Zhang, "Efficient patch-wise non-uniform deblurring for a single image," *IEEE Transactions on Multimedia*, vol. 16, no. 6, pp. 1510–1524, 2014.
- [75] Y.-W. Tai, P. Tan, and M. Brown, "Richardson-Lucy deblurring for scenes under a projective motion path," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1603–1618, 2011.
- [76] N. Joshi, S. B. Kang, C. L. Zitnick, and R. Szeliski, "Image deblurring using inertial measurement sensors," *ACM Transactions on Graphics*, vol. 29, no. 4, p. 1, 2010.
- [77] Y.-W. Tai, H. Du, M. Brown, and S. Lin, "Correction of spatially varying image and video motion blur using a hybrid camera," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 6, pp. 1012–1028, 2010.
- [78] X. Zhang and F. Sun, "Blind nonuniform deblur under projection motion path," *Journal of Electronic Imaging*, vol. 22, no. 3, p. 033034, 2013.
- [79] O. Whyte, J. Sivic, A. Zisserman, and J. Ponce, "Non-uniform deblurring for shaken images," *International Journal of Computer Vision*, vol. 98, no. 2, pp. 168–186, 2012.
- [80] —, "Non-uniform deblurring for shaken images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 491–498.
- [81] Z. Hu and M. hsuan Yang, "Fast non-uniform deblurring using constrained camera pose subspace," in *Proceedings of the British Machine Vision Conference*, 2012, pp. 136.1–136.11.
- [82] O. Whyte, J. Sivic, and A. Zisserman, "Deblurring shaken and partially saturated images," *International Journal of Computer Vision*, vol. 110, no. 2, pp. 185–201, 2014.
- [83] S. Kim, Y.-W. Tai, S. J. Kim, M. Brown, and Y. Matsushita, "Nonlinear camera response functions and image deblurring," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2012, pp. 25–32.
- [84] M. A. T. Figueiredo and J. M. Bioucas-Dias, "Restoration of Poissonian images using alternating direction optimization," *IEEE Trans. Image Process.*, vol. 19, no. 12, pp. 3133–3145, Dec. 2010.
- [85] P. Rodriguez, R. Rojas, and B. Wohlberg, "Mixed Gaussian-impulse noise image restoration via total variation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 1077–1080.
- [86] S. Cho, J. Wang, and S. Lee, "Handling outliers in non-blind image deconvolution," in *IEEE International Conference on Computer Vision (ICCV)*, Nov. 2011, pp. 495–502.
- [87] F. Šroubek and P. Milanfar, "Robust multichannel blind deconvolution via fast alternating minimization," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1687–1700, Apr. 2012.
- [88] W. Li, J. Zhang, and Q.-H. Dai, "Robust blind motion deblurring using near-infrared flash image," *Journal of Visual Communication and Image Representation*, vol. 24, no. 8, pp. 1394–1413, 2013.
- [89] S. D. Babacan, J. Wang, R. Molina, and A. K. Katsaggelos, "Bayesian blind deconvolution from differently exposed image pairs," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2874–2888, Nov. 2010.

- [90] M. Tallón, J. Mateos, S. D. Babacan, R. Molina, and A. K. Katsaggelos, “Space-variant blur deconvolution and denoising in the dual exposure problem,” *Information Fusion*, vol. 14, no. 4, pp. 396–409, 2013.
- [91] J. Brailean and A. Katsaggelos, “Simultaneous recursive displacement estimation and restoration of noisy-blurred image sequences,” *IEEE Trans. Image Process.*, vol. 4, no. 9, pp. 1236–1251, Sep 1995.
- [92] X. Deng, Y. Shen, M. Song, D. Tao, J. Bu, and C. Chen, “Video-based non-uniform object motion blur estimation and deblurring,” *Neurocomputing*, vol. 86, pp. 170–178, 2012.
- [93] Y. Xu, X. Hu, and S. Peng, “Blind motion deblurring using optical flow,” *Optik*, vol. 126, no. 1, p. 87?4, Jan. 2015.
- [94] X. Zhu and P. Milanfar, “Removing atmospheric turbulence via space-invariant deconvolution,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 157–170, Jan. 2013.
- [95] W.-Z. Shao and M. Elad, “Simple, accurate, and robust nonparametric blind super-resolution,” *ArXiv e-prints*, Mar. 2015. [Online]. Available: <http://arxiv.org/abs/1503.03187>
- [96] H. Zhang and L. Carin, “Multi-shot imaging: Joint alignment, deblurring, and resolution-enhancement,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2925–2932.
- [97] C. Vijay, C. Paramanand, A. Rajagopalan, and R. Chellappa, “Non-uniform deblurring in HDR image reconstruction,” *IEEE Trans. Image Process.*, vol. 22, no. 10, pp. 3739–3750, Oct. 2013.
- [98] O. Šindelář and F. Šroubek, “Image deblurring in smartphone devices using built-in inertial measurement sensors,” *Journal of Electronic Imaging*, vol. 22, no. 1, pp. 011 003–011 003, 2013.
- [99] O. Šindelář, F. Šroubek, and P. Milanfar, “A smartphone application for removing handshake blur and compensating rolling shutter,” in *IEEE International Conference on Image Processing (ICIP)*, 2014.
- [100] W. Jiang, D. Zhang, and H. Yu, “Sensor-assisted image deblurring of consumer photos on smartphones,” in *2014 IEEE International Conference on Multimedia and Expo (ICME)*, Jul. 2014, pp. 1–6.
- [101] S. H. Park and M. Levoy, “Gyro-based multi-image deconvolution for removing handshake blur,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 3366–3373.
- [102] A. K. Katsaggelos, “A multiple input image restoration approach,” *Journal of Visual Communication and Image Representation*, vol. 1, pp. 93–103, 1990.
- [103] NVIDIA Corporation, “NVIDIA CUDA fast Fourier transform,” 2015. [Online]. Available: <http://docs.nvidia.com/cuda/cufft/index.html>
- [104] T. Mazanec, A. Hermanek, and J. Kamenicky, “Blind image deconvolution algorithm on NVIDIA CUDA platform,” in *2010 IEEE 13th International Symposium on Design and Diagnostics of Electronic Circuits and Systems (DDECS)*, April 2010, pp. 125–126.
- [105] F. Šroubek and J. Flusser, “Multichannel blind deconvolution of spatially misaligned images,” *IEEE Trans. Image Process.*, vol. 14, no. 7, pp. 874–883, Jul. 2005.
- [106] Y. Zhang, J. He, and J. Yuan, “A video deblurring optimization algorithm based on motion detection,” in *The 3rd International Conference on Multimedia Technology (ICMT-13)*, 2013.
- [107] S. Holder and G. Lin, “Acceleration of image restoration algorithms for dynamic measurements in coordinate metrology by using OpenCV GPU framework,” 2014.
- [108] J. Klosowski and S. Krishnan, “Real-time image deconvolution on the GPU,” in *SPIE Conference: Parallel Processing for Imaging Applications*, Jan. 2011.

Chapter 4

Multispectral Image Classification (I). Image Processing for Classification

4.1 Interactive Classification Oriented Superresolution of Multispectral Images

- **P. Ruiz**, J.V. Talens, J. Mateos, R. Molina, and A.K. Katsaggelos, “Interactive Classification Oriented Superresolution of Multispectral Images” in *7th International Workshop Data Analysis in Astronomy (DAA2011)*, edited by Livio Scarsi and Vito Di Gesù, 77-85, Erice (Italy), April 2011.

– Status: Published

INTERACTIVE CLASSIFICATION ORIENTED SUPERRESOLUTION OF MULTISPECTRAL IMAGES

P. Ruiz^{1*}, *J. V. Talents*², *J. Mateos*¹, *Rafael Molina*¹ and *Aggelos K. Katsaggelos*³

¹ Dpto. de Ciencia de la Computación e I.A. Universidad de Granada.

² Image Processing Laboratory (IPL). Universitat de Valencia

³ Dpt. of Electrical Engineering and Computer Science. Northwestern University.

*e-mail:mataran@decsai.ugr.es

ABSTRACT

Classification techniques are routinely utilized on satellite images. Pansharpening techniques can be used to provide super resolved multispectral images that can improve the performance of classification methods. So far, these pansharpening methods have been explored only as a preprocessing step. In this work we address the problem of adaptively modifying the pansharpening method in order to improve the precision and recall figures of merit of the classification of a given class without significantly deteriorating the performance of the classifier over the other classes. The validity of the proposed technique is demonstrated using a real Quickbird image.

Index Terms—Pansharpening, super-resolution, classification, LDA, SVM.

1. INTRODUCTION

Satellite images are of great interest due to the numerous applications they can be utilized. Drawing maps, delimitation of parcels, studies on hydrology, forest or agriculture are just a few examples where these images are used.

Due to physical and technological constraints, satellites usually have sensors that capture two types of images. One sensor captures a multispectral (MS) image composed of several spectral bands with low spatial resolution (LR). The other sensor captures a high spatial resolution (HR) image, named panchromatic (PAN) image, with a low spectral resolution. While the first image allows to distinguish features spectrally but not spatially, the second allows to distinguish features spatially but not spectrally.

Pansharpening is an image fusion approach that combines the LR MS and PAN images to obtain an image with the spectral resolution of the MS image and the spatial resolution of the PAN image. Many techniques have been proposed

in the literature to carry out the pansharpening procedure (see Ref. [1] for a complete review of pansharpening methods).

Many satellite image applications involve the classification of pixels in an image into a number of classes. In supervised classification, starting from a small set of samples previously labeled by the user, classification is carried out automatically by the classifiers. Bruzzone *et al.* [2] showed that the use of pansharpening methods that do not introduce significant spectral distortion helps the classifier to obtain higher accuracy, especially for pixels at the borders of objects.

While in the past pansharpening techniques have only been used as a preprocessing step, in this work we address the problem of adaptively modifying the pansharpening method in order to improve the precision and recall figures of merit of the classification of a given class without deteriorating the performance of the classifier over the other classes.

The rest of paper is organized as follows: In section 2 we describe the pansharpening technique we use. The used classifiers are briefly explained in section 3. The proposed method to estimate the pansharpening parameters to improve the performance of the classifier on a given class is described in section 4. Section 5 presents experimental results on real data. Finally, section 6 concludes the paper.

2. PANSHARPENING ALGORITHM

In this paper we use the pansharpening method proposed by Amro *et al.* [3] and the parameter estimation procedure described in Ref. [1]. This method makes use of the non-subsampled contourlet transform [4] (NSCT) to decompose the details of the PAN and each band of the MS image into different scales and different directions. Then, the hierarchical Bayesian framework is used to model those observations and their relations with the original high resolution multispectral image and Bayesian inference is applied to estimate the HR MS image and the model parameters. Let us now explain in detail the used pansharpening method.

The used contourlet based pansharpening algorithm takes as input the PAN image, x , of size $p = m \times n$, and the ob-

This work has been supported in part by the Comisión Nacional de Ciencia y Tecnología under contract TIN2010-15137, CEI BioTic at the University of Granada, and the Department of Energy grant DE-NA0000457.

served LR MS image, Y , with B bands, $Y_b, b = 1, \dots, B$, each of size $P = M \times N$ pixels with $M < m$ and $N < n$. Initially, each band of the LR MS image Y is upsampled to the size of the PAN image by bicubic interpolation. We will denote by s_b each band b of the $p = m \times n$ upsampled image.

Then, using the NTSC transform we can write the PAN and the upsampled MS images as:

$$x = x^r + \sum_{l=1}^L \sum_{d=1}^D x^{ld}, \quad s_b = s_b^r + \sum_{l=1}^L \sum_{d=1}^D s_b^{ld}, \quad b = 1, \dots, B \quad (1)$$

where the superscript r denotes the residual (low pass filtered version) NSCT coefficient band and the superscript ld refers to the detail bands, with $l = 1, \dots, L$, representing the scale and $d = 1, \dots, D$, representing the direction for each coefficient band. The pansharpening goal is to estimate the HR MS image coefficients y_b^{ld} from the observed x^{ld} and s_b^{ld} coefficients. Finally, each band of the pansharpened HR MS image will be obtained by the inverse NSCT from the corresponding residual band of the upsampled MS image s_b^r and the estimated detail bands y_b^{ld} .

We will model the coefficient bands using the hierarchical Bayesian framework. This framework has two stages. In the first stage, knowledge about the structural form of the noise in the coefficients bands and the structural behavior of the HR MS image coefficients is used in forming $p(s_b^{ld}, x^{ld} | y_b^{ld}, \Omega_b^{ld})$ and $p(y_b^{ld} | \Omega_b^{ld})$, respectively. These noise and image models depend on the unknown parameters Ω_b^{ld} that need to be estimated. In the second stage a hyperprior on the parameters is defined, thus allowing the incorporation of information about these hyperparameters into the process. Let us define the probability distribution involved in each stage.

Following Refs. [3, 5], we chose a prior model based on the Total Variation (TV) for the HR MS image coefficient bands, y_b^{ld} , given by

$$p(y_b^{ld} | \alpha_b^{ld}) \propto (\alpha_b^{ld})^{p/2} \exp \left\{ -\alpha_b^{ld} TV(y_b^{ld}) \right\}, \quad (2)$$

with $TV(y_b^{ld}) = \sum_{i=1}^p \sqrt{(\Delta_i^h(y_b^{ld}))^2 + (\Delta_i^v(y_b^{ld}))^2}$ where $\Delta_i^h(y_b^{ld})$ and $\Delta_i^v(y_b^{ld})$ represent the horizontal and vertical first order differences at pixel i , respectively, and α_b^{ld} is the model parameter of the MS band b coefficients at level l and direction d . The idea behind this model is to consider the coefficient bands as a set of relatively smooth regions separated by strong edges, such as the coefficients of the NSCT.

Since the MS bands coefficients and the PAN image coefficients are independent given the HR MS image coefficients, we define $p(s_b^{ld}, x^{ld} | y_b^{ld}, \Omega_b^{ld}) = p(s_b^{ld} | y_b^{ld}, \Omega_b^{ld}) \times p(x^{ld} | y_b^{ld}, \Omega_b^{ld})$. The conditional distribution of the upsampled MS coefficients given the HR MS coefficients is defined as [3]

$$p(s_b^{ld} | y_b^{ld}, \beta_b^{ld}) \propto (\beta_b^{ld})^{p/2} \exp \left\{ -\frac{1}{2} \beta_b^{ld} \|s_b^{ld} - y_b^{ld}\|^2 \right\}, \quad (3)$$

where β_b^{ld} is the inverse of the unknown noise variance of the detail band at level l and direction d of the MS band b . The relationship between the HRMS band coefficients and the PAN image is modeled by the conditional probability distribution

$$p(x^{ld} | y_b^{ld}, \gamma_b^{ld}) \propto (\gamma_b^{ld})^{p/2} \exp \left\{ -\frac{1}{2} \gamma_b^{ld} \|x^{ld} - y_b^{ld}\|^2 \right\}. \quad (4)$$

where γ_b^{ld} is the inverse of the unknown noise variance at each NSCT decomposition level, l , and direction, d , of PAN image. Note that, with this modeling, we have decoupled each one of the bands of the contourlet transform and, since they are uncorrelated, we can do the estimation of each band independently of the other bands.

In the second stage of the hierarchical Bayesian framework we define the distribution on the parameters by using a gamma distribution

$$p(w | a_w, c_w) = \Gamma(w | a_w, c_w), \quad (5)$$

where $w > 0, w \in \Omega_b^{ld} = (\alpha_b^{ld}, \beta_b^{ld}, \gamma_b^{ld})$ denotes a hyperparameter, and $a_w > 0$ and $c_w > 0$ are, respectively, the shape and the inverse scale parameters of the distribution.

Finally, combining the first and second stages of the problem modeling, and defining $\Omega_b^{ld} = \{\alpha_b^{ld}, \beta_b^{ld}, \gamma_b^{ld}\}$, we have the global distribution

$$p(\Omega_b^{ld}, y_b^{ld}, x^{ld}, s_b^{ld}) = p(\alpha_b^{ld}) p(\beta_b^{ld}) p(\gamma_b^{ld}) p(y_b^{ld} | \alpha_b^{ld}) \times p(s_b^{ld} | y_b^{ld}, \beta_b^{ld}) p(x^{ld} | y_b^{ld}, \gamma_b^{ld}), \quad (6)$$

where $p(y_b^{ld} | \alpha_b^{ld})$, $p(s_b^{ld} | y_b^{ld}, \beta_b^{ld})$ and $p(x^{ld} | y_b^{ld}, \gamma_b^{ld})$ are given in Eqs. (2), (3), and (4), respectively.

The Bayesian paradigm dictates that inference on the parameters and the image, $(\Omega_b^{ld}, y_b^{ld})$, should be based on $p(\Omega_b^{ld}, y_b^{ld} | s_b^{ld}, x^{ld}) = p(\Omega_b^{ld}, y_b^{ld}, s_b^{ld}, x^{ld}) / p(s_b^{ld}, x^{ld})$. Since $p(s_b^{ld}, x^{ld})$ cannot be calculated analytically, then $p(\Omega_b^{ld}, y_b^{ld} | s_b^{ld}, x^{ld})$ can not be found in closed form. We apply the variational methodology to approximate the posterior distribution by another distribution, $q(\Omega_b^{ld}, y_b^{ld})$, that minimizes the Kullback-Leibler(KL) divergence. We choose to approximate the posterior distribution $p(\Omega_b^{ld}, y_b^{ld} | s_b^{ld}, x^{ld})$ by the distribution $q(\Omega_b^{ld}, y_b^{ld}) = q(\Omega_b^{ld}) q(y_b^{ld})$, where $q(y_b^{ld})$ and $q(\Omega_b^{ld})$ denote distributions on y_b^{ld} and Ω_b^{ld} , respectively.

The estimation of the parameters and the image is done iteratively. First, an estimation of each parameter $w \in \Omega_b^{ld}$ is selected as the mean of the posterior gamma distribution $q(w)$ and then the estimation of the Gaussian distribution of the HR MS coefficients, $q(y_b^{ld})$, is performed.

3. CLASSIFICATION

Once the pansharpened image has been obtained, its classification is carried out. The approach we follow (which will be described later) to improve the classification rate of one class will be tested on two classification methods which, for

completeness, are briefly described now: linear discriminant analysis (LDA) and support vector machines (SVM).

LDA is an effective subspace technique that optimizes Fisher’s score [6]. Subspace methods are a particular class of algorithms focused on finding projections of the original hyperdimensional space to a lower dimensional space where class separation is maximized. In addition, LDA does not require the tuning of free parameters. These good attributes have resulted in its extensive use and practical exploitation in remote sensing applications mainly focused on image classification and band selection. LDA is related to Fisher’s linear discriminant and, roughly speaking, both aim at finding a linear combination of features that characterize or separate two or more classes.

SVM is one of the most successful examples of kernel methods, being a linear classifier that implements maximum margin separation between classes in a high dimensional Hilbert space \mathcal{H} . Kernel methods embed the data observed in the input space \mathcal{X} into a higher dimensional space, the feature space \mathcal{H} , where the data are more likely to be linearly separable. Therefore, it is possible to build an efficient linear classifier in \mathcal{H} , that translates into a nonlinear classifier in the input space. The mapping function to perform such an embedding is denoted as $\Phi : \mathcal{X} \rightarrow \mathcal{H}$. Computing the explicit mappings $\Phi(\mathbf{x})$ of all the observed data points can be computational demanding, especially if the dimensionality of \mathcal{H} is high. To avoid this problem and build efficient algorithms, kernel methods compute the similarity between training samples $\{\mathbf{x}_i\}_{i=1}^n$ using inner products between mapped samples instead of computing the dot product in the higher dimensional space explicitly. The so-called kernel matrix $\mathbf{K}_{ij} = K(\mathbf{x}_i, \mathbf{x}_j) = \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle$ contains all necessary information to perform many classical linear algorithms in the feature space, which are non-linear in the input space [7].

It is important to note that, both for training and using the SVM for testing, one only needs to work with a valid kernel function, which should accurately reflect the similarity between samples. Valid kernels are functions representing a dot product in \mathcal{H} . The radial basis function (RBF), $K(\mathbf{x}, \mathbf{z}) = \exp(-\|\mathbf{x} - \mathbf{z}\|^2/2\sigma^2)$, $\sigma \in \mathbb{R}^+$ was the kernel function selected in this work. To implement SVM for multi-class problems we used the one-versus-all strategy given the particular characteristics of the proposed scheme.

4. IMPROVING THE CLASSIFICATION PERFORMANCE FOR A SINGLE CLASS

Once the pansharpening method described in section 2 has been used on a LR MS image and one of the classification methods described above has been applied, the user may be interested in boosting the performance of the classifier on a given class. In this section we propose to recalculate the parameters of the pansharpening method in order to obtain a new pansharpened image with an improved classification rate

for the class of interest.

By examining the HR classified image, both visually and numerically (using for instance the confusion matrix), the user selects a class to improve its classification figures of merits. A new estimation of the image and parameters is performed. Utilizing the already estimated pansharpened image, the parameters for the new reconstruction are estimated utilizing only the pixels belonging to the class of interest in this image. Using those parameters a new pansharpened image is obtained. No iteration between parameter and image estimates is required.

This result in an estimation of the image whose spectral and spatial characteristics are more tailored to the pixels in the class of interest and, hence, will hopefully increase the classification performance for the elements of the class. Note however that this may imply, as we will see in the experimental section, that the classification performance on the other classes may decrease.

5. EXPERIMENTAL RESULTS

Experiments were run on a Quickbird image. The MS image, depicted in real color in Figure 1a, has a spatial resolution of 256×256 pixels with each pixel covering a square area with a side of 2.4 m and four spectral bands: blue (450-520 nm), green (520-600 nm), red (630-690 nm), near-IR (760-900 nm). The PAN image (see Fig. 1b) has a resolution of 1024×1024 pixels with a size of 0.6 m covering the whole spectral interval (405-1053 nm). The result of the pansharpening process, with the parameters automatically estimated using all the MS and PAN images, is shown in Fig. 1d.

Using the MS and PAN images, a small number of pixels were classified into ten different classes (cars, water, forest, ...). This set of pixels, depicted in Fig. 1c, is considered our ground truth. We randomly chose 20% of the samples of each class to train the LDA and SVM classifiers and the rest was used for testing. In order to incorporate both spectral and spatial characteristics for each pixel into the classification process, we used a descriptor composed of the value of each pixel under consideration and its four nearest neighbors. Since each pixel has associated five values, four corresponding to the MS bands and another one for the panchromatic, the descriptor for each pixel has 25 components.

The classification quality is measured using the precision and recall values on a given class defined as

$$recall = \frac{TP}{TP + FN}, \quad precision = \frac{TP}{TP + FP} \quad (7)$$

where TP is the number of pixels in class correctly classified, FN is the number of pixels in the class incorrectly classified and FP is number of pixels not belonging to a class incorrectly classified as belonging to the class. Table 1 presents the figures of merit for each classifier on the pansharpened image in Fig. 1d. This image presents a very high level of

Table 1. Recall and Precision values obtained using the pansharpened image with parameters estimated from all the pixels in the image.

Class	LDA		SVM	
	recall	precision	recall	precision
1. Asphalt	0.89	0.91	0.99	0.98
2. Dense Forest	0.75	0.72	0.91	0.93
3. Forest	0.87	0.98	0.99	0.98
4. Bare Soil	0.93	0.86	0.99	0.99
5. Building	0.84	0.93	0.99	0.98
6. Grass	0.82	0.82	0.96	0.94
7. Dry Grass	0.99	0.77	0.99	0.99
8. Car	0.63	0.66	0.81	0.98
9. Water	0.93	0.74	0.97	0.98
10. Isolated Tree	0.89	0.29	0.82	0.89

spatial detail with no chromatic distortion. The classification figures show that SVM outperforms LDA although both classifiers perform well for all the classes except classes 10 and 8 where they perform poorly, especially the LDA classifier.

A class is now selected to improve its classification figures. In this case class 10 (isolated tree) was selected although similar results were obtained when selecting the other classes. Using only the pixels of the MS and PAN image belonging to the selected class, the parameters were estimated using the procedure described in section 4 and a new pansharpened image, depicted in Fig. 1e, was obtained.

Using this image, the classifiers were trained and a new classification step was performed obtaining the results presented in Table 2. Although the reconstructed images using the parameters estimated from the whole image (Fig. 1d) and the parameters estimated using only the pixels of the class 10 (Fig. 1e) are very similar from a visual point of view, the classification figures show a higher precision and recall for the selected class 10 and, also, for many others. Note however, that some classes, like classes 6 or 8, perform slightly worse with those parameters.

6. CONCLUSIONS

In this paper we have shown that pansharpening techniques can be used to increase the performance of classification methods when are applied to MS images. We have addressed the problem of adaptively modifying a pansharpening method in order to improve the precision and recall figures of merit of the classification on a given class without deteriorating the performance of the classifier over the other classes. The validity of the proposed technique has been demonstrated using a real Quickbird image. Work is being currently carried out to theoretically justify the used approach.

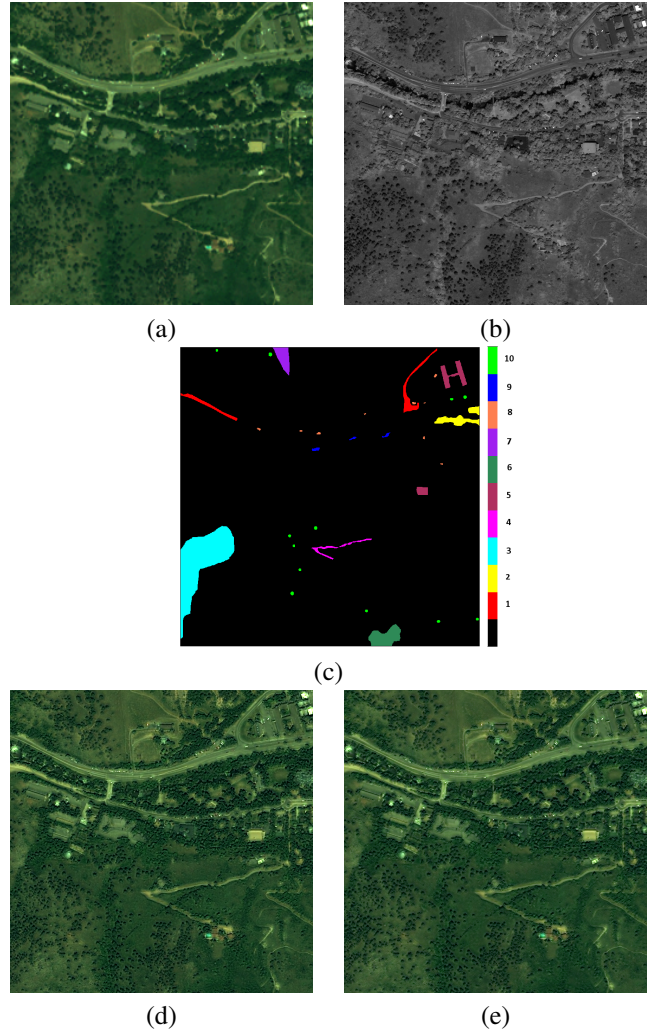


Fig. 1. (a) MS and (b) PAN images. (c) Ground truth. (d) Pansharpened image using the super resolution method described in section 2 with the parameters estimated using the whole image. (e) Pansharpened image utilizing only the pixels of the training set in class 10 to estimate the model parameters.

Table 2. Recall and Precision values obtained using the pan-sharpened image with parameters estimated only from pixels of the class 10.

Class	LDA		SVM	
	recall	precision	recall	precision
1. Asphalt	0.89	0.90	0.99	0.97
2. Dense Forest	0.72	0.73	0.92	0.93
3. Forest	0.88	0.99	0.99	0.99
4. Bare Soil	0.94	0.87	0.99	0.99
5. Building	0.82	0.93	0.99	0.98
6. Grass	0.83	0.80	0.96	0.95
7. Dry Grass	0.99	0.80	0.99	0.99
8. Car	0.54	0.56	0.75	0.94
9. Water	0.94	0.72	0.99	0.99
10. Isolated Tree	0.89	0.30	0.88	0.92

[7] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*, (Cambridge University Press, 2004).

Acknowledgements

This work has been supported by the Consejería de Innovación, Ciencia y Empresa of the Junta de Andalucía under contract P07-FQM-02701, by the Comisión Nacional de Ciencia y Tecnología under contract TIN2010-15137, and the Spanish research program Consolider Ingenio 2010: MIPRCV (CSD2007-00018).

7. REFERENCES

[1] I. Amro, Multispectral Image fusion using Multiscale and Super-resolution methods, PhD thesis, Dept. of Computer Science and Artificial Intelligence, Universidad de Granada 2011.

[2] L. Bruzzone, L. Carlin, L. Alparone, S. Baronti, A. Garzelli and F. Nencini, Can multiresolution fusion techniques improve classification accuracy? *Image and Signal Processing for Remote Sensing XII*, **6365** 2006.

[3] I. Amro, J. Mateos, M. Vega, General contourlet pan-sharpening method using Bayesian inference, in *2010 European Signal Processing Conference (EUSIPCO-2010)*, 2010.

[4] A. L. da Cunha, J. Zhou and, M. N. Do, The nonsub-sampled contourlet transform: theory, design, and applications, in *IEEE Trans. Image Proc.*, **15**, 3089 (2006).

[5] M. Vega, J. Mateos, R. Molina and A.K.. Katsaggelos, Super resolution of multispectral images using TV image models, in *2th Int. Conf. on Knowledge-Based and Intelligent Information & Eng. Sys.*, 2008.

[6] R. Duda and P. Hart, *Pattern classification and scene analysis* (Wiley, New York, USA, 1973).

4.2 Learning Filters in Gaussian Process Classification Problems

- **P. Ruiz**, J. Mateos, R. Molina, and A.K. Katsaggelos, “Learning Filters in Gaussian Process Classification Problems” in *IEEE International Conference on Image Processing (ICIP 2014)*, 2913-2917, Paris (France), October 2014.
 - Status: Published
 - Indexed in CORE Conference Ranking as CORE B
 - H index: 35 (Q1:81/1201)

LEARNING FILTERS IN GAUSSIAN PROCESS CLASSIFICATION PROBLEMS

Pablo Ruiz^{1}, Javier Mateos¹, Rafael Molina¹ and Aggelos K. Katsaggelos²*

¹ Dpto. de Ciencia de la Computación e I.A. Universidad de Granada.

² Dpt. of Electrical Engineering and Computer Science. Northwestern University.

*e-mail:mataran@decsai.ugr.es

ABSTRACT

Many real classification tasks are oriented to sequence (neighbor) labeling, that is, assigning a label to every sample of a signal while taking into account the sequentiality (or neighborhood) of the samples. This is normally approached by first filtering the data and then performing classification. In consequence, both processes are optimized separately, with no guarantee of global optimality. In this work we utilize Bayesian modeling and inference to jointly learn a classifier and estimate an optimal filterbank. Variational Bayesian inference is used to approximate the posterior distributions of all unknowns, resulting in an iterative procedure to estimate the classifier parameters and the filterbank coefficients. In the experimental section we show, using synthetic and real data, that the proposed method compares favorably with other classification/filtering approaches, without the need of parameter tuning.

Index Terms— Gaussian Process classification, filter estimation, analysis representation.

1. INTRODUCTION

Many real classification tasks assign a label to every sample of a signal (or pixel of an image) while taking into account the sequentiality (or vicinity) of the samples. This task is normally approached by first filtering the data and then performing classification. For instance, a super resolution method can be applied to a multispectral image [1] followed by a classification method on the improved multispectral image [2]; or an improved passive millimeter-wave image can be obtained [3] followed by an object detection procedure [4].

Using filtering as a pre-processing step before learning a classifier does not guarantee optimal joint performance. To solve this problem, we propose a Bayesian framework to learn a classifier, at the same time estimate an optimal filterbank to improve the classifier performance.

Let us assume that we have access to a multichannel sequential signal or multichannel sequential features extracted from the signal. We use the term “multichannel features” to refer to both concepts for simplicity. Let $\mathbf{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_N]$ be the matrix including these original input features, where each feature \mathbf{z}_i is of length B . Instead of performing classification directly on the features \mathbf{Z} , we would like to compute new features $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_N]$ so as to optimize the classification performance. \mathbf{Z} and \mathbf{X} can be related in two different ways. The first method, based on the analysis representation, obtains \mathbf{X} as a linear transformation of \mathbf{Z} , leading to $\mathbf{X} = \mathbf{AZ}$ where \mathbf{A} defines a linear filterbank whose coefficients must be estimated. The

analysis representation appears in many signal reconstruction problems. For instance, it can be used to improve the classification of EEG data in brain-computer interfaces [5], or to discover causality interaction in functional MRI [6].

In the second method, based on the synthesis representation, \mathbf{Z} is represented using a dictionary \mathbf{D} that has to be learnt from a set of samples, that is, $\mathbf{Z} = \mathbf{DX}$. The new features \mathbf{X} are to be used to classify the samples. The synthesis representation model is related, for instance, to the use of discriminative Gaussian Process Latent Variable Models (GPLVM) [7], where a linear discriminant prior on the latent variables is introduced and bears some connections with learning discriminative dictionaries (see, for instance, [8, 9]). In this work we use the analysis representation.

The idea of jointly optimizing a filter and a classifier dates back to the 1990s within the field of artificial neural networks. It was, for instance, used in convolutional networks [10] or to define a neural model for temporal processing [11, 12]. Recently, the same principle is used in [13] where filters are learnt jointly with a support vector machine (SVM) to perform classification.

In this work the filtering/classification tasks are formulated as a single Bayesian inference problem. Variational inference is used to learn the classifier and the optimal filterbank coefficients as well as the model parameters. The rest of this paper is organized as follows. In Section 2 Bayesian modeling is presented to use analysis representation on images. Variational Inference is performed in Section 3. The classification rule is introduced in Section 4. In Section 5 results for both synthetic and real experiments are presented and finally Section 6 concludes the paper.

2. HIERARCHICAL BAYESIAN MODELING

Let us assume that during the training phase we have access to $\mathbf{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_N]$, where each \mathbf{z}_i is of length B , and their corresponding labels $\mathbf{y} = [y_1, \dots, y_N]^T$ with $y_i \in \{0, 1\}$. To obtain the new features each band is filtered with a spatial filter $\mathbf{a}_i \in \mathbb{R}^{k^2}$, $i = 1, \dots, B$, producing

$$\mathbf{X} = \mathbf{AZ} = \begin{bmatrix} \mathbf{a}_1^T & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{a}_2^T & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{a}_B^T \end{bmatrix} \begin{bmatrix} \mathbf{z}_{1,1} & \mathbf{z}_{1,2} & \dots & \mathbf{z}_{1,N} \\ \mathbf{z}_{2,1} & \mathbf{z}_{2,2} & \dots & \mathbf{z}_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{z}_{B,1} & \mathbf{z}_{B,2} & \dots & \mathbf{z}_{B,N} \end{bmatrix}, \quad (1)$$

where each $\mathbf{z}_{i,j}$ is a column vector of size k^2 containing the neighborhood of the j -th sample in the i -th band. To reduce the number of coefficients in \mathbf{A} to be estimated, we only perform intraband filtering. Interband filtering is not performed because the classifier utilizes multiband information.

This work has been supported in part by the Comisión Nacional de Ciencia y Tecnología under contract TIN2010-15137, CEI BioTic at the University of Granada, and the Department of Energy grant DE-NA0000457.

To model the classification function relating each sample \mathbf{x}_i to its corresponding label y_i , we follow a two stage procedure. First, we introduce a latent variable f_i which is related to y_i by a sigmoidal function $y_i = \sigma(f_i) = 1/(1 + e^{-f_i})$. Let $\mathbf{f} = [f_1, \dots, f_N]$ be the values of the latent function at $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_N]$, then the joint likelihood factorizes to

$$p(\mathbf{y}|\mathbf{f}) = \prod_{i=1}^N [\sigma(f_i)]^{y_i} [1 - \sigma(f_i)]^{1-y_i}. \quad (2)$$

In the second stage, to model \mathbf{f} , we define on f_i a Gaussian Process, which depends on \mathbf{X} , and so we write

$$p(\mathbf{f}|\mathbf{X}, \mu, \gamma, \sigma) = \mathcal{N}(\mathbf{f}|\mu\mathbf{1}, \mathbf{C}), \quad (3)$$

where $\mathbf{C} = \gamma\mathbf{K}_{\mathbf{X}} + \sigma\mathbf{I}$, and $\mathbf{K}_{\mathbf{X}} = (\mathbf{k}(\mathbf{x}_i, \mathbf{x}_j))$, $i, j = 1, \dots, N$, is the kernel used. In this work linear and Gaussian kernels are considered (see [14] for details).

To model \mathbf{X} , instead of enforcing Eq. (1), we consider a weaker constraint by defining the following pseudo-observation model

$$p(\mathbf{X}|\mathbf{Z}, \mathbf{A}) \propto \exp\left(-\frac{\beta}{2}\|\mathbf{X} - \mathbf{AZ}\|_{\text{F}}^2\right), \quad (4)$$

where $\|\cdot\|_{\text{F}}$ is the Frobenius norm. When $\beta \rightarrow \infty$ we obtain the constraint $\mathbf{X} = \mathbf{AZ}$. In Sect. 3 we explain how to configure the penalty β .

With no much prior information on the filterbank coefficients, we follow the approach in [13] and use the following prior on \mathbf{A} ,

$$p(\mathbf{A}|\boldsymbol{\alpha}) = \prod_{i=1}^B p(\mathbf{a}_i|\alpha_i) = \prod_{i=1}^B \mathcal{N}(\mathbf{a}_i|\mathbf{0}, \alpha_i^{-1}\mathbf{I}_{k_2}), \quad (5)$$

where $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_B)^T$ are the precision coefficients, which are modeled using Gamma distributions, that is,

$$p(\boldsymbol{\alpha}) = \prod_{i=1}^B p(\alpha_i) \propto \prod_{i=1}^B \alpha_i^{a_i-1} \exp(-b_i\alpha_i). \quad (6)$$

The parameters a_i and b_i are treated as deterministic whose values are set to small values (e.g., 10^{-5}) to obtain broad hyperpriors.

Finally, the joint distributions factorizes as

$$p(\mathbf{y}, \Theta) = p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{X}, \mu, \gamma, \sigma)p(\mathbf{X}|\mathbf{A}, \mathbf{Z}, \beta)p(\mathbf{A}|\boldsymbol{\alpha})p(\boldsymbol{\alpha}).$$

where $\Theta = \{\mathbf{f}, \mathbf{X}, \mathbf{A}, \boldsymbol{\alpha}, \mu, \gamma, \sigma\}$, and \mathbf{Z} is fixed.

3. BAYESIAN INFERENCE AND VARIATIONAL APPROXIMATION

In our Bayesian framework, unknown variables are estimated from the posterior distribution $p(\Theta|\mathbf{y}) = p(\mathbf{y}, \Theta)/p(\mathbf{y})$. However this distribution is not tractable because $p(\mathbf{y})$ can not be calculated. To alleviate this problem, variational methods are used to approximate it by a tractable distribution of the form

$$q(\Theta) = q(\mathbf{f})q(\mathbf{X})q(\mu)q(\gamma)q(\sigma) \prod_{i=1}^B q(\mathbf{a}_i)q(\alpha_i). \quad (7)$$

The variational criterion used to find $q(\Theta)$ is the minimization of the Kullback-Leibler (KL) divergence [14], given by

$$C_{\text{KL}}(q(\Theta)||p(\Theta|\mathbf{y})) = \int q(\Theta) \log \frac{q(\Theta)}{p(\mathbf{y}, \Theta)} d\Theta + \text{const} \quad (8)$$

which is always non negative and equal zero if and only if the distributions $q(\Theta)$ and $p(\Theta|\mathbf{y})$ coincide.

Due to the form of the joint likelihood defined in Eq. (2), the KL divergence cannot be evaluated. To solve this problem we bound the joint likelihood in Eq. (2), using the variational lower bound [14, 15]

$$\ln(1 + e^u) \leq \lambda(\xi)(u^2 - \xi^2) + \frac{u - \xi}{2} + \ln(1 + e^\xi), \quad (9)$$

where $\lambda(\xi) = \frac{1}{2\xi} \left(\frac{1}{1+e^{-\xi}} - \frac{1}{2} \right)$. Thus the joint likelihood is bounded as:

$$p(\mathbf{y}|\mathbf{f}) \geq \exp\left\{ \left(\mathbf{y} - \frac{1}{2}\mathbf{1}\right)^T \mathbf{f} - \mathbf{f}^T \Lambda \mathbf{f} \right\} \times \exp\left\{ \boldsymbol{\xi}^T \Lambda \boldsymbol{\xi} + \frac{1}{2}\mathbf{1}^T \boldsymbol{\xi} \right\} \prod_{i=1}^N \sigma(-\xi_i) = \mathbf{H}(\mathbf{y}, \mathbf{f}, \boldsymbol{\xi}), \quad (10)$$

where $\boldsymbol{\xi} = (\xi_1, \dots, \xi_N)^T$, and $\Lambda = \text{Diag}(\lambda(\xi_1), \dots, \lambda(\xi_N))$. The inequality in Eq. (10) leads to the following lower bound for the joint probability distribution:

$$p(\mathbf{y}, \Theta) \geq \mathbf{M}(\mathbf{y}, \Theta, \boldsymbol{\xi}) = \mathbf{H}(\mathbf{y}, \mathbf{f}, \boldsymbol{\xi})p(\mathbf{f}|\mathbf{X}, \mu, \gamma, \sigma)p(\mathbf{X}|\mathbf{A}, \mathbf{Z}, \beta)p(\mathbf{A}|\boldsymbol{\alpha})p(\boldsymbol{\alpha}). \quad (11)$$

Finally, the KL divergence in Eq. (8) is majorized by

$$C_{\text{KL}}(q(\Theta)||p(\Theta|\mathbf{y})) \leq C_{\text{KL}}(q(\Theta)||\mathbf{M}(\mathbf{y}, \Theta, \boldsymbol{\xi})) + \text{const}. \quad (12)$$

Although a new set of unknowns $\boldsymbol{\xi}$ has been included, now the KL divergence between $q(\Theta)$ and $\mathbf{M}(\mathbf{y}, \Theta, \boldsymbol{\xi})$ is mathematically tractable, and it can be used to calculate the posterior distribution $q(\Theta)$. The optimal posterior distribution approximation is the given by [14]

$$q(\theta) \propto \exp[\langle \log \mathbf{M}(\mathbf{y}, \Theta, \boldsymbol{\xi}) \rangle_{q(\Theta_\theta)}], \quad (13)$$

where $\theta \in \Theta$, the set Θ_θ represents the set difference $\Theta \setminus \{\theta\}$ and the operator $\langle \cdot \rangle_{q(\Theta_\theta)}$ denotes expected value with respect to the distribution $q(\Theta_\theta)$. For simplicity we use $\langle \mathbf{u} \rangle$ to denote $\langle \mathbf{u} \rangle_{q(\mathbf{u})}$. In this paper we assume that $q(\mathbf{X})$, $q(\mu)$, $q(\gamma)$ and $q(\sigma)$ are degenerate distributions. No constraints are imposed on $q(\mathbf{f})$, $q(\mathbf{a}_i)$ and $q(\alpha_i)$.

Since $\langle \log \mathbf{M}(\mathbf{y}, \Theta, \boldsymbol{\xi}) \rangle_{q(\Theta_\mathbf{f})}$ is a quadratic function on \mathbf{f} , its posterior distribution approximation is a Gaussian distribution with parameters

$$\boldsymbol{\mu}_{\mathbf{f}} = \Sigma_{\mathbf{f}} \left[\mathbf{y} - \frac{1}{2}\mathbf{1} + \mu\mathbf{C}^{-1}\mathbf{1} \right], \quad \Sigma_{\mathbf{f}} = (\mathbf{C}^{-1} + 2\Lambda)^{-1}. \quad (14)$$

The value where $q(\mu)$ is degenerate is obtained by solving

$$\hat{\mu} = \arg \min_{\mu} \langle \log \mathbf{M}(\mathbf{y}, \Theta, \boldsymbol{\xi}) \rangle_{q(\Theta_\mu)}. \quad (15)$$

By differentiating $\langle \log \mathbf{M}(\mathbf{y}, \Theta, \boldsymbol{\xi}) \rangle_{q(\Theta_\mu)}$ with respect to μ and equating to zero we obtain

$$\hat{\mu} = \frac{(\boldsymbol{\mu}_{\mathbf{f}})^T \mathbf{C}^{-1} \mathbf{1}}{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{1}}. \quad (16)$$

Following the same procedure for γ in Eq. (3) we obtain

$$(\mu\mathbf{1} - \boldsymbol{\mu}_{\mathbf{f}})^T \mathbf{C}^{-1} \mathbf{K}_{\mathbf{X}} \mathbf{C}^{-1} (\mu\mathbf{1} - \boldsymbol{\mu}_{\mathbf{f}}) - \text{Tr}[\mathbf{C}^{-1} \Sigma_{\mathbf{f}} \Lambda \mathbf{K}_{\mathbf{X}}] = 0,$$

where γ is included in \mathbf{C} . Then we use the following fixed point algorithm (see [2] for details) to update γ

$$\gamma = \frac{\gamma (\mu\mathbf{1} - \boldsymbol{\mu}_{\mathbf{f}})^T \mathbf{C}^{-1} \mathbf{K}_{\mathbf{X}} \mathbf{C}^{-1} (\mu\mathbf{1} - \boldsymbol{\mu}_{\mathbf{f}})}{2 \text{Tr}[\mathbf{C}^{-1} \Sigma_{\mathbf{f}} \Lambda \mathbf{K}_{\mathbf{X}}]}, \quad (17)$$

where the old value of γ is used in the right hand side to obtain an updated value in the left hand side. The same procedure is used on σ to obtain the updating rule

$$\sigma = \frac{\sigma(\boldsymbol{\mu}_f - \boldsymbol{\mu}_f)^T \mathbf{C}^{-1} \mathbf{C}^{-1} (\boldsymbol{\mu}_f - \boldsymbol{\mu}_f)}{2 \text{Tr}[\mathbf{C}^{-1} \boldsymbol{\Sigma}_f \boldsymbol{\Lambda}]} \quad (18)$$

To estimate $\boldsymbol{\xi}$ we solve the optimization problems

$$\hat{\boldsymbol{\xi}}_i = \arg \min_{\boldsymbol{\xi}_i} \langle \log \mathbf{M}(\mathbf{y}, \Theta, \boldsymbol{\xi}) \rangle_{q(\Theta)} \quad (19)$$

Differentiating and equating to zero we obtain

$$\hat{\boldsymbol{\xi}}_i = \sqrt{\langle \boldsymbol{\mu}_f \rangle_i^2 + \langle \boldsymbol{\Sigma}_f \rangle_{ii}} \quad (20)$$

Since $\langle \log \mathbf{M}(\mathbf{y}, \Theta, \boldsymbol{\xi}) \rangle_{q(\Theta_{\mathbf{a}_i})}$ is a quadratic function on \mathbf{a}_i , $q(\mathbf{a}_i)$ is a Gaussian distribution with parameters

$$\langle \mathbf{a}_i \rangle = \beta \boldsymbol{\Sigma}_i \mathbf{Z}_i (\mathbf{X}_i)^T, \quad \boldsymbol{\Sigma}_i = (\beta \mathbf{Z}_i \mathbf{Z}_i^T + \langle \alpha_i \rangle \mathbf{I}_{k_2})^{-1}, \quad (21)$$

where \mathbf{X}_i , $i = 1 \dots, B$, represent the i -th row of \mathbf{X} .

The posterior density of α_i becomes a Gamma distribution with mean

$$\langle \alpha_i \rangle = \frac{2a_i + k^2}{2b_i + \text{Tr}(\boldsymbol{\Sigma}_i + \langle \mathbf{a}_i \rangle \langle \mathbf{a}_i \rangle^T)} \quad (22)$$

Finally, to estimate \mathbf{X} we solve

$$\hat{\mathbf{X}} = \arg \min_{\mathbf{X}} \langle \log \mathbf{M}(\mathbf{y}, \Theta, \boldsymbol{\xi}) \rangle_{q(\Theta_{\mathbf{X}})} \quad (23)$$

For a linear kernel we have

$$\mathbf{X}^T = \left[\gamma (\mathbf{C}^{-1} \boldsymbol{\Sigma}_f 2\boldsymbol{\Lambda}) - \gamma \mathbf{C}^{-1} (\boldsymbol{\mu}_f - \boldsymbol{\mu}_f) (\boldsymbol{\mu}_f - \boldsymbol{\mu}_f)^T \mathbf{C}^{-1} + \beta \mathbf{I}_N \right]^{-1} \beta \mathbf{Z}^T \langle \mathbf{A} \rangle^T \quad (24)$$

For a Gaussian kernel case with a fixed scale parameter, we obtain an update rule for each component of \mathbf{X} . Thus, for the p -th component of \mathbf{x}_i we obtain

$$x_{pi} = \frac{\frac{\gamma}{s^2} \sum_{t \neq i}^N (v_i v_t - w_{it} (\lambda(\xi_i) + \lambda(\xi_t))) e_{it} \mathbf{x}_t(p) + \beta \mathbf{a}_p^T \mathbf{z}_{p,i}}{\frac{\gamma}{s^2} \sum_{t \neq i}^N (v_i v_t - w_{it} (\lambda(\xi_i) + \lambda(\xi_t))) e_{it} + \beta}, \quad (25)$$

where s is the scale parameter of the Gaussian kernel, $\mathbf{v} = \mathbf{C}^{-1} (\boldsymbol{\mu}_f - \boldsymbol{\mu}_f)$, $w_{ij} = (\mathbf{I} + 2\sigma\boldsymbol{\Lambda} + 2\gamma \mathbf{K}_{\mathbf{X}\boldsymbol{\Lambda}})^{-1}$, and $e_{ij} = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2s^2)$.

To configure the proximity operator penalty, β , we multiply Eq. (24) by $\frac{\gamma + \beta}{\gamma + \beta}$ and define $\tau = \frac{\beta}{\gamma + \beta}$. Hence, Equation (24) can then be written as

$$\mathbf{X}^T = \left[-(1 - \tau) \mathbf{C}^{-1} (\boldsymbol{\mu}_f - \boldsymbol{\mu}_f) (\boldsymbol{\mu}_f - \boldsymbol{\mu}_f)^T \mathbf{C}^{-1} + (1 - \tau) (\mathbf{C}^{-1} \boldsymbol{\Sigma}_f 2\boldsymbol{\Lambda}) + \tau \mathbf{I}_N \right]^{-1} \tau \mathbf{Z}^T \langle \mathbf{A} \rangle^T \quad (26)$$

Note that $\tau \in [0, 1]$ and when $\tau = 1$, we obtain $\mathbf{X} = \langle \mathbf{A} \rangle \mathbf{Z}$ in Eq. (26). For Gaussian kernels we proceed in the same manner but multiplying Eq. (25) by $\frac{\gamma/s^2 + \beta}{\gamma/s^2 + \beta}$.

Let us now summarize the estimation procedure. Starting with $\mathbf{X}^0 = \mathbf{A}^0 \mathbf{Z}$, $(\mathbf{K}_{\mathbf{X}}^0)_{ij} = \mathbf{k}(\mathbf{x}_i^0, \mathbf{x}_j^0)$, $\mathbf{a}_i^0 = \text{identity filter}$, $\alpha_i^0 = 1$, $\boldsymbol{\mu}^0 = 0$, $\boldsymbol{\gamma}^0 = 1$, $\boldsymbol{\sigma}^0 = 1$, $\mathbf{C}^0 = \boldsymbol{\gamma}^0 \mathbf{K}_{\mathbf{X}}^0 + \boldsymbol{\sigma}^0 \mathbf{I}$, and $\boldsymbol{\xi}_i^0 = 1$, the method iterates until convergence between Eqs. (14), (16), (17), (18), (20), (21), (22) and (23). We use the old value of the parameter in the right hand side of the estimations to obtain the new values in the left hand side. For the value for τ , we have experimentally found that using $\tau^{n+1} = \min(\tau^n + 0.001, 1)$ made the iterative process to first concentrate on the estimation of the model parameters and then proceed to estimate the filter coefficients. See the experimental section to determine the initial value of τ . At convergence of the estimation procedure we obtain the classifier and the filterbank coefficients.

4. CLASSIFICATION OF NEW PIXELS

In order to classify a new sample \mathbf{z} we transform it using the equation $\mathbf{x} = \langle \mathbf{A} \rangle \mathbf{z}$ where $\langle \mathbf{A} \rangle$ has been obtained at convergence of the training phase and denote by $f_{\mathbf{x}}$ its associated latent variable. Then $p(f_{\mathbf{x}} | \mathbf{f}, \mathbf{X}, \mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\gamma}, \boldsymbol{\sigma})$ is a Gaussian distribution with mean and variance

$$\langle f_{\mathbf{x}} | \mathbf{f}, \mathbf{X}, \mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\gamma}, \boldsymbol{\sigma} \rangle = \boldsymbol{\mu} + \mathbf{h}^T \mathbf{C}^{-1} (\mathbf{f} - \boldsymbol{\mu}_f),$$

$$\text{var}(f_{\mathbf{x}} | \mathbf{f}, \mathbf{X}, \mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\gamma}, \boldsymbol{\sigma}) = c - \mathbf{h}^T \mathbf{C}^{-1} \mathbf{h},$$

where $c = \boldsymbol{\gamma} \mathbf{k}(\mathbf{x}, \mathbf{x})$ and $\mathbf{h} = \boldsymbol{\gamma} (\mathbf{k}(\mathbf{x}, \mathbf{x}_1), \dots, \mathbf{k}(\mathbf{x}, \mathbf{x}_N))^T$ and $\boldsymbol{\gamma}$, $\boldsymbol{\sigma}$ and $\boldsymbol{\mu}$ have been provided by the proposed method at convergence.

We then have

$$p(f_{\mathbf{x}} | \mathbf{y}, \mathbf{X}, \mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\gamma}, \boldsymbol{\sigma}) = \int_{\mathbf{f}} p(f_{\mathbf{x}} | \mathbf{f}, \mathbf{X}, \mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\gamma}, \boldsymbol{\sigma}) \mathcal{N}(\mathbf{f} | \boldsymbol{\mu}_f, \boldsymbol{\Sigma}_f) d\mathbf{f},$$

which is a Gaussian distribution with parameters

$$\langle f_{\mathbf{x}} | \mathbf{y}, \mathbf{X}, \mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\gamma}, \boldsymbol{\sigma} \rangle = \boldsymbol{\mu} + \mathbf{h}^T \mathbf{C}^{-1} (\boldsymbol{\mu}_f - \boldsymbol{\mu}_f),$$

$$\text{var}(f_{\mathbf{x}} | \mathbf{y}, \mathbf{X}, \mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\gamma}, \boldsymbol{\sigma}) = \mathbf{h}^T \mathbf{C}^{-1} \boldsymbol{\Sigma}_f \mathbf{C}^{-1} \mathbf{h} + c - \mathbf{h}^T \mathbf{C}^{-1} \mathbf{h},$$

This leads to the following classification procedure

$$y_{\mathbf{x}} = \begin{cases} 1 & \text{if } \boldsymbol{\mu} + \mathbf{h}^T \mathbf{C}^{-1} (\boldsymbol{\mu}_f - \boldsymbol{\mu}_f) \geq 0 \\ 0 & \text{if } \boldsymbol{\mu} + \mathbf{h}^T \mathbf{C}^{-1} (\boldsymbol{\mu}_f - \boldsymbol{\mu}_f) < 0 \end{cases} \quad (27)$$

5. EXPERIMENTAL RESULTS

In this section synthetic and real experiments are conducted to evaluate the performance of the proposed method, named GPF. In both experiments, we used filters of size 3×3 , 5×5 , 7×7 and 9×9 . We ran the proposed method for different values of τ^0 in the interval $[0.1, 0.9]$ with step 0.01 and selected the one giving the best classification results. GPF was compared with the SVMF method [13] which jointly learn a SVM classifier and estimates a filterbank as well as a GP classifier which does not filter the data. To do this, SVM objective function is augmented with a regularization term on the filters. Hence, in addition to the cost parameter of the SVM (C), the regularization coefficient of the filter (λ) has to be selected. We ran SVMF on $(C, \lambda) \in \{0.01, 0.05, 0.1, 0.5, 1, 5, 10, 25, 50, 100\}^2$ and selected the values producing the best performance.

To obtain unbiased conclusions from the results, ten independent repetitions of the experiments were carried out. For each of them, a training set of 40 randomly selected samples (20 from each class) and a test set of 2000 samples were used. The Overall Accuracy (OA), the estimated Cohen's kappa statistic (κ -index) and Z-score are used as measures of accuracy and class agreement. We also report the computational cost in seconds of each algorithm, implemented using MATLAB[®] on a i7 at 2.80 GHz.

5.1. Synthetic data experiment

In the synthetic data experiment, we generated a 500×500 binary image where black and white pixels alternate in a checkerboard fashion. Observations in the class \mathcal{C}_0 (black pixels) are generated by a Gaussian distribution of mean 0.25 and standard deviation 0.4, observations of pixels in the class \mathcal{C}_1 (white pixels) are generated by a Gaussian distribution of mean 0.75 and standard deviation 0.4. Figure 1a shows a zoom of the observation dataset. Notice that it is hard to decide the class of some pixels by considering only their values.

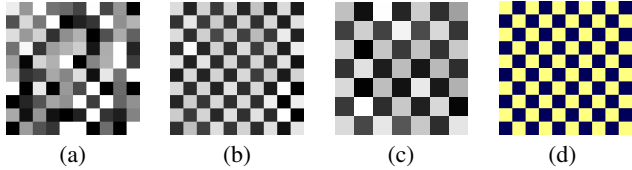


Fig. 1. (a) A set of observations of the synthetic dataset. (b) Filtered observations with the estimated kernel. (c) Estimated 7×7 kernel. (d) Classification map: dark blue C_0 , light yellow C_1 .

Table 1. Figures of merit for the synthetic experiment.

Sizes	GPF				SVMF						
	τ^0	OA	κ	Z	Time	C	λ	OA	κ	Z	Time
3×3	0.86	96.06	0.9212	106.75	0.26	0.1	5	95.84	0.9168	103.80	0.30
5×5	0.86	99.72	0.9943	452.77	0.31	0.5	1	99.69	0.9938	444.85	0.56
7×7	0.87	100	1	∞	1.02	0.5	1	100	1	∞	0.50
9×9	0.87	100	1	∞	1.69	0.5	1	100	1	∞	0.60
No Filter	-	71.06	0.4314	21.39	0.05	0.1	-	72.72	0.4537	22.74	0.003

However, in the filtered image, shown in Fig. 1b, it is easier to distinguish the class of each pixel. It is worth noting that the estimated filter, depicted in Fig. 1c, alternates positive coefficients, in the position of pixels belonging to the class of kernel central pixel, with negative coefficient, in the remainder positions. Figure 1d displays the classification map for the image in Fig. 1a, with a 100% OA.

Mean values for OA, κ -index and Z-score and the value of τ providing the best classification results are reported in Table 1. The proposed GPF method obtained an OA above 96% for all considered filter sizes and, for sizes of 7×7 and 9×9 , the estimated filter is capable to linearly separate both class and a 100% OA is obtained. In all the cases, an improvement of almost 30% is obtained over the base case where the data are not filtered (see the last row of Table 1). The computational cost of the algorithm is very limited needing only between 0.26 and 1.69 seconds to perform both training and classification tasks. The figures of merit for the SVMF method are very similar to those of the GPF although the proposed method scored slightly better for the kernel sizes of 3×3 and 5×5 .

5.2. Real data experiment

The dataset was extracted from a 7-bands satellite image of city of Naples (Italy) captured by the Landsat TM sensor in 1995 in the Urban Expansion Monitoring project (UrbEx) [16]. A small RGB region of this image is displayed in Fig. 2a. A reference land cover map was also provided by the Italian Institute of Statistics (ISTAT). The goal is the discrimination of urban (C_1) versus non-urban (C_0) land-cover classes. The reference land cover map for the image in Fig. 2a is shown in Fig. 2b. Light yellow color represents urban class, dark blue color represents non-urban and red corresponds to pixels whose class is unknown.

In this experiment we used a Gaussian kernel with parameter $s = 100$. This value was selected as the one giving the best results for SVMF. Table 2 shows the mean values for OA, κ -index and Z-score for GPF and SVMF method. Baseline case results, when no filtering is used, are also reported. GPF obtained over a 95% OA, values above 0.90 of κ -index and Z-score values close to 100 for all filter sizes, while the running time moved from 0.48 to 1.62 seconds as the kernel size increased. Those figures of merit indicates a small but significant improvement over the baseline case. In this real situation, GPF consistently obtained better results than SVMF for all kernel sizes. Also, the proposed GPF method ran much faster than

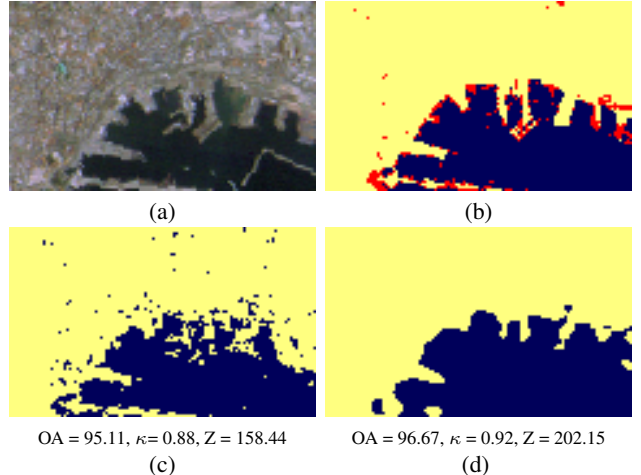


Fig. 2. (a) RGB representation of a small region of the real image. (b) Its reference land cover map. (c) Classification map without filtering. (d) Classification map with filtering.

Table 2. Figures of merit for the real experiment.

Sizes	GPF				SVMF						
	τ^0	OA	κ	Z	Time	C	λ	OA	κ	Z	Time
3×3	0.89	95.18	0.9036	95.02	0.48	50	100	93.92	0.8785	86.02	6.25
5×5	0.87	95.78	0.9156	103.03	0.64	100	50	93.16	0.8633	77.55	20.22
7×7	0.90	95.64	0.9127	100.42	0.81	100	50	93.09	0.8618	76.93	51.15
9×9	0.85	95.25	0.9049	95.50	1.62	100	50	92.88	0.8558	74.61	79.81
No Filter	-	93.21	0.8542	77.15	0.34	0.01	-	92.88	0.8577	74.93	0.11

SVMF (more than 50 times faster in some cases).

To better understand the role of filtering in the proposed method, Figures 2c and 2d depict the classification map for the image in Fig. 2a when no filtering was applied and when kernels of size 5×5 were estimated, respectively. The classification map when the image is not filtered is quite noisy, specially at the boundary of urban and non-urban areas, while the one for the filtered image exhibit more homogeneous regions and, although some pixels are misclassified, class boundaries are much better delimited. The figures of merit for this particular area are shown under their corresponding map. Although the OA for the filtered case is only a 1.5% better than the one for the unfiltered case, filtering allows for a significantly better class agreement reflected in a higher κ -index and Z-score.

6. CONCLUSIONS

In this work we have presented a new method to jointly filter and classify a signal or an image. Using Bayesian modeling and variational inference we have developed an iterative procedure to jointly estimate the classifier parameters, the filterbank and the model parameters. In the experimental section we have shown that the estimated filters helps to improve the classifier performance. The proposed method has been compared with other classification/filtering approaches, and experimental results have shown that the proposed method is more accurate and efficient.

7. REFERENCES

- [1] I. Amro, J. Mateos, M. Vega, R. Molina, and A.K. Katsaggelos, "A survey of classical methods and new trends in pansharpen-

- ing of multispectral images,” *EURASIP Journal on Advances in Signal Processing*, vol. 2011, pp. 2011:79, September 2011.
- [2] P. Ruiz, J. Mateos, G. Camps-Valls, R. Molina, and A.K. Katsaggelos, “Bayesian active remote sensing image classification,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 4, pp. 2186–2196, 2014.
- [3] B. Amizic, L. Spinoulas, R. Molina, and A. K. Katsaggelos, “Compressive sampling with unknown blurring function: application to passive millimeter-wave imaging,” in *IEEE International Conference on Image Processing*. Orlando, Florida, October 2012, pp. 925–928.
- [4] O. Martinez, L. Ferraz, X. Binefa, I. Gomez, and C. Dorronsoro, “Concealed object detection and segmentation over millimetric waves images,” in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2010, pp. 31–37.
- [5] D. Gutiérrez and D. I. Escalona-Vargas, “EEG data classification through signal spatial redistribution and optimized linear discriminants,” *Computer Methods and Programs in Biomedicine*, vol. 97, no. 1, pp. 39–47, 2010.
- [6] S. Ryali, K. Supekar, T. Chen, and V. Menon, “Multivariate dynamical systems models for estimating causal interactions in fMRI,” *NeuroImage*, vol. 54, no. 2, pp. 807–823, 2011.
- [7] N. D. Lawrence, “Gaussian process latent variable models for visualisation of high dimensional data,” in *Advances in Neural Information Processing Systems*, 2004.
- [8] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, “Discriminative learned dictionaries for local image analysis,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2008*, 2008, pp. 1–8.
- [9] I. Ramirez, P. Sprechmann, and G. Sapiro, “Classification and clustering via dictionary learning with structured incoherence and shared features,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2010*, 2010, pp. 3501–3508.
- [10] Y. LeCun and Y. Bengio, “Convolutional networks for images, speech, and time series,” in *The Handbook of Brain Theory and Neural Networks*, Michael A. Arbib, Ed., pp. 255–258. MIT Press, 1998.
- [11] B. De Vries and J. C. Principe, “The gamma model—A new neural model for temporal processing,” *Neural Networks*, vol. 5, pp. 565–576, 1992.
- [12] S. Lawrence, A. C. Tsoi, and A. D. Back, “The gamma MLP for speech phoneme recognition,” in *Advances in Neural Information Processing Systems*. 1996, pp. 785–791, MIT Press.
- [13] R. Flamary, D. Tuia, B. Labbe, G. Camps-Valls, and A. Rakotomamonjy, “Large margin filtering,” *IEEE Transactions on Signal Processing*, vol. 60, no. 2, pp. 648–659, 2012.
- [14] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer, 2007.
- [15] G. Bouchard, “Efficient bounds for the softmax function, applications to inference in hybrid models,” in *2007 Neural Information Processing Systems Conference, NIPS 2007*, 2007, vol. 6239.
- [16] P. Castracane, F. Iavarone, S. Mica, E. Sottile, C. Vignola, and O. Arino, “Monitoring urban sprawl and its trends with EO data. UrbEx, a prototype national service from a WWF-ESA joint effort,” in *2nd GRSS/ISPRS Joint Workshop on Remote Sensing and Data Fusion over Urban Areas*, 2003, pp. 245–248.

Chapter 5

Multispectral Image Classification (II). Active Learning

5.1 A Bayesian Active Learning Framework for a Two-Class Classification Problem

- **P. Ruiz**, J. Mateos, R. Molina, and A.K. Katsaggelos, “A Bayesian Active Learning Framework for a Two-Class Classification Problem” in *MUSCLE International Workshop on Computational Intelligence for Multimedia Understanding*, edited by Emanuele Salerno, A. Enis Çetin and Ovidio Salvetti, vol. LNCS-7252, 42-53, Pisa (Italy), 2012.

– Status: Published

A Bayesian Active Learning Framework for a Two-Class Classification Problem

Pablo Ruiz¹, Javier Mateos¹, Rafael Molina¹, and Aggelos K. Katsaggelos²

¹ University of Granada, 18071 Granada, Spain

`mataran@decsai.ugr.es`,

WWW home page: <http://decsai.ugr.es/vip>

² Northwestern University, Evanston, IL, USA

Abstract. In this paper we present an active learning procedure for the two-class supervised classification problem. The utilized methodology exploits the Bayesian modeling and inference paradigm to tackle the problem of kernel-based data classification. This Bayesian methodology is appropriate for both finite and infinite dimensional feature spaces. Parameters are estimated, using the kernel trick, following the evidence Bayesian approach from the marginal distribution of the observations. The proposed active learning procedure uses a criterion based on the entropy of the posterior distribution of the adaptive parameters to select the sample to be included in the training set. A synthetic dataset as well as a real remote sensing classification problem are used to validate the followed approach.

1 Introduction

In many real applications large collections of data are extracted whose class is unknown. Those applications include, for instance, most image classification applications, text processing, speech recognition, and biological research problems. While extracting the samples is straightforward and inexpensive, classifying each one of those samples is a tedious and often expensive task. Active learning is a supervised learning technique that attempts to overcome the labeling bottleneck by asking queries in the form of unlabeled samples to be labeled by an *oracle* (e.g., a human annotator) [10]. An active learning procedure queries only the most informative samples from the whole set of unlabeled samples. The objective is to obtain a high classification performance using as few labeled samples as possible, minimizing, this way, the cost of obtaining labeled data.

Kernel methods in general and Support Vector Machines (SVMs) in particular dominate the field of discriminative data classification [8]. This problem has also been approached from a Bayesian point of view. For example, the relevance vector machine [13] assumes a Gaussian prior over the adaptive parameters and uses the EM algorithm to estimate them. In practice, this prior enforces sparsity because the posterior distribution of many adaptive parameters is sharply peaked around zero. Lately, Gaussian Process Classification [7] has received much attention. Adopting the least-squares SVM formulation may alternatively allow to

perform Bayesian inference on SVMs [12]. A huge benefit is obtained by applying Bayesian inference on these machines since hyperparameters may be learned directly from data using a consistent theoretical framework.

In this paper we make use of the Bayesian paradigm to tackle the problem of active learning on kernel-based two-class data classification. The Bayesian modeling and inference approach to the kernel-based classification we propose in this paper allows us to derive efficient closed-form expressions for parameter estimation and active learning.

The general two-class supervised classification problem [2] we tackle here implies a classification function of the form:

$$y(\mathbf{x}) = \boldsymbol{\phi}^\top(\mathbf{x})\mathbf{w} + b + \epsilon, \quad (1)$$

where the mapping $\boldsymbol{\phi} : \mathcal{X} \rightarrow \mathcal{H}$ embeds the observed $\mathbf{x} \in \mathcal{X}$ into a higher L -dimensional (possibly infinite) feature space \mathcal{H} . The output $y(\mathbf{x}) \in \{0, 1\}$ consists of a binary coding representation of its classification, \mathbf{w} is a vector of size $L \times 1$ of adaptive parameters to be estimated, b represents the bias in the classification function, and ϵ is an independent realization of the Gaussian distributions $\mathcal{N}(0, \sigma^2)$.

While kernel-based classification in *static* scenarios has been extensively studied, the problem related to the emerging field of *active learning* [10] is still unsolved. Let us assume that we have access to P vectors in the feature space denoted by $\boldsymbol{\phi}(\mathbf{x}_i), i = 1, \dots, P$ for which the corresponding output $y(\mathbf{x}_i), i = 1, \dots, P$ can be provided by an oracle. The key is to decide which elements \mathbf{x}_i to acquire from the set of P possible samples in order to build an optimal compact classifier. Active learning aims at efficiently sampling the observations space to improve the model performance by *incrementally* building training sets. Such sets are obtained by selecting from the available samples the best ones according to a selection strategy and querying the oracle only for the label of those samples. Many selection strategies have been devised in the literature, which are based on different heuristics: 1) large margin, 2) expert committee, and 3) posterior probability (see [10] for a comprehensive review). The first two approaches typically exploit SVM methods. The latter requires classifiers that can provide posterior probabilities.

In [6], a Bayesian active learning procedure for finite dimensional feature spaces is proposed. Assuming that $\boldsymbol{\phi}(\mathbf{x}_i), i = 1, \dots, P$ has L components, the design matrix $\boldsymbol{\Phi}_{:,}$ is of size $P \times L$, whose i^{th} row, $i = 1, \dots, P$ is given by $\boldsymbol{\phi}(\mathbf{x}_i)^\top$. Then, a subset of size C of the L columns of $\boldsymbol{\Phi}_{:,}$, denoted by $\boldsymbol{\Phi}_{:,I_C}$, is selected using the differential entropy instead of the response functions $y(\mathbf{x}_i)$ [6]. Notice that this approach is in contrast to other basis selection techniques which make explicit use of the response functions, for example, [3] in the context of SVM, [4] in the context of sparse representation, and [1] considering compressive sensing. To select the rows of $\boldsymbol{\Phi}_{:,I_C}$, for which the response associated to $\boldsymbol{\phi}(\mathbf{x}_i)$ will be queried, a criterion based again on differential entropy is utilized (see [6] for details). See also [5] for the general theory and [9] for the use of the approach in compressive sensing.

Here, the Bayesian modeling and inference paradigm is applied to two-class classification problems which utilize kernel-based classifiers. This paradigm is used to tackle both active learning and parameter estimation for infinite dimensional feature spaces, and consequently for problems where basis selection cannot be carried out explicitly. As we will see later, the proposed approach will make extensive use of the marginal distribution of the observations to avoid dealing with infinite dimensional feature spaces and the posterior distribution of the infinite dimensional \mathbf{w} .

The rest of the paper is organized as follows. Section 2 introduces the models we use in our Bayesian framework. Then, in section 3, Bayesian inference is performed. We calculate the posterior distribution of \mathbf{w} , and propose a methodology for parameter estimation, active learning, and class prediction. Experiments illustrating the performance of the proposed approach on a synthetic and a real remote sensing classification problem are presented in section 4. Finally, section 5 concludes the paper.

2 Bayesian modeling

Let us assume that the target variable $y(\mathbf{x}_i)$ follows the model in Eq. (1). If we already know the classification output $y(\mathbf{x}_i)$ associated with the feature samples $\phi(\mathbf{x}_i)$, $i = 1, \dots, M$, with M the number of samples, we can then write

$$p(\mathbf{y}|\mathbf{w}, \sigma^2) = \prod_{i=1}^M \mathcal{N}(y(\mathbf{x}_i) | \phi^\top(\mathbf{x}_i)\mathbf{w} + b, \sigma^2). \quad (2)$$

Since \mathbf{x}_i , $i = 1, \dots, M$, will always appear as conditioning variable, for the sake of simplicity, we have removed the dependency on $\mathbf{x}_1, \dots, \mathbf{x}_M$ in the left-hand side of the equation. We note that, for infinite dimensional feature vectors $\phi(\mathbf{x}_i)$, \mathbf{w} is infinite dimensional.

The Bayesian framework allows us to introduce information about the possible value of \mathbf{w} in the form of a prior distribution. In this work we assume that each component of \mathbf{w} independently follows a Gaussian distribution $\mathcal{N}(0, \gamma^2)$. When the feature vectors are infinite dimensional, we will not make explicit use of this prior distribution but still we will be able to carry out parameter estimation and active learning tasks.

3 Bayesian inference

Bayesian inference extracts conclusions from the posterior distribution $p(\mathbf{w}|\mathbf{y}, \gamma^2, \sigma^2)$. The posterior distribution of \mathbf{w} is given by [2]

$$p(\mathbf{w}|\mathbf{y}, \gamma^2, \sigma^2) = \mathcal{N}(\mathbf{w} | \Sigma_{\mathbf{w}|\mathbf{y}, \gamma^2, \sigma^2} \sigma^{-2} \Phi^\top (\mathbf{y} - b\mathbf{1}), \Sigma_{\mathbf{w}|\mathbf{y}, \gamma^2, \sigma^2}), \quad (3)$$

where

$$\Sigma_{\mathbf{w}|\mathbf{y}, \gamma^2, \sigma^2} = (\sigma^{-2} \Phi^\top \Phi + \gamma^{-2} \mathbf{I})^{-1}$$

and Φ is the design matrix whose i^{th} row is $\phi(\mathbf{x}_i)^\top$.

It is important to note that we do not need to know the form of Φ explicitly to calculate this posterior distribution. We only need to know the Gram matrix $\mathbf{K} = \Phi\Phi^\top$, which is an $M \times M$ symmetric matrix with elements $\mathbf{K}_{nm} = k(\mathbf{x}_n, \mathbf{x}_m) = \phi(\mathbf{x}_n)^\top \phi(\mathbf{x}_m)$, which has to be a positive semidefinite matrix [8]. This leads to the construction of kernel functions $k(\mathbf{x}, \mathbf{x}')$ for which the Gram matrix \mathbf{K} is positive semidefinite for all possible choices of the set $\{\mathbf{x}_n\}$ [11]. Note that, even if Φ has an infinite number of columns, which correspond to the case of \mathbf{x}_i being an infinite dimensional feature vector, we can still calculate \mathbf{K} of size $M \times M$ by means of the kernel function. Note also that we are somewhat abusing the notation here because \mathbf{w} is infinite dimensional for infinite dimensional feature vectors.

3.1 Parameter Estimation

To estimate the values of γ^2 and σ^2 we use the Evidence Bayesian approach without any prior information on these parameters. According to it, we maximize the marginal distribution obtained by integrating out the vector of adaptive parameters \mathbf{w} . It can easily be shown, see for instance [2], that

$$p(\mathbf{y}|\gamma^2, \sigma^2) = \mathcal{N}(\mathbf{y}|b\mathbf{1}, \Sigma_{\mathbf{y}|\gamma^2, \sigma^2}), \quad (4)$$

where

$$\Sigma_{\mathbf{y}|\gamma^2, \sigma^2} = \gamma^2 \Phi\Phi^\top + \sigma^2 \mathbf{I}.$$

The value of b can be easily obtained from Eq. (4) as

$$b = \frac{1}{M} \sum_{i=1}^M y(\mathbf{x}_i). \quad (5)$$

Differentiating $2 \ln p(\mathbf{y}|\gamma^2, \sigma^2)$ with respect to γ^2 and equating to zero, we obtain

$$\begin{aligned} \text{tr}[(\gamma^2 \Phi\Phi^\top + \sigma^2)^{-1} \Phi\Phi^\top] = \\ \text{tr}[(\mathbf{y} - b\mathbf{1})^\top (\gamma^2 \Phi\Phi^\top + \sigma^2 \mathbf{I})^{-1} \Phi\Phi^\top (\gamma^2 \Phi\Phi^\top + \sigma^2 \mathbf{I})^{-1} (\mathbf{y} - b\mathbf{1})]. \end{aligned} \quad (6)$$

Diagonalizing $\Phi\Phi^\top$, we obtain $\mathbf{U}\Phi\Phi^\top\mathbf{U}^\top = \mathbf{D}$, where \mathbf{U} is an orthonormal matrix and \mathbf{D} is a diagonal matrix with entries $\lambda_i, i = 1, \dots, M$. We can then rewrite the above equation as

$$\sum_{k=1}^M \frac{\lambda_k}{\gamma^2 \lambda_k + \sigma^2} = \sum_{i=1}^M z_i^2 \frac{\lambda_i}{(\gamma^2 \lambda_i + \sigma^2)^2}, \quad (7)$$

where $\mathbf{U}(\mathbf{y} - b\mathbf{1}) = \mathbf{z}$ with components $z_i, i = 1, \dots, M$.

Multiplying both sides of the above equation by γ^2 we have

$$\gamma^2 = \sum_{i=1}^M \frac{\frac{\lambda_i}{\gamma^2 \lambda_i + \sigma^2}}{\sum_{k=1}^M \frac{\lambda_k}{\gamma^2 \lambda_k + \sigma^2}} \frac{\gamma^2 z_i^2}{\gamma^2 \lambda_i + \sigma^2} = \sum_{i=1}^M \mu_i \frac{\gamma^2 z_i^2}{\gamma^2 \lambda_i + \sigma^2}, \quad (8)$$

where

$$\mu_i = \frac{\frac{\lambda_i}{\gamma^2 \lambda_i + \sigma^2}}{\sum_{k=1}^M \frac{\lambda_k}{\gamma^2 \lambda_k + \sigma^2}}. \quad (9)$$

Note that $\mu_i \geq 0$ and $\sum_{i=1}^M \mu_i = 1$.

Similarly, differentiating $2 \ln p(\mathbf{y}|\gamma^2, \sigma^2)$ with respect to σ^2 and equating it to zero, we obtain

$$\sum_{k=1}^M \frac{1}{\gamma^2 \lambda_k + \sigma^2} = \sum_{i=1}^M z_i^2 \frac{1}{(\gamma^2 \lambda_i + \sigma^2)^2}. \quad (10)$$

Following the same steps we already performed to estimate γ^2 , we obtain

$$\sigma^2 = \sum_{i=1}^M \nu_i \frac{\sigma^2 z_i^2}{\gamma^2 \lambda_i + \sigma^2}, \quad (11)$$

where

$$\nu_i = \frac{1}{\frac{\gamma^2 \lambda_i + \sigma^2}{\sum_{k=1}^M \frac{1}{\gamma^2 \lambda_k + \sigma^2}}}. \quad (12)$$

Note that, again, $\nu_i \geq 0$ and $\sum_{i=1}^M \nu_i = 1$.

To obtain estimates of γ^2 and σ^2 we use an iterative procedure where the values of the old estimates of γ^2 and σ^2 are used on the right hand side of Equations (8) and (11) to obtain the updated values of the parameters in the left hand side of these equations. Although we have not formally established the convergence and unicity of the solution, we have not observed any convergence problems in the performed experiments. Note that to estimate γ^2 and σ^2 we have not made use of the posterior distribution of the components of \mathbf{w} .

3.2 Active Learning

Active learning starts with a small set of observations whose class is already known. From these observations, the posterior distribution of \mathbf{w} and the parameters b , γ^2 and σ^2 can be estimated using the procedure described in the previous sections. Now we want that the system learns new observations incrementally. Let us assume that we want to add a new observation associated to $\phi(\mathbf{x}_+)$, whose corresponding $y(\mathbf{x}_+)$ will be learned by querying the oracle. The covariance matrix of the posterior distribution of \mathbf{w} when $\phi(\mathbf{x}_+)$ is added is given by

$$\Sigma_{\mathbf{w}|\mathbf{y}, \gamma^2 \sigma^2}^{\mathbf{x}_+} = (\sigma^{-2}(\Phi^\top \Phi + \phi(\mathbf{x}_+) \phi^\top(\mathbf{x}_+)) + \gamma^{-2} \mathbf{I})^{-1}.$$

Since we have a set of observations that could be added and whose class is unknown (but can be learned by querying the oracle), the objective of active learning is to select the observation that maximizes the performance of the system, minimizing in this way the number of queries answered by the oracle. To

select this new feature vector, in this paper, we propose to maximize the difference between the entropies of the posterior distribution before and after adding the new feature vector (see [6, 9]) to obtain

$$\mathbf{x}_+ = \arg \max_{\mathbf{x}} \frac{1}{2} \log |\boldsymbol{\Sigma}_{\mathbf{w}|\mathbf{y}, \gamma^2, \sigma^2}| |\boldsymbol{\Sigma}_{\mathbf{w}|\mathbf{y}, \gamma^2, \sigma^2}^{\mathbf{x}}|^{-1}. \quad (13)$$

Then we have

$$\begin{aligned} & \frac{1}{2} \log |\boldsymbol{\Sigma}_{\mathbf{w}|\mathbf{y}, \gamma^2, \sigma^2}| |\boldsymbol{\Sigma}_{\mathbf{w}|\mathbf{y}, \gamma^2, \sigma^2}^{\mathbf{x}}|^{-1} \\ &= \frac{1}{2} \log |\mathbf{I} + \sigma^{-2} \phi(\mathbf{x}) \phi^\top(\mathbf{x}) (\sigma^{-2} \boldsymbol{\Phi}^\top \boldsymbol{\Phi} + \gamma^{-2} \mathbf{I})^{-1}| \\ &= \frac{1}{2} \log (1 + \sigma^{-2} \phi^\top(\mathbf{x}) (\sigma^{-2} \boldsymbol{\Phi}^\top \boldsymbol{\Phi} + \gamma^{-2} \mathbf{I})^{-1} \phi(\mathbf{x})), \end{aligned}$$

and using

$$(\sigma^{-2} \boldsymbol{\Phi}^\top \boldsymbol{\Phi} + \gamma^{-2} \mathbf{I})^{-1} = \gamma^2 \mathbf{I} - \gamma^4 \boldsymbol{\Phi}^\top (\sigma^2 \mathbf{I} + \gamma^2 \boldsymbol{\Phi} \boldsymbol{\Phi}^\top)^{-1} \boldsymbol{\Phi}, \quad (14)$$

we can finally write

$$\begin{aligned} & \frac{1}{2} \log |\boldsymbol{\Sigma}_{\mathbf{w}|\mathbf{y}, \gamma^2, \sigma^2}| \cdot |\boldsymbol{\Sigma}_{\mathbf{w}|\mathbf{y}, \gamma^2, \sigma^2}^{\mathbf{x}}|^{-1} \\ &= \frac{1}{2} \log (1 + \sigma^{-2} \gamma^2 \phi^\top(\mathbf{x}) \phi(\mathbf{x}) - \sigma^{-2} \gamma^4 \phi^\top(\mathbf{x}) \boldsymbol{\Phi}^\top (\sigma^2 \mathbf{I} + \gamma^2 \boldsymbol{\Phi} \boldsymbol{\Phi}^\top)^{-1} \boldsymbol{\Phi} \phi(\mathbf{x})) \\ &= \frac{1}{2} \log \left(1 + \sigma^{-2} \gamma^2 \phi^\top(\mathbf{x}) \phi(\mathbf{x}) - \sigma^{-2} \gamma^4 \phi^\top(\mathbf{x}) \boldsymbol{\Phi}^\top \boldsymbol{\Sigma}_{\mathbf{y}|\gamma^2, \sigma^2}^{-1} \boldsymbol{\Phi} \phi(\mathbf{x}) \right). \end{aligned} \quad (15)$$

Consequently, all needed quantities to select \mathbf{x}_+ can be calculated without knowledge of the feature vectors and the posterior distribution of the possibly infinite dimensional adaptive parameters and using only kernel functions and the marginal distribution of the observations.

Notice that, given $\boldsymbol{\Sigma}_{\mathbf{y}|\gamma^2, \sigma^2}^{-1}$, we can easily calculate the new precision matrix $\boldsymbol{\Sigma}_{\mathbf{y}, \mathbf{y}(\mathbf{x}_+)|\gamma^2, \sigma^2}^{-1}$ of the marginal distribution of \mathbf{y} when the observation corresponding to \mathbf{x}_+ has been added. We have

$$\boldsymbol{\Sigma}_{\mathbf{y}, \mathbf{y}(\mathbf{x}_+)|\gamma^2, \sigma^2}^{-1} = \begin{pmatrix} \mathbf{M} & -\mathbf{M} \mathbf{v} d^{-1} \\ -d^{-1} \mathbf{v}^\top \mathbf{M} & d^{-1} + d^{-2} \mathbf{v}^\top \mathbf{M} \mathbf{v} \end{pmatrix}, \quad (16)$$

with $\mathbf{v} = \gamma^2 \boldsymbol{\Phi} \phi(\mathbf{x}_+)$, $d = \sigma^2 + \gamma^2 \phi^\top(\mathbf{x}_+) \phi(\mathbf{x}_+)$, and $\mathbf{M} = (\boldsymbol{\Sigma}_{\mathbf{y}|\gamma^2, \sigma^2} - d^{-1} \mathbf{v} \mathbf{v}^\top)^{-1}$.

To calculate \mathbf{M} we use the Sherman-Morrison-Woodbury formula to obtain

$$\mathbf{M} = \boldsymbol{\Sigma}_{\mathbf{y}|\gamma^2, \sigma^2}^{-1} - \frac{1}{-d + \mathbf{v}^\top \boldsymbol{\Sigma}_{\mathbf{y}|\gamma^2, \sigma^2}^{-1} \mathbf{v}} \boldsymbol{\Sigma}_{\mathbf{y}|\gamma^2, \sigma^2}^{-1} \mathbf{v} \mathbf{v}^\top \boldsymbol{\Sigma}_{\mathbf{y}|\gamma^2, \sigma^2}^{-1},$$

and consequently $\boldsymbol{\Sigma}_{\mathbf{y}, \mathbf{y}(\mathbf{x}_+)|\gamma^2, \sigma^2}^{-1}$ can be calculated from the previous $\boldsymbol{\Sigma}_{\mathbf{y}|\gamma^2, \sigma^2}^{-1}$ in a straightforward manner.

Hence, starting with an initial estimation of the parameters, to perform active learning we alternate between the selection of a new sample using Eq. (13) and the estimation of the unknown parameters b , γ^2 , and σ^2 using the procedure described in section 3.1.

3.3 Prediction

Once the system has been trained, we want to predict the value of $y(\mathbf{x}_*)$ for a new value of \mathbf{x} , denoted by \mathbf{x}_* . To calculate this predicted value, we make use of the distribution of $\phi^\top(\mathbf{x}_*)\mathbf{w} + b$ where the posterior distribution of \mathbf{w} is given in Eq. (3). Its mean value, $\phi^\top(\mathbf{x}_*)\mathbb{E}[\mathbf{w}] + b$, is given by

$$\phi^\top(\mathbf{x}_*)\mathbb{E}[\mathbf{w}] + b = \phi^\top(\mathbf{x}_*)\Sigma_{\mathbf{w}|\mathbf{y},\gamma^2,\sigma^2}\sigma^{-2}\Phi^\top(\mathbf{y} - b\mathbf{1}) + b, \quad (17)$$

where we have made use of Eq. (14) to obtain

$$\begin{aligned} \phi^\top(\mathbf{x}_*)\mathbb{E}[\mathbf{w}] + b &= \gamma^2\sigma^{-2}\phi^\top(\mathbf{x}_*)\Phi^\top(\mathbf{y} - b\mathbf{1}) \\ &\quad - \gamma^4\sigma^{-2}\phi^\top(\mathbf{x}_*)\Phi^\top(\sigma^2\mathbf{I} + \gamma^2\Phi\Phi^\top)^{-1}\Phi\Phi^\top(\mathbf{y} - b\mathbf{1}) + b, \end{aligned} \quad (18)$$

which can be calculated without knowing the feature vectors if the kernel function is known.

4 Experimental Results

We have tested the proposed active learning algorithm on a synthetic dataset and a real remote sensing classification problem. The synthetic data set, due to Paisley [6], consists of 200 observations, 100 from each one of the two classes, in a bi-dimensional space. The data, plotted in figure 1, is composed of two classes defined by two manifolds, which are not linearly separable in this bi-dimensional space.

We have compared the proposed active learning method with random sampling and the recently proposed Bayesian method in [6]. Random sampling was implemented using the proposed method but, instead of selecting the samples according to Eq. (13), samples are selected randomly from the available training set. In all cases, a Gaussian kernel was used, whose optimal width parameter was selected by maximizing the standard cross-validation accuracy.

We divided the full set of 200 samples into two disjoint sets of 100 randomly selected samples each, one for training and the other for testing. We started our active learning process with a seed, a single labeled sample, randomly selected from the data set, that is, $M = 1$ at the beginning and the rest of the training set was used to simulate the oracle queries. We run the three algorithms for 99 iterations adding one sample at each iteration, that is, querying the oracle one sample each time so, at the end, $M = 100$. To obtain meaningful results, the process was repeated 10 times with different randomly selected training and test sets.

The performance of the algorithms is measured utilizing the samples in the test set using the mean confusion matrix, the mean overall accuracy (OA) and OA variance, and the mean kappa index. Each cell (i, j) of the mean confusion matrix contains the mean number of samples, over the ten executions of the algorithms using the different training and test sets, belonging to the j -th class, classified in the i -th class. The overall accuracy is the proportion of correctly

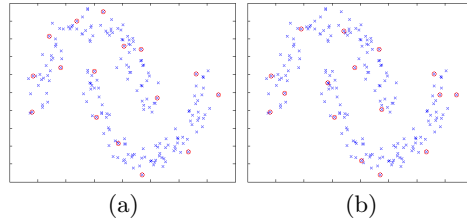


Fig. 1. First 15 selected samples for (a) the method in [6] and (b) the proposed method.

classified samples over the total number of samples. The mean OA averages the ten OA results of the ten different algorithm executions. The variance of the OA in all the executions is reported as OA variance. The kappa index is a statistical measure, which reflects agreement between the obtained accuracy and the accuracy that would be expected by randomly classifying the samples. Unlike the Overall Accuracy, the kappa index avoids the chance effect. A value of the kappa index greater than 0.8 is considered to be "very good". Since ten runs of the algorithm are performed, the mean kappa over all the executions index is used.

In Figure 1 we show the first 15 selected samples for the method in [6] and the proposed method. It can be seen that both algorithms select samples that efficiently represent the two manifolds. Figure 2 shows the average learning curves for random sampling, the method in [6] and the proposed method. From the figure, it is clear that random sampling provides the lowest convergence rate, while the method in [6] and the proposed method have a similar learning rate to the full set overall accuracy. At convergence, when 100 samples are included in the training set, all methods have the same accuracy but the proposed method reaches this value with 18.4 samples on the average while the method in [6] needs 28.2 samples and random sampling needs 36.4 samples.

In the second experiment a real remote sensing dataset was used. Satellite or airborne mounted sensors usually capture a set of images of the same area in several wavelengths or spectral channels forming a multispectral image. This multispectral image allows for the classification of the pixels in the scene into different classes to obtain classification maps used for management, policy making and monitoring. A critical problem in remote sensing image segmentation is that few labeled pixels are typically available: in such cases, active learning may be very helpful [14].

We evaluated the methods on a real Landsat 5 TM image, whose RGB bands are depicted in Fig. 3a. The region of interest is a 1024×1024 pixels area centered in the city of Granada, in the south of Spain. The Landsat TM sensor provides a six bands multispectral image that covers RGB, near-infrared and mid-infrared ranges with a spatial resolution of 30 meters per pixel, that is, each pixels captures the energy reflected by the Earth in a square area of side equal to 30 meters. The dataset, created by the RSGIS Laboratory at the University of Granada, divides the scene into two classes, vegetation and no-vegetation. Note

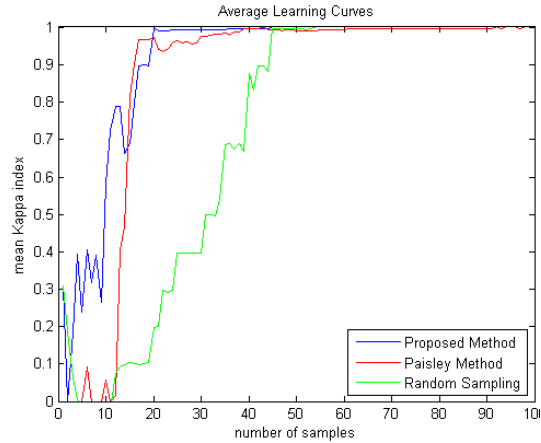


Fig. 2. Average learning curves for the active learning techniques using random sampling, the Bayesian method in [6] (Paisley method), and the proposed method for the synthetic experiment.

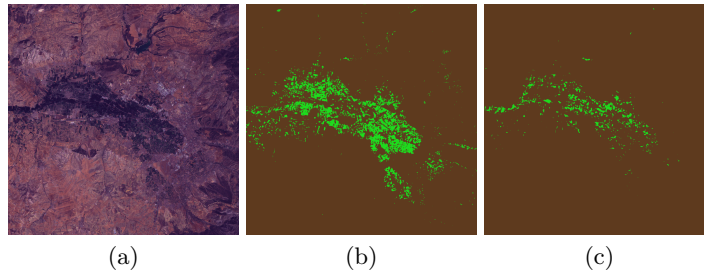


Fig. 3. (a) Multispectral image, (b) classification map with the proposed method, and (c) classification map with the method in [6]. Pixels classified as vegetation are shown in green color and pixels classified as no-vegetation are shown in brown.

that the no-vegetation class includes bare soil that has a very similar spectral signature to vegetation making the correct classification of the pixels a challenging problem.

A total of 336 samples, whose class is precisely known by visual inspection of the images and by terrain inspection, were selected from the image, 174 samples corresponding to the vegetation class and 162 samples corresponding to the no-vegetation class. Each sample has six characteristics, each one corresponding to the mean value of a 3×3 area centered in the pixel under study for each one of the six bands that comprise the multispectral information provided by the Landsat TM satellite. Again, the same Gaussian kernel was used for all methods.

From the labeled dataset a test set of 150 samples was randomly selected, and the remaining 186 samples were used to simulate the oracle queries. We run the experiments 10 times with different training and test sets. All the algorithms

Table 1. Mean confusion matrix, mean kappa index, mean overall accuracy and its variance for ten runs of the method in [6] on different test sets.

Predicted/actual	vegetation	no-vegetation	Mean Kappa = 0.9453
vegetation	74.4	3.5	Mean OA = 97.27%
no-vegetation	0.6	71.5	OA variance = 4.39×10^{-5}

Table 2. Mean confusion matrix, mean kappa index, mean overall accuracy and its variance for ten runs of the proposed method on different test sets.

Predicted/actual	vegetation	no-vegetation	Mean Kappa = 0.96
vegetation	74.4	2.4	Mean OA = 98.00%
no-vegetation	0.6	72.6	OA variance = 9.87×10^{-5}

were run for 185 iterations, starting from a training set with a single labeled pixel, that is $M = 1$, and adding one pixel to the training set at each iteration (query).

Again, the proposed method is compared with random sampling and the Bayesian method in [6]. For the method in [6] we did not perform the basis selection step. We want to note that, since this basis selection procedure discards features from the samples, better results are expected when all the features are used although the computational cost will be higher.

Figure 4 shows the average learning curves. The method in [6] provides a lower convergence rate to the full set overall accuracy than the proposed method. However, the method in [6] starts learning faster than the proposed one. It may be due to the fact that the active learning is carried out in an M -dimensional feature space while the proposed method works in an infinite-dimensional space. However, at convergence, when 186 samples have been included in the training set, the proposed method performs better than the method in [6]. Note also that, at convergence, random sampling obtains the same results with the proposed method, obtaining better classification accuracy than the method in [6]. This was expected since it uses the same classification procedure as the proposed method, except for the active learning selection procedure. Note, however, that the convergence rate is much slower than the other two methods.

Figures 3b and 3c depict the classification map for the full image using the proposed method and the method in [6]. The random sampling classification is not shown since, at convergence, coincides with the proposed method. The mean of the confusion matrices as well as the mean kappa index, the mean overall accuracy, and the overall accuracy variance are shown in Tables 1 and 2, for the method in [6] and the proposed method, respectively. From these figures of merit it is clear that the proposed method discriminates better between vegetation and no-vegetation than the method in [6].

All compared methods were implemented using Matlab[©] and run on a Intel i7 @ 2.67GHz. The proposed method took 1.23 sec to complete the 185 iterations while the method in [6] took 48.44 sec and random sampling took 1.01 sec. It is worth noting that computing the precision matrix $\Sigma_{y,y(\mathbf{x}_+)}^{-1} | \gamma^2, \sigma^2$ in Eq. (16)

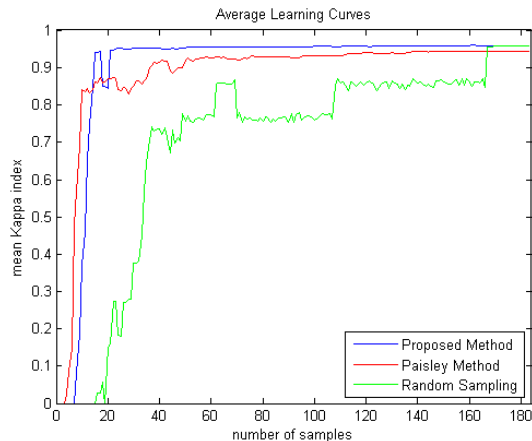


Fig. 4. Learning curve for the active learning techniques using random sampling, the Bayesian method in [6] (Paisley method), and the proposed method for the real remote sensing dataset.

takes most of the time, which explains the similar cost between the proposed method and random sampling. It is worth noting that the proposed method provided better figures of merit than the method in [6] for both mean kappa index and mean overall accuracy, learning with less interaction with the oracle and, also, with a much lower computational cost.

5 Conclusions

We presented an active learning procedure that exploits Bayesian learning and parameter estimation to tackle the problem of two-class kernel-based data classification. Using the Bayesian modeling and inference, we developed a Bayesian method for classification both finite and infinite dimensional feature spaces. The proposed method allows us to derive efficient closed-form expressions for parameter estimation and incremental and active learning. The method was experimentally compared to other methods and its performance was assessed on remote sensing multispectral image as well as synthetic data.

Acknowledgments

This work has been supported by the Spanish research programme Consolider Ingenio 2010: MIPRCV (CSD2007-00018) and the “Consejería de Innovación, Ciencia y Empresa of the Junta de Andalucía” under contract P07-TIC-02698. We want to thank V. F. Rodríguez-Galiano and Prof. M. Chica from the RSGIS laboratory (Group RNM122 of the Junta de Andalucía), who are supported by the Spanish MICINN (CGL2010-17629), for the image of the neighborhood of

the city of Granada and the classified samples that conformed the real dataset used in this paper.

References

1. Babacan, D., Molina, R., Katsaggelos, A.: Bayesian compressive sensing using Laplace priors. *IEEE Transactions on Image Processing* 19(1), 53–63 (2010)
2. Bishop, C.M.: *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer (2007)
3. Burges, C.J.: A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery* 2, 121–167 (1998)
4. Elad, M.: *Sparse and Redundant Representations - From Theory to Applications in Signal and Image Processing*. Springer (2010)
5. MacKay, D.J.C.: Information-based objective functions for active data selection. *Neural Computation* 4(4), 590–604 (1992)
6. Paisley, J., Liao, X., Carin, L.: Active learning and basis selection for kernel-based linear models: A Bayesian perspective. *IEEE Transactions on Signal Processing* 58, 2686–2700 (2010)
7. Rasmussen, C.E., Williams, C.K.: *Gaussian Processes for Machine Learning*. MIT Press, NY (2006)
8. Schölkopf, B., Smola, A.: *Learning with Kernels*. MIT Press, Cambridge, MA (2002)
9. Seeger, M.W., Nickisch, H.: Compressed sensing and Bayesian experimental design. In: *International Conference on Machine Learning* 25 (2008)
10. Settles, B.: *Active learning literature survey*. Computer Sciences Technical Report 1648, University of Wisconsin–Madison (2009)
11. Shawe-Taylor, J., Cristianini, N.: *Kernel Methods for Pattern Analysis*. Cambridge Univ. Press (2004)
12. Suykens, J.A.K., Van Gestel, T., De Brabanter, J., De Moor, B., Vandewalle, J.: *Least Squares Support Vector Machines*. World Scientific, Singapore (2002)
13. Tipping, M.E.: The relevance vector machine. *Journal of Machine Learning Research* 1, 211–244 (2001)
14. Tuia, D., Volpi, M., Copa, L., Kanevski, M., Muñoz-Marí, J.: A survey of active learning algorithms for supervised remote sensing image classification. *IEEE J. Sel. Topics Signal Proc.* 4, 606–617 (2011)

5.2 Bayesian Active Remote Sensing Image Classification

- **P. Ruiz**, J. Mateos, G. Camps-Valls, R. Molina, and A.K. Katsaggelos, “Bayesian Active Remote Sensing Image Classification”, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 4, 2186-2196, April 2014.
 - Status: Published
 - Impact Factor (JCR 2013): 2.933
 - Subject Category: Engineering, Electrical & Electronic (Q1: 31/248), Geochemistry & Geophysics (Q1: 17/80), Imaging Science & Photographic Technology (Q1: 3/23), Remote Sensing (Q1: 3/27)

Bayesian Active Remote Sensing Image Classification

Pablo Ruiz, *Student Member, IEEE*, Javier Mateos, Gustavo Camps-Valls, *Senior Member, IEEE*, Rafael Molina, *Member, IEEE*, and Aggelos K. Katsaggelos, *Fellow, IEEE*

Abstract—In recent years kernel methods and in particular support vector machines (SVMs) have been successfully introduced to remote sensing image classification. Their properties make them appropriate for dealing with high number of image features and low number of available labeled spectra. The introduction of alternative approaches based on (parametric) Bayesian inference has been quite scarce in the more recent years. Assuming a particular prior data distribution may lead to poor results in remote sensing problems because of the specificities and complexity of the data. In this context, the emerging field of non-parametric Bayesian methods constitutes a proper theoretical framework to tackle the remote sensing image classification problem.

This paper exploits the Bayesian modeling and inference paradigm to tackle the problem of kernel-based remote sensing image classification. This Bayesian methodology is appropriate for both finite and infinite dimensional feature spaces. The particular problem of active learning is addressed by proposing an incremental/active learning approach based on three different approaches: the maximum differential of entropies, the minimum distance to decision boundary, and the minimum normalized distance. Parameters are estimated by using the evidence Bayesian approach, the kernel trick, and the marginal distribution of the observations instead of the posterior distribution of the adaptive parameters. This approach allows us to deal with infinite dimensional feature spaces. The proposed approach is tested on the challenging problem of urban monitoring from multispectral and synthetic aperture radar (SAR) data and in multiclass land cover classification of hyperspectral images, in both purely supervised and active learning settings. Similar results are obtained when compared to SVMs in supervised mode, with the advantage of providing posterior estimates for classification and automatic parameter learning. Comparison with random sampling, and standard active learning methods, such as margin sampling and entropy-query-by-bagging reveal a systematic overall accuracy gain and faster convergence with the number of queries.

Index Terms—Supervised classification, incremental/active learning, multispectral image segmentation, Bayesian inference

P. Ruiz, J. Mateos and R. Molina are with Dpt. Ciencias de la Computación e I. A. E.T.S. Ing. Informática y Telecomunicación. Universidad de Granada, 18071 Granada, Spain. (e-mail: mataran@decsai.ugr.es, jmd@decsai.ugr.es, rms@decsai.ugr.es).

G. Camps-Valls is with the Image Processing Laboratory (IPL), University of Valencia, Parc Científic Universitat de València, C/ Cat. A. Escardino, 46980 Paterna, València, Spain. (e-mail: gustavo.camps@uv.es).

A. K. Katsaggelos is with the Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL 60208 USA (e-mail: aggk@eecs.northwestern.edu).

This paper has been partially supported by the Spanish Ministry for Education and Science under projects, TIN2010-15137, EODIX/AYA2008-05965-C04-03 and CONSOLIDER/CSD2007-00018, and a grant from the U.S. Department of Energy (DE-NA0000457).

Manuscript received March, 2012; revised April, 2013.

I. INTRODUCTION

CURRENTLY, kernel methods in general and support vector machines (SVMs) in particular dominate the field of *discriminative* data classification models [1]. During the last years, the methods have been successfully introduced in the field of remote sensing image classification [2], [3]. Kernel methods deal efficiently with low-sized datasets of potentially high dimensionality, as in the case of hyperspectral images. The use of the kernel trick [4], as is known in the literature, allows kernel methods to work in higher dimensional (possibly infinite-dimensional) spaces requiring the knowledge of only a kernel function which calculates an inner product in the new space using the original data. Also, since kernel methods do not assume an explicit prior data distribution but are inherently non-parametric models, they cope well with remote sensing data specificities and complexities. Alternative Bayesian approaches to remote sensing processing problems also exist and have been introduced as well to Earth observation applications. For example, the relevance vector machine (RVM) [5] assumes a Gaussian prior over the weights to enforce sparsity and uses expectation-maximization to infer the parameters. In [6], [7], the RVM was used for multispectral image segmentation and landmine detection using ground penetrating radar, while in [8] the model was used for adaptive biophysical parameter retrieval. Lately, Gaussian Processes [9] have received much attention in the field of machine learning, and some applications and developments have been introduced in remote sensing data processing as well, both for classification [10], [11] and parameter retrieval [12] settings.

In this paper, we restrict ourselves to the classification problem. Due to the particular characteristics of remote sensing data, namely potentially high-dimensionality, low number of labeled samples and different noise sources, assuming a particular prior data distribution may lead to poor classification results. In this context, the emerging field of *non-parametric Bayesian methods* constitutes a proper theoretical framework to tackle the problem [13], [9], [14]¹. This paper follows a Bayesian modeling and inference paradigm to tackle the problem of *kernel-based* remote sensing image classification. This Bayesian methodology is appropriate for both finite and infinite dimensional feature spaces, and hence robustness to the aforementioned problems in remote sensing is achieved. In two-class classification problems, the goal is to estimate a function and use as decision boundary the points where

¹Excellent online lectures are available at: http://videlectures.net/mlss09uk_teh_nbm/ and http://videlectures.net/mlss09uk_orbanz_fnbm/

the function is zero, to decide whether a sample belongs to a given class. In its simplest form, and given a training set, this is equivalent to estimate a linear function on a transformed feature space to separate samples from both classes. SVMs approach this problem through the concept of margin which is defined as the smallest distance between the decision boundary and any of the samples. On the other hand, Bayesian modeling and inference approach the problem by introducing information on the hyperplane coefficients using a prior model which in combination with the likelihood of the labeled samples leads to both a posterior distribution of the hyperplane coefficients and a Bayesian classification procedure. The use of the Bayesian paradigm allows for the calculation of the uncertainty of the estimated parameters and also the determination of the certainty of the estimated label for a given sample. It also allows for the estimation of all the model parameters in a rigorous and sound manner.

Relations between SVMs and Bayesian inference is not new. Note that adopting the least-squares SVM formulation may alternatively allow to perform Bayesian inference on SVMs [15]. Bayesian inference on these machines yields some relevant benefits: hyperparameters may be learned directly from data using a consistent theoretical framework, and posterior probabilities for the predictions can be obtained. Consequently, non-parametric Bayesian methods may deal with uncertainties in the data and naturally allow us developing intuitive incremental/active learning methods. The presented Bayesian kernel-based classifier permits to derive efficient closed-form expressions for parameter estimation, as well as to perform incremental, adaptive and active learning in a consistent, principled way.

While kernel-based classification in *static* scenarios has been extensively studied, the problem of *on-line* and *incremental* classification is still unsolved. The most effective schemes so far make use of both incremental and online SVMs [16], [17], [18]. Most of these approaches are based on growing and pruning strategies to create and update a *dictionary* of (representative) support vectors. Unfortunately, the algorithms require tuning several heuristic parameters. Alternatively, Bayesian kernel machines, such as Gaussian processes, have been successfully reformulated to deal with online and sparse settings [19], [20]. These methods typically rely again on a sequential generation of datasets of relevant samples. Nevertheless the framework nicely allows for both a propagation of predictions and Bayesian error estimates.

The previous online/incremental approaches are actually related to the emerging field of *active* learning [21]. Active learning aims at building efficient training sets by *iteratively* improving the model performance through sampling. Many query strategies have been devised in the literature, which are based on different heuristics: 1) large margin, 2) expert committee, and 3) posterior probability (see [21] for a comprehensive review). The first approach typically exploits SVM methods, while the second one can be adopted by any classifier. The latter requires classifiers that can provide posterior probabilities. While Platt's solution [22] of including a sigmoid link in SVMs could do the job, some theoretical concerns have been raised about the true meaning of such

posteriors. In Bayesian active learning, the prior over the hypotheses space is updated after seeing new data. For example, in [23], the expected Kullback-Leibler divergence between the current and the revised posterior distributions is maximized, while in [24], the authors proposed a Bayesian framework to tackle the active learning problem, which is utilized in Remote sensing in [25]. In [26], a Bayesian framework is also used and the posterior distribution is obtained as a Multinomial Logistic Regression model. Other basis selection techniques make explicit use of the response functions [27], [28], [29]. See also [30] for the basis selection general theory and [31] for the use of the approach in compressive sensing.

The field of remote sensing image classification has experienced a growing interest in active learning. Most of the introduced methods rely on smart sampling strategies over the SVM margin [32], [33], [34]. Some alternative approaches to work with batches of selections per iteration have been presented, and mainly rely on the concept of *diversity* between candidate pixels [35], [33] or with respect to the current model [36], or both [37]. Recent papers deal with new applications of active learning algorithms: in [38], [39], active learning is used to select the most useful unlabeled pixels to train a *semisupervised* classifier, while in [11], [40] active queries are used to correct for dataset shift in different areas of images. A complete review of the field of active learning in remote sensing can be found in [41].

In this paper, the Bayesian modeling and inference paradigm is applied to kernel-based classifiers. This paradigm is used to tackle both passive and active learning, as well as to address the problem of parameter estimation for infinite dimensional feature spaces, and consequently for problems where basis selection cannot be carried out explicitly. The current work presents the novel introduction of nonparametric Bayesian learning for remote sensing image classification both in purely supervised and active learning settings. This approach proposes an iterative procedure to maximize the marginal of the observations and, to the best of our knowledge, this is the first paper where nonparametric Bayesian methods are used in Active Remote Sensing Images Classification. The presented methods actually go one step further by extending standard nonparametric large margin techniques, such as SVM, which are typically used for image segmentation applications. Non-parametric Bayesian modeling and inference paradigms are introduced here to tackle the problem of kernel-based remote sensing image classification with the resulting major advantage of automatically learning the values of the (hyper)parameters from the data and thus no ad hoc cross-validation tuning schemes are necessary. This Bayesian methodology is appropriate for both finite and infinite dimensional feature spaces. The particular problem of active learning is addressed by proposing an incremental/active learning approach based on three different approaches: the maximum differential of entropies, the minimum distance to decision boundary, and the minimum normalized distance. Comparison with random sampling and standard active learning methods, such as margin sampling, or entropy-query-by-bagging, reveals a systematic overall accuracy gain and faster convergence with the number of queries.

The remainder of the paper is organized as follows. Section II introduces the basic notation to perform Bayesian modeling. Section III presents the Bayesian inference framework proposed in this paper. We first introduce the basic tools and then the novel formulations for parameter estimation, active learning data classification and prediction. Section V illustrates the performance of the proposed method in multispectral image segmentation. Conclusions are outlined in Section VI.

II. PROBLEM STATEMENT AND BAYESIAN MODELING

Let us introduce the basic problem formulation and notation. Let n be the number of pixels of a d -dimensional hyperspectral image, $\{\mathbf{x}_i | i = 1, \dots, n\}$, $\mathbf{x} \in \mathbb{R}^d$ we want to classify. The general two-class supervised classification problem we tackle here defines a classification function of the form

$$y(\mathbf{x}) = \phi^\top(\mathbf{x})\mathbf{w} + b + \epsilon, \quad (1)$$

where the mapping $\phi : \mathcal{X} \rightarrow \mathcal{H}$ maps the observed data point (samples, spectra) $\mathbf{x} \in \mathcal{X}$ into a higher L -dimensional (possibly infinite) Hilbert feature space \mathcal{H} . Note that for a K -class problem, the decision function implies K independent classification functions of the form $y_k(\mathbf{x}) = \phi^\top(\mathbf{x})\mathbf{w}_k + b_k + \epsilon_k$, $k = 1, \dots, K$ [4].

For the sake of simplicity of the notation, we will focus here on the binary case. However, its extension to multiclass scenarios is straightforward². Therefore, for a data point, \mathbf{x} , the output $y(\mathbf{x}) \in \{0, 1\}$ consists of a binary coding representation of its classification as belonging to class \mathcal{C}_0 or \mathcal{C}_1 , respectively, \mathbf{w} is a vector of size $L \times 1$ of adaptive parameters to be estimated, b represents the bias in the classification function, and ϵ is an independent realization of the Gaussian distribution $\mathcal{N}(0, \sigma^2)$.

For a training set, we already know the classification output $y(\mathbf{x}_i)$ associated with the feature samples $\phi(\mathbf{x}_i)$, $i = 1, \dots, M$, with M the number of samples, and therefore we can write

$$p(\mathbf{y} | \mathbf{w}, b, \sigma^2) = \prod_{i=1}^M \mathcal{N}(y(\mathbf{x}_i) | \phi^\top(\mathbf{x}_i)\mathbf{w} + b, \sigma^2), \quad (2)$$

where $\mathbf{y} = (y(\mathbf{x}_1), y(\mathbf{x}_2), \dots, y(\mathbf{x}_M))^\top$. Since \mathbf{x}_i , $i = 1, \dots, M$, will always appear as conditioning variable, for the sake of simplicity, we have removed the dependency on $\mathbf{x}_1, \dots, \mathbf{x}_M$ in the left-hand side of the equation. We note that, for infinite dimensional feature vectors $\phi(\mathbf{x}_i)$, \mathbf{w} is infinite dimensional.

The Bayesian framework allows us to introduce information about the possible value of \mathbf{w} in the form of a prior distribution. Since the likelihood function defined in Eq. (2) is the exponential of a quadratic function of \mathbf{w} , its corresponding conjugate prior should be a Gaussian distribution [4] so that the posterior will also be Gaussian. In this work, we consider a particular form of the Gaussian prior in which each component

²Extension to multiclass problems can be accomplished in many different ways by following standard schemes: one-versus-all, one-versus-one, pure multiclass schemes, or even sophisticated puncturing alternatives. We suggest here the use of a one-versus-all scheme, which typically gives rise to simpler and highly competitive results [42].

of \mathbf{w} independently follows a Gaussian distribution $\mathcal{N}(0, \gamma^2)$. Notice that this distribution can also be obtained utilizing the Gaussian Process framework [4]. When the feature vectors are infinite dimensional, we will not make explicit use of this prior distribution but still we will be able to carry out parameter estimation, prediction, and active learning tasks.

III. PROPOSED BAYESIAN INFERENCE METHOD

Due to the possible use of infinite dimensional feature spaces we will mainly use the marginal distribution of the observations to perform inference tasks, that is, parameter estimation, prediction and active learning and avoid, when possible, the use of the posterior distribution of the adaptive parameters, \mathbf{w} , since it cannot be calculated for infinite dimensional spaces. However, when a finite dimensional space is used, we will also calculate the posterior distribution in this section.

A. Marginal Distribution of \mathbf{y}

The marginal distribution of \mathbf{y} can be obtained by integrating out the vector of adaptive parameters \mathbf{w} . It can easily be shown, see for instance [4], that

$$p(\mathbf{y} | b, \gamma^2, \sigma^2) = \mathcal{N}(\mathbf{y} | b\mathbf{1}, \mathbf{C}), \quad (3)$$

with

$$\mathbf{C} = \gamma^2 \Phi \Phi^\top + \sigma^2 \mathbf{I}, \quad (4)$$

where Φ is the design matrix whose i -th row is $\phi^\top(\mathbf{x}_i)$, and $\mathbf{1}$ is a column vector with all its M components equal to 1.

It is important to note that we do not need to know the form of Φ explicitly to calculate this marginal distribution. We only need to know the Gram matrix $\mathbf{K} = \Phi \Phi^\top$, which is an $M \times M$ symmetric matrix with elements $\mathbf{K}_{nm} = k(\mathbf{x}_n, \mathbf{x}_m) = \phi^\top(\mathbf{x}_n)\phi(\mathbf{x}_m)$. It has to be a positive semidefinite matrix (see [1]), i.e., we only need to know the kernel function $k(\cdot, \cdot)$ that represents the inner product in the new feature space to calculate the marginal distribution. This leads to the construction of kernel functions $k(\mathbf{x}, \mathbf{x}')$ for which the Gram matrix \mathbf{K} is positive semidefinite for all possible choices of the set $\{\mathbf{x}_n\}$. Note that, even if Φ has an infinite number of columns, which corresponds to the case of $\phi(\mathbf{x}_i)$ being an infinite dimensional feature vector, we can still calculate \mathbf{K} of size $M \times M$ by means of the kernel function. Consequently, the new feature space dimension depends of the selected kernel function.

It is also worth noting that the above marginal distribution can be obtained by assuming that \mathbf{y} consists of independent additive noisy observations, with variance γ^2 , of a Gaussian process with mean b and covariance \mathbf{K} .

For a new sample \mathbf{x}_* the distribution of

$$\mathbf{y}_{M+1} = \begin{pmatrix} \mathbf{y} \\ y(\mathbf{x}_*) \end{pmatrix}, \quad (5)$$

has the form

$$p(\mathbf{y}_{M+1} | b, \gamma^2, \sigma^2) = \mathcal{N}(\mathbf{y}_{M+1} | b\mathbf{1}_{M+1}, \mathbf{C}_{M+1}), \quad (6)$$

with $\mathbf{C}_{M+1} = \gamma^2 \Phi_{M+1} \Phi_{M+1}^\top + \sigma^2 \mathbf{I}_{M+1}$, which can be written as

$$\mathbf{C}_{M+1} = \begin{pmatrix} \mathbf{C} & \mathbf{k} \\ \mathbf{k}^\top & c \end{pmatrix}, \quad (7)$$

where \mathbf{C} has been defined in Eq. (4) and

$$\mathbf{k}^\top = \gamma^2 \phi^\top(\mathbf{x}_*) \Phi^\top, \quad (8)$$

$$c = \gamma^2 \phi^\top(\mathbf{x}_*) \phi(\mathbf{x}_*) + \sigma^2. \quad (9)$$

Furthermore, the conditional distribution $p(y(\mathbf{x}_*)|\mathbf{y})$ is a Gaussian distribution with mean $m(\mathbf{x}_*)$ and variance $v(\mathbf{x}_*)$ given by

$$m(\mathbf{x}_*) = b + \gamma^2 \phi^\top(\mathbf{x}_*) \Phi^\top \mathbf{C}^{-1} (\mathbf{y} - b\mathbf{1}), \quad (10)$$

$$v(\mathbf{x}_*) = c - \mathbf{k}^\top \mathbf{C}^{-1} \mathbf{k}. \quad (11)$$

B. Posterior Distribution of \mathbf{w}

When the feature space is finite dimensional we can also calculate the posterior distribution of \mathbf{w} , which is given by (see [4]),

$$p(\mathbf{w}|\mathbf{y}, b, \gamma^2, \sigma^2) = \mathcal{N}(\mathbf{w} | \Sigma_{\mathbf{w}|\mathbf{y}, \gamma^2, \sigma^2} \sigma^{-2} \Phi^\top (\mathbf{y} - b\mathbf{1}), \Sigma_{\mathbf{w}|\mathbf{y}, \gamma^2, \sigma^2}), \quad (12)$$

where

$$\Sigma_{\mathbf{w}|\mathbf{y}, \gamma^2, \sigma^2} = (\sigma^{-2} \Phi^\top \Phi + \gamma^{-2} \mathbf{I})^{-1}.$$

Notice that $m(\mathbf{x}_*)$ defined in Eq. (10) can be expressed in terms of $\mathbb{E}[\mathbf{w}]$ as

$$m(\mathbf{x}_*) = \phi(\mathbf{x}_*)^\top \mathbb{E}[\mathbf{w}] + b. \quad (13)$$

C. Parameter Estimation

The last step in the Bayesian inference we are carrying out is the estimation of the parameters involved in the models, that is, the estimation of the values of γ^2 , σ^2 , and b . The value of b can be easily obtained from Eq. (3) as

$$b = \frac{1}{M} \sum_{i=1}^M y(\mathbf{x}_i). \quad (14)$$

To estimate the values of γ^2 and σ^2 we use the Evidence Bayesian approach without any prior information on these parameters. The Evidence Bayesian approach [43], see [44], [45] for other possible names, determines the values of the parameters γ^2 and σ^2 by maximizing the marginal distribution in Eq. (3) obtained by integrating out the vector of adaptive parameters \mathbf{w} . Intuitively, by integrating over \mathbf{w} we are searching for the best value of γ^2 and σ^2 for all possible values of \mathbf{w} . Differentiating $2 \ln p(\mathbf{y}|b, \gamma^2, \sigma^2)$ with respect to γ^2 and equating the result to zero, we obtain

$$\text{tr}[\mathbf{C}^{-1} \Phi \Phi^\top] = \text{tr}[(\mathbf{y} - b\mathbf{1})^\top \mathbf{C}^{-1} \Phi \Phi^\top \mathbf{C}^{-1} (\mathbf{y} - b\mathbf{1})]. \quad (15)$$

Diagonalizing $\Phi \Phi^\top$, we obtain $\mathbf{U} \Phi \Phi^\top \mathbf{U}^\top = \mathbf{D}$, where \mathbf{U} is an orthonormal matrix and \mathbf{D} is a diagonal matrix with entries $\lambda_i, i = 1, \dots, M$. We can then rewrite the above equation as

$$\sum_{k=1}^M \frac{\lambda_k}{\gamma^2 \lambda_k + \sigma^2} = \sum_{i=1}^M z_i^2 \frac{\lambda_i}{(\gamma^2 \lambda_i + \sigma^2)^2}, \quad (16)$$

where $\mathbf{U}(\mathbf{y} - b\mathbf{1}) = \mathbf{z}$ with components $z_i, i = 1, \dots, M$.

Multiplying both sides of the above equation by γ^2 we have

$$\gamma^2 = \sum_{i=1}^M \frac{\frac{\lambda_i}{\gamma^2 \lambda_i + \sigma^2}}{\sum_{k=1}^M \frac{\lambda_k}{\gamma^2 \lambda_k + \sigma^2}} \frac{\gamma^2 z_i^2}{\gamma^2 \lambda_i + \sigma^2} = \sum_{i=1}^M \mu_i \frac{\gamma^2 z_i^2}{\gamma^2 \lambda_i + \sigma^2}, \quad (17)$$

where

$$\mu_i = \frac{\frac{\lambda_i}{\gamma^2 \lambda_i + \sigma^2}}{\sum_{k=1}^M \frac{\lambda_k}{\gamma^2 \lambda_k + \sigma^2}}. \quad (18)$$

Note that $\mu_i \geq 0$ and $\sum_{i=1}^M \mu_i = 1$.

Similarly, differentiating $2 \ln p(\mathbf{y}|\gamma^2, \sigma^2)$ with respect to σ^2 and equating the result to zero, we obtain

$$\sum_{k=1}^M \frac{1}{\gamma^2 \lambda_k + \sigma^2} = \sum_{i=1}^M z_i^2 \frac{1}{(\gamma^2 \lambda_i + \sigma^2)^2}. \quad (19)$$

Following the same steps we already performed to estimate γ^2 , we obtain

$$\sigma^2 = \sum_{i=1}^M \nu_i \frac{\sigma^2 z_i^2}{\gamma^2 \lambda_i + \sigma^2}, \quad (20)$$

where

$$\nu_i = \frac{\frac{1}{\gamma^2 \lambda_i + \sigma^2}}{\sum_{k=1}^M \frac{1}{\gamma^2 \lambda_k + \sigma^2}}. \quad (21)$$

Note that, again, $\nu_i \geq 0$ and $\sum_{i=1}^M \nu_i = 1$.

Equations (17) and (20) suggest the iterative procedure described in Alg. 1 to estimate the parameters where the old value of the parameters is used in the right hand side of the equations to obtain a new estimate of the parameters in the left hand side of the equations.

Algorithm 1 Parameter estimation

Using Eq. (14), compute $b = \frac{1}{M} \sum_{i=1}^M y(\mathbf{x}_i)$.

Compute \mathbf{U} and $\lambda_i, i = 1, \dots, M$, as the eigenvector matrix and eigenvalues of $\Phi \Phi^\top$, respectively.

Set $\mathbf{z} = \mathbf{U}(\mathbf{y} - b\mathbf{1})$.

Initialize $\gamma^2 = 1, \sigma^2 = 1$.

repeat

 Set $\gamma_{old}^2 = \gamma^2, \sigma_{old}^2 = \sigma^2$.

 Set $\gamma^2 = \sum_{i=1}^M \mu_i \gamma_{old}^2 z_i^2 / (\gamma_{old}^2 \lambda_i + \sigma_{old}^2)$.

 Set $\sigma^2 = \sum_{i=1}^M \nu_i \sigma_{old}^2 z_i^2 / (\gamma_{old}^2 \lambda_i + \sigma_{old}^2)$.

until $(\gamma^2 - \gamma_{old}^2)^2 / (\gamma_{old}^2)^2 < 10^{-6}$ and $(\sigma^2 - \sigma_{old}^2)^2 / (\sigma_{old}^2)^2 < 10^{-6}$.

D. Classification

Once the system has been trained, we want to assign a class to a new value of \mathbf{x} , denoted by \mathbf{x}_* . We already know that the conditional distribution $p(y(\mathbf{x}_*)|\mathbf{y})$ is a Gaussian distribution with mean $m(\mathbf{x}_*)$ and variance $v(\mathbf{x}_*)$ given in Eqs. (10) and (11). We classify \mathbf{x}_* utilizing $m(\mathbf{x}_*)$ and write

$$\mathbf{x}_* \text{ is assigned to } \begin{cases} \mathcal{C}_1 & \text{if } m(\mathbf{x}_*) \geq 0.5 \\ \mathcal{C}_0 & \text{if } m(\mathbf{x}_*) < 0.5 \end{cases}. \quad (22)$$

Notice that the classification of \mathbf{x}_* is based on the proximity of the mean value of $p(y(\mathbf{x}_*)|\mathbf{y})$ to the value zero or one that represents the classes \mathcal{C}_0 and \mathcal{C}_1 , respectively.

IV. PROPOSED ACTIVE LEARNING METHOD

Active learning starts with a small set of observations whose class is already known. From these observations, the marginal distribution of \mathbf{y} , the conditional distribution of \mathbf{w} given \mathbf{y} , and the parameters b , γ^2 , and σ^2 are estimated using the procedure described in the previous sections. In order to improve the performance of the classifier we want to select a new training sample \mathbf{x}_+ , whose corresponding $y(\mathbf{x}_+)$ will be learned by querying the oracle. Let us now examine different ways to select the new training sample.

A. Method 1: Maximum differential of entropies

Utilizing Eq. (10) and (11) we observe that, for a sample \mathbf{x} not already present in the training set, the distribution of $y(\mathbf{x})$ given the set of observations \mathbf{y} has variance

$$v(\mathbf{x}) = \gamma^2 \boldsymbol{\phi}^\top(\mathbf{x}) \boldsymbol{\phi}(\mathbf{x}) + \sigma^2 - \gamma^4 \boldsymbol{\phi}^\top(\mathbf{x}) \boldsymbol{\Phi}^\top \mathbf{C}^{-1} \boldsymbol{\Phi} \boldsymbol{\phi}(\mathbf{x}), \quad (23)$$

and consequently we can select the new training sample as the one maximizing the variance of the prediction, that is,

$$\mathbf{x}_+ = \arg \max_{\mathbf{x}} v(\mathbf{x}). \quad (24)$$

Notice that using this criterion amounts to selecting the sample the classifier is less certain about the class it belongs to.

Let us relate this active method procedure to the one proposed in [24], [31] for finite dimensional feature spaces. The covariance matrix of the posterior distribution of \mathbf{w} when a new \mathbf{x} is added to the training set is given by

$$\boldsymbol{\Sigma}_{\mathbf{w}|\mathbf{y}, \gamma^2, \sigma^2}^{\mathbf{x}} = (\sigma^{-2}(\boldsymbol{\Phi}^\top \boldsymbol{\Phi} + \boldsymbol{\phi}(\mathbf{x}) \boldsymbol{\phi}^\top(\mathbf{x})) + \gamma^{-2} \mathbf{I})^{-1}. \quad (25)$$

For finite dimensional feature spaces it is proposed in [24], [31] to add to the training set the sample with maximum difference between the entropies of the posterior distribution before and after adding the new sample, that is,

$$\mathbf{x}_+ = \arg \max_{\mathbf{x}} \frac{1}{2} \log |\boldsymbol{\Sigma}_{\mathbf{w}|\mathbf{y}, \gamma^2, \sigma^2}^{\mathbf{x}}|^{-1}. \quad (26)$$

Let us first express this criterion in terms of the marginal distribution of the observations in order to remove the need of using finite dimensional feature spaces. We note that

$$\begin{aligned} \log |\boldsymbol{\Sigma}_{\mathbf{w}|\mathbf{y}, \gamma^2, \sigma^2}^{\mathbf{x}}|^{-1} &= \log |\mathbf{I} + \sigma^{-2} \boldsymbol{\phi}(\mathbf{x}) \boldsymbol{\phi}^\top(\mathbf{x}) (\sigma^{-2} \boldsymbol{\Phi}^\top \boldsymbol{\Phi} + \gamma^{-2} \mathbf{I})^{-1}| \\ &= \log(1 + \sigma^{-2} \boldsymbol{\phi}^\top(\mathbf{x}) (\sigma^{-2} \boldsymbol{\Phi}^\top \boldsymbol{\Phi} + \gamma^{-2} \mathbf{I})^{-1} \boldsymbol{\phi}(\mathbf{x})), \end{aligned} \quad (27)$$

and using

$$(\sigma^{-2} \boldsymbol{\Phi}^\top \boldsymbol{\Phi} + \gamma^{-2} \mathbf{I})^{-1} = \gamma^2 \mathbf{I} - \gamma^4 \boldsymbol{\Phi}^\top \mathbf{C}^{-1} \boldsymbol{\Phi}, \quad (28)$$

we can write Eq. (27) in terms of the marginal distribution of the observations as

$$\begin{aligned} \log |\boldsymbol{\Sigma}_{\mathbf{w}|\mathbf{y}, \gamma^2, \sigma^2}^{\mathbf{x}}|^{-1} &= \log(1 + \sigma^{-2} \gamma^2 \boldsymbol{\phi}^\top(\mathbf{x}) \boldsymbol{\phi}(\mathbf{x}) - \sigma^{-2} \gamma^4 \boldsymbol{\phi}^\top(\mathbf{x}) \boldsymbol{\Phi}^\top \mathbf{C}^{-1} \boldsymbol{\Phi} \boldsymbol{\phi}(\mathbf{x})) \\ &= \log(1 + \sigma^{-2} \gamma^2 \boldsymbol{\phi}^\top(\mathbf{x}) \boldsymbol{\phi}(\mathbf{x}) \\ &\quad - \sigma^{-2} \gamma^4 \boldsymbol{\phi}^\top(\mathbf{x}) \boldsymbol{\Phi}^\top \boldsymbol{\Sigma}_{\mathbf{y}|\gamma^2, \sigma^2}^{-1} \boldsymbol{\Phi} \boldsymbol{\phi}(\mathbf{x})). \end{aligned} \quad (29)$$

Consequently, all needed quantities to select \mathbf{x}_+ can be calculated without knowledge of the feature vectors and the posterior distribution of the possibly infinite dimensional adaptive parameters and using only kernel functions and the marginal distribution of the observations.

Furthermore we have

$$\log |\boldsymbol{\Sigma}_{\mathbf{w}|\mathbf{y}, \gamma^2, \sigma^2}^{\mathbf{x}}|^{-1} = \log(\sigma^{-2} v(\mathbf{x})), \quad (30)$$

and consequently both criteria coincide. Notice that, as we have already mentioned, we have also shown that the maximum differential of entropies criterion can be utilized over infinite dimensional feature spaces.

B. Method 2: Minimum distance to decision boundary

In our classification problem the decision boundary corresponds to the set

$$\boldsymbol{\Pi} = \{\mathbf{x} \in \mathcal{X} : \boldsymbol{\phi}^\top(\mathbf{x}) \mathbb{E}[\mathbf{w}] + b - 0.5 = 0\}. \quad (31)$$

We can then select the next sample to be included in the training set by using

$$\begin{aligned} \mathbf{x}_+ &= \arg \min_{\mathbf{x}} d^2(\mathbf{x}, \boldsymbol{\Pi}) \\ &= \arg \min_{\mathbf{x}} \frac{(\boldsymbol{\phi}^\top(\mathbf{x}) \mathbb{E}[\mathbf{w}] + b - 0.5)^2}{\|\mathbb{E}[\mathbf{w}]\|^2} \\ &= \arg \min_{\mathbf{x}} (m(\mathbf{x}) - 0.5)^2. \end{aligned} \quad (32)$$

Note that this method provides a Bayesian formulation of the SVM margin sampling heuristic (see [41]).

C. Method 3: Minimum Normalized Distance

The two active learning methods described above take into consideration only partial aspects of the conditional distribution $p(y(\mathbf{x}_*)|\mathbf{y})$. While maximum differential of entropies utilizes the variance of this distribution, it does not use the distance to the decision boundary. On the other hand, the minimum distance to the decision boundary criterion is based on the mean of this conditional distribution and does not take into account the uncertainty of the distribution. It is obviously very easy to imagine scenarios where these two criteria will not select the best sample, either because it is too far from the decision boundary and, hence, having large variance does not represent a problem, or because, although the sample is the closest to the decision boundary, its uncertainty is very small and consequently it may not be the best sample to be included in the training set.

We can then use the following active learning procedure which combines precision and proximity to the decision boundary

$$\mathbf{x}_+ = \arg \min_{\mathbf{x}} \mathbb{E} \left[\frac{(y(\mathbf{x}) - 0.5)^2}{v(\mathbf{x})} \right], \quad (33)$$

where the expected value is calculated utilizing the conditional distribution $p(y(\mathbf{x})|\mathbf{y})$ defined in Eqs. (10) and (11).

Notice that since

$$\mathbb{E} \left[\frac{(y(\mathbf{x}) - 0.5)^2}{v(\mathbf{x})} \right] = 1 + \frac{(m(\mathbf{x}) - 0.5)^2}{v(\mathbf{x})}, \quad (34)$$

we can rewrite this criterion as

$$\mathbf{x}_+ = \arg \min_{\mathbf{x}} \frac{(m(\mathbf{x}) - 0.5)^2}{v(\mathbf{x})} \quad (35)$$

D. Multiclass Extension of the Active Learning Methods

Here we extend the proposed active learning methods to deal with K -class problems. Recently, architectures for multiclass active learning have been proposed. For instance, in [33] authors propose the MCLU technique which selects the most uncertain samples according to a confidence score based on the distances to all separation hyperplanes. Note, however, that this approach is specific to maximum margin algorithms like SVM, which is not our case. In this paper, nevertheless, we will use the classical one-versus-all strategy for tackling multiclass problems. Hence, for each candidate \mathbf{x} , K different pair of values $\{m_k(\mathbf{x}), v_k(\mathbf{x})\}_{k=1, \dots, K}$ are obtained. These values are used in Eqs. (24), (32) or (35), depending on the selected method, that is finally optimized with respect to \mathbf{x} and k .

V. EXPERIMENTAL RESULTS

In this section, the proposed method is applied to both purely supervised and active remote sensing image classification settings. The method is compared to the standard SVM algorithm in the case of supervised classification when few labeled samples are available. This problem is typically encountered in remote sensing image classification, in which active learning can improve performance. Comparison to random sampling and standard active learning methods, such as margin sampling and entropy-query-by-bagging is then performed. In all cases we provide the overall accuracy, the estimated Cohen's kappa statistic and Z -score³ as measures of accuracy and class agreement, respectively. All experiments were implemented using Matlab[©] and run on an Intel[©] i7@2.67GHz. The Matlab[©] source code of the proposed method is available at <http://decsai.ugr.es/vip/resources/BAL.html> for the interested reader. Additionally, a video demonstration of the method is available at the same location.

A. Study area and data collection

Two multispectral images are used in our experiments for supervised and active learning classification:

- *Supervised classification with Landsat imagery.* The image was acquired in the context of the Urban Expansion Monitoring project [46] over the city of Rome (Italy) by the Landsat TM sensor in 1999. An external Digital Elevation Model (DEM) and a reference land cover map provided by the Italian Institute of Statistics (ISTAT) were also available. The available features were the seven Landsat bands, two SAR backscattering intensities (0–35 days), and the SAR interferometric coherence.

Since image features come from different sensors, the first step was to perform a specific processing and conditioning of optical and SAR data, and to co-register all

images [46]. In particular, the seven bands of Landsat TM were co-registered with the ISTAT classification data, and resampled to 30×30 m with the Nearest-Neighbor algorithm. The registration for the multi-source images was performed at the sub-pixel level obtaining a root-mean-squared error of about 10 m, which potentially enables good urban classification ability. We also appended two SAR features: the estimated coherence, Co , and a spatially filtered version of the coherence, FCo , which is specially designed to increase the urban areas discrimination [46]. After this preprocessing, all features were stacked at the pixel level, and each feature was standardized. The goal is the discrimination of urban (C_1) versus non-urban (C_0) land-cover classes.

- *Active classification with ROSIS imagery.* The second image was acquired by the DAIS7915 sensor over the city of Pavia (Italy), and constitutes a challenging 9-class urban classification problem dominated by directional features and relatively high spatial resolution (5 meters pixels). We took into account only 40 spectral bands of reflective energy in the range $[0.5, 1.76] \mu\text{m}$, thus skipping thermal infrared bands and middle infrared bands above 1958 nm. We carried out a Principal Components Analysis (PCA) to reduce the dimensionality of the problem and considered the 10 first components for each pixel that have provided good classification performance in previous works (see, for instance, [47]).

B. Supervised Classification Results

For the case of supervised classification, we report results both on the binary classification problem of the Rome scene and the multiclass classification problem of the Pavia scene and compare the performance of our approach to the standard SVM approach.

From the Rome image, of size 1440×930 pixels, a training set of 500 randomly selected pixels was obtained, and results are given in a representative test set of 10000 samples. To obtain unbiased conclusions from the results, the process was repeated 10 times with different randomly selected training and test sets, and the average accuracies are given. In all cases, a Gaussian kernel was used. Using 3-fold cross-validation with the SVM as classifier, a kernel lengthscale $\sigma = 100$ was selected. Although we could have used Bayesian inference to estimate the kernel parameter (see, for instance, [4]) we decided to use the same kernel parameter on both methods and concentrate on the remaining model parameters. Notice that this decision slightly favors SVM since the kernel parameter is estimated seeking the best SVM performance. For the case of SVMs, the regularization parameter C was tuned by 3-fold cross-validation on the training dataset. Our method does not need any heuristic tuning since hyperparameters are estimated automatically in the training phase. The proposed method needed 0.33 seconds to complete the training while the SVM needed 1.94 seconds.

Table I shows the obtained results in the 10 independent realizations and their average and variance. Although SVM obtains better results in many cases, the differences are not

³ Z -score is defined as the ratio between the estimated kappa statistic and its standard deviation.

TABLE I

CLASSIFICATION ACCURACY FOR SVM AND THE PROPOSED METHOD IN THE ROME (1999) SCENE. OVERALL ACCURACY, ESTIMATED COHEN'S STATISTIC AND Z-SCORE RESULTS ARE GIVEN FOR ALL 10 REALIZATIONS AND AVERAGED.

Realization	Overall accuracy, OA[%]		Kappa statistic, κ		Z-score, Z	
	Proposed	SVM	Proposed	SVM	Proposed	SVM
1	96.66	96.95	0.895	0.905	158.04	169.50
2	96.48	96.61	0.890	0.896	154.57	161.07
3	97.27	96.96	0.914	0.905	177.60	168.35
4	96.24	96.54	0.883	0.894	150.20	160.00
5	97.10	96.54	0.909	0.892	172.39	157.37
6	96.64	95.99	0.893	0.874	156.69	142.59
7	96.86	96.58	0.905	0.898	170.81	165.14
8	96.76	96.72	0.895	0.896	156.32	158.44
9	97.02	96.71	0.901	0.900	174.92	167.36
10	96.99	97.00	0.906	0.908	170.54	173.04
Average	96.80	96.66	0.90	0.90	164.21	162.29
Variance	0.0957	0.0867	$< 10^{-5}$	$< 10^{-5}$	98.95	74.66

TABLE II

MEAN CONFUSION MATRIX FOR SVM AND THE PROPOSED METHOD (IN BRACKETS). WE SHOW THE AVERAGE KAPPA STATISTIC, ALONG WITH ITS VARIANCE, Z-SCORE AND CONFIDENCE INTERVALS FOR BOTH METHODS.

	\hat{C}_0	\hat{C}_1
\hat{C}_0	7802.80 (7846.30)	169.00 (198.30)
\hat{C}_1	165.00 (121.50)	1863.20 (1833.90)
OA [%]	SVM 96.66%	Proposed 96.80%
κ	0.90	0.90
σ_κ^2	3.07e-05	3.02e-05
Z-score	162.29	164.21
κ CI	[0.886,0.908]	[0.889,0.910]

statistically significant, as assessed by the average values of the three measures. Table II shows the *average* confusion matrices for the 10 realizations, along with its variance, Z-score and confidence intervals for both methods. These results also confirm the numerical and statistical similarity of the results. Finally, Fig. 1 shows the classification maps obtained by SVM and the proposed method in a particular realization. Visual results match the previous numerical accuracies as no difference is obtained. The statistical significance of the kappa statistic also confirms this issue.

A second experiment was performed on the 9-class urban classification problem of the Pavia scene depicted in Fig. 2a, which has 400×400 pixels. Training was done on 1260 randomly selected pixels (140 from each class), and a test set of 13314 representative samples was used. Again, ten different realizations were used to obtain unbiased conclusions from the results. We used a Gaussian kernel, whose lengthscale $\sigma = 500$ was selected using 3-fold cross-validation with the SVM as classifier. As in the previous experiment, the regularization parameter C for the SVM was tuned by 3-fold cross-validation on the training dataset while the proposed method estimated all hyperparameters automatically in the training phase. The proposed method needed 14.32 seconds to complete the training while the SVM needed 9.09 seconds. This is explained by the fact that SVM estimates a single value of C for all classifiers while the proposed method has to estimate the value of the hyperparameters for each classifier.

Table III shows the obtained results in the 10 independent

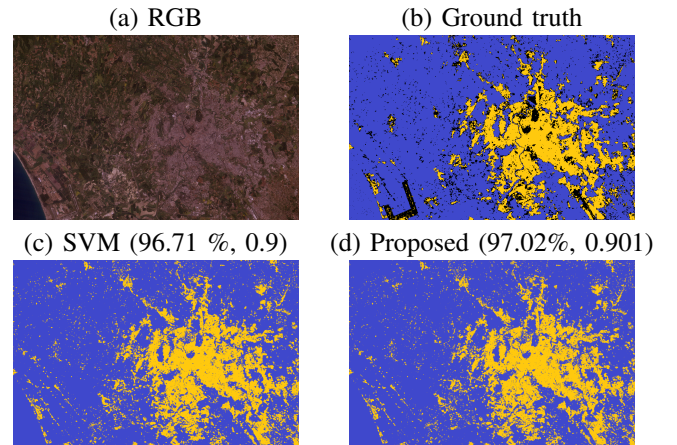


Fig. 1. (a) RGB composite of the Landsat multispectral image, (b) ground truth showing the urban (yellow), non-urban (blue) classes and background (black), (c) classification map with SVMs, and (d) classification map with the proposed method. Overall accuracy and kappa statistic are given in parentheses.

TABLE III

CLASSIFICATION ACCURACY FOR SVM AND THE PROPOSED METHOD IN THE PAVIA SCENE. OVERALL ACCURACY, ESTIMATED COHEN'S STATISTIC AND Z-SCORE RESULTS ARE GIVEN FOR ALL 10 REALIZATIONS AND AVERAGED.

Realization	Overall accuracy, OA[%]		Kappa statistic, κ		Z-score, Z	
	Proposed	SVM	Proposed	SVM	Proposed	SVM
1	98.24	98.10	0.979	0.977	705.28	678.51
2	97.75	98.13	0.973	0.977	622.88	683.29
3	98.31	98.28	0.979	0.979	721.51	712.40
4	98.42	98.42	0.981	0.981	743.00	744.51
5	98.46	98.11	0.981	0.977	754.33	680.18
6	97.95	98.36	0.975	0.980	653.39	730.60
7	98.48	98.27	0.981	0.979	760.01	710.51
8	98.29	98.23	0.979	0.978	714.96	701.91
9	98.37	98.30	0.980	0.979	733.20	715.85
10	98.18	97.87	0.978	0.974	693.48	640.06
Average	98.25	98.21	0.979	0.978	710.20	699.78
Variance	0.0545	0.0253	$< 10^{-6}$	$< 10^{-6}$	1926.64	906.97

realizations and their average and variance. The proposed method provides better results in almost all cases, although the differences are not statistically significant, as assessed by the Z score of the κ statistic for both classifiers. Unlike the overall accuracy, the kappa statistic avoids the chance effect, and a value above 0.8 is typically considered to be a 'very good' agreement. The kappa index confidence interval is [0.975, 0.980] for the proposed method and [0.975, 0.981] for the SVM. These results also confirm the numerical and statistical similarity of the results. Finally, Fig. 2 shows the classification maps obtained by SVM and the proposed method in a particular realization. Visual results match the previous numerical accuracies as no difference is obtained.

C. Active Learning Results

In this second battery of experiments, we illustrate the capabilities of the proposed active learning methods. Classification experiments are conducted using the Rome (Italy) scene acquired in 1999 whose RGB bands are depicted in Fig. 1a. The proposed Bayesian active learning methods are

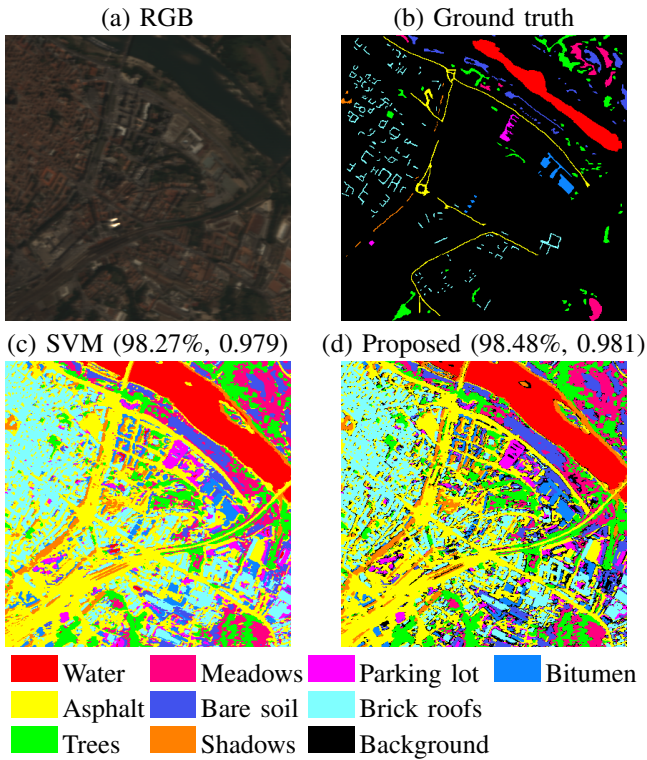


Fig. 2. (a) False color Pavia multispectral image composed by bands [8, 4, 1], (b) ground truth showing classes in colors and background in black, (c) classification map with SVMs, and (d) classification map with the proposed method. Overall accuracy and kappa statistic are given in parentheses.

identified as follows: maximum differential of entropies (BAL-1), the minimum distance to decision boundary (BAL-2), and the minimum normalized distance (BAL-3). They are compared to SVM-based approaches following similar heuristics: margin sampling (MS) [21] and entropy-query-by-bagging (EQB) [36]. The naïve (passive) approach of random sampling (RS) is included here as baseline.

Figure 3 shows the average accuracy curves over 10 realizations with different randomly selected training, pool and test sets as a function of the number of training samples. The initial training set is formed by only 7 labeled pixels for each class, while the pool set has 986 spectra, and the test set is formed by 10000 samples. Although the proposed method can be used for the selection of a batch of samples, in the experiments we report results by adding one sample at each iteration (query). At each iteration the SVM model was retrained using 3-fold cross validation on the current training dataset to tune the regularization parameter C . The parameters for the proposed method were automatically estimated using Eqs. (17) and (20). For the EQB method, six classifiers were used. The compared methods perform remarkably differently from the very beginning: while all of them start from approximately $Z = 80$, a fast convergence is observed for all methods but RS, as expected. MS and EQB show very similar performance, and both outperform our proposed BAL-1. The curves also reveal better results at convergence for the BAL-2 and BAL-3 methods. Nevertheless, for a low number of iterations (between 25-50), BAL-3 shows much

TABLE IV
FIGURES OF MERIT AT CONVERGENCE IN THE ROME (1999) SCENE (AFTER 100 SAMPLES WERE ADDED) FOR ALL LEARNING METHODS.

Methods	Avg. OA	σ_{OA}^2	Avg. kappa	σ_{κ}^2	Z-score	κ CI
SVM-RS	95.09	0.7520	0.8467	0.0008	128.79	[0.83,0.86]
SVM-MS	97.08	0.0894	0.9095	0.0001	175.23	[0.90,0.92]
SVM-EQB	97.06	0.1009	0.9094	0.0001	175.53	[0.90,0.92]
BAL-1	96.41	0.1847	0.8869	0.0002	152.87	[0.88,0.90]
BAL-2	97.31	0.0921	0.9166	0.0001	183.82	[0.91,0.93]
BAL-3	97.34	0.0412	0.9173	$< 10^{-4}$	184.28	[0.91,0.93]

TABLE V
TOTAL RUNNING TIME IN SECONDS FOR ALL ACTIVE LEARNING METHODS IN THE ROME(1999) SCENE.

SVM-RS	SVM-MS	SVM-EQB	BAL-1	BAL-2	BAL-3
179	185	235	9	9	9

better results. The dashed line represents the upper bound for $OA=97.45$ and $Z\text{-score}=187.62$. Table IV gives the accuracy, kappa and Z agreement scores after the full iterative process, when 100 samples were added, and confirms the suitability of the proposed methods, specifically BAL-2 and BAL-3, which show higher accuracies and lower variance. Table V shows the total running time in seconds, after 100 queries, for the compared methods, including the initial learning stage and the parameter estimation at each query. It is worth mentioning that the running time for SVM based methods, MS and EQB, is much higher than the time for the proposed Bayesian active learning methods.

In addition, a multiclass active learning experiment was performed in the Pavia scene. In this experiment, we compare the multiclass extension of the proposed methods with the multiclass versions of RS, MS and EQB. Also, the Multiclass Level Uncertainty method (MCLU) [33] was included in the comparison. Figure 4 shows the average accuracy curves over 10 realizations with different randomly selected training, pool and test sets as a function of the number of training samples. The initial training set is formed by only 5 labeled pixels for each class, while the pool set has 13076 spectra, and the test set is formed by 1453 samples. For the parameter selection we followed the same procedure as in the previous experiment. The proposed methods start with an advantage of 2% with respect to the SVM based methods that, in the case of the proposed BAL-2 and BAL-3 methods, is kept until iteration 40. After that, MS, EQB, MCLU, BAL-2 and BAL-3 have a similar behavior. We think that this is due to the way the parameters are estimated. SVM methods use cross-validation to estimate the parameters and, when the training set is small, it does not provide accurate results. However, the proposed method provides a precise estimation even if the number of training samples is very small. BAL-1 performs similarly to RS which confirms that maximizing the variance of the prediction is not a good selection method by itself but, in some cases, helps when combined with the minimum distance to the decision boundary, as in BAL-3 method. Table VI shows the numerical results when 100 samples were added. From those figures of merit we observe that MCLU provides slightly better results than MS, EQB, BAL-2 and BAL-3. The

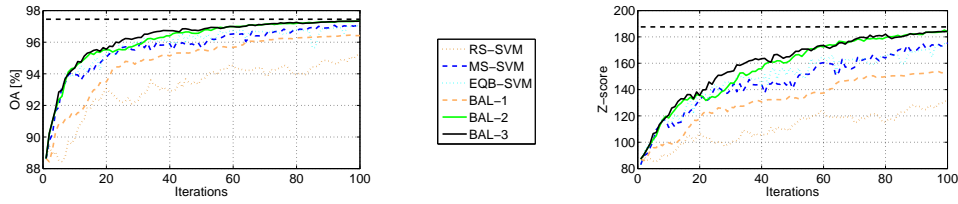


Fig. 3. Average accuracy (left) and Z-score (right) learning curves in the Rome(1999) scene.

TABLE VI

FIGURES OF MERIT AT CONVERGENCE IN THE PAVIA SCENE (AFTER 100 SAMPLES WERE ADDED) FOR ALL ACTIVE LEARNING METHODS.

Methods	Avg. OA	σ_{OA}^2	Avg. kappa	σ_{κ}^2	Z-score	κ CI
SVM-RS	95.09	0.3812	0.9407	$< 10^{-4}$	139.62	[0.93,0.95]
SVM-MS	97.56	0.2055	0.9706	$< 10^{-4}$	202.10	[0.96,0.98]
SVM-MCLU	98.12	0.0934	0.9774	$< 10^{-4}$	230.84	[0.96,0.99]
SVM-EQB	97.90	0.1213	0.9746	$< 10^{-4}$	217.61	[0.96,0.98]
BAL-1	94.51	1.7123	0.9338	0.0003	133.54	[0.92,0.95]
BAL-2	97.92	0.2092	0.9749	$< 10^{-4}$	220.04	[0.97,0.98]
BAL-3	97.69	0.2791	0.9720	$< 10^{-4}$	208.67	[0.96,0.98]

TABLE VII

TOTAL RUNNING TIME IN SECONDS FOR ALL ACTIVE LEARNING METHODS IN THE PAVIA SCENE.

SVM-RS	SVM-MS	SVM-MCLU	SVM-EQB	BAL-1	BAL-2	BAL-3
380	397	401	812	148	165	183

dashed line represents the upper bound for OA=98.50 and Z-score=260.99. Table VII shows, for the compared methods, the total running time in seconds after 100 queries, including the initial learning stage and parameter estimation at each query. Again the running time for SVM based methods is much higher (from 2 to 5 times depending on the method) than the time required by the proposed Bayesian active learning methods.

VI. CONCLUSIONS

This paper presented a non-parametric Bayesian learning approach based on kernels for remote sensing image classification. The Bayesian methodology efficiently tackles purely supervised and active learning approaches, and shows competitive performance when compared to SVMs and recent active learning approaches. For the latter setting, an incremental learning approach based on three different approaches was presented: the maximum differential of entropies, the minimum distance to decision boundary, and the minimum normalized distance. Automatic parameter estimation is solved by using the evidence Bayesian approach, the kernel trick, and the marginal distribution of the observations instead of the posterior distribution of the adaptive parameters.

The proposed approach was tested in several scenes dealing with the urban monitoring problem from multispectral and SAR data. We observed that, while similar results are obtained by SVMs in supervised mode, an improvement in accuracy and convergence is observed for the active learning scenario. Interestingly our methods do not only provide point-wise class predictions but confidence intervals.

Future work will deal with the application to more challenging multitemporal image segmentation and change detection problems, in which a confidence map could be readily exploited. Also, it is interesting to study the performance of the model in the presence of a reduced number of labeled samples and much higher dimensionality scenarios.

ACKNOWLEDGMENT

The authors would like to thank Dr. J. Muñoz-Marí from the Universitat de València (Spain) and Dr. Devis Tuia at the EPFL (Switzerland) for sharing the code of the RS, MS, and EQB-SVM active learning methods compared in this paper. We would also like to thank Dr. L. Gómez-Chova from the Universitat de València (Spain) for the preprocessing of the images used in this work. Finally, we would like to thank Dr. J. Malo from the Universitat de València (Spain) for the fruitful discussions and his valuable suggestions.

REFERENCES

- [1] B. Schölkopf and A. Smola, *Learning with Kernels – Support Vector Machines, Regularization, Optimization and Beyond*, MIT Press Series, Cambridge, MA, USA, 2002.
- [2] G. Camps-Valls and L. Bruzzone, “Kernel-based methods for hyperspectral image classification,” *IEEE Trans. on Geoscience and Remote Sensing*, vol. 43, no. 6, pp. 1351–1362, June 2005.
- [3] G. Camps-Valls and L. Bruzzone, *Kernel Methods for Remote Sensing Data Analysis*, John Wiley and Sons, 2009.
- [4] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer, 2007.
- [5] M. E. Tipping, “The relevance vector machine,” *Journal of Machine Learning Research*, vol. 1, pp. 211–244, 2001.
- [6] P. Torrione and L.M. Collins, “Texture features for antitank landmine detection using ground penetrating radar,” *IEEE Trans. on Geoscience and Remote Sensing*, vol. 45, no. 7, pp. 2374–2382, July 2007.
- [7] F. A. Mianji and Y. Zhang, “Robust hyperspectral classification using relevance vector machine,” *IEEE Trans. on Geoscience and Remote Sensing*, vol. 49, no. 6, pp. 2100–2112, June 2011.
- [8] G. Camps-Valls, L. Gómez-Chova, J. Vila-Francés, J. Amorós-López, J. Muñoz-Marí, and J. Calpe-Maravilla, “Retrieval of oceanic chlorophyll concentration with relevance vector machines,” *Remote Sensing of Environment*, vol. 105, no. 1, pp. 23–33, Nov 2006.
- [9] C.E. Rasmussen and C.K. Williams, *Gaussian Processes for Machine Learning*, MIT Press, NY, 2006.
- [10] Y. Bazi and F. Melgani, “Gaussian process approach to remote sensing image classification,” *IEEE Trans. on Geoscience and Remote Sensing*, vol. 48, no. 1, pp. 186–197, January 2010.
- [11] G. Jun and J. Ghosh, “Spatially adaptive classification of land cover with remote sensing data,” *IEEE Trans. on Geoscience and Remote Sensing*, vol. 49, no. 7, pp. 2662–2673, July 2011.
- [12] J. Verrelst, L. Alonso, G. Camps-Valls, J. Delegido, and J. Moreno, “Retrieval of vegetation biophysical parameters using gaussian process techniques,” *IEEE Trans. on Geoscience and Remote Sensing*, vol. PP, no. 99, pp. 1–12, 2011.
- [13] A. O’Hagan, *Bayesian Inference*, vol. 2B, chapter 10, Arnold, 1994.
- [14] P. Orbanz and Y.-W. Teh, *Bayesian Nonparametric Models*, Springer, 2010.

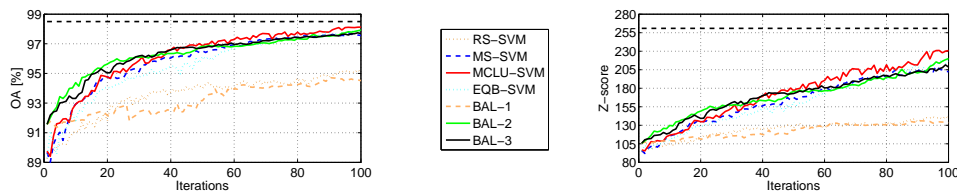


Fig. 4. Average accuracy (left) and Z-score (right) learning curves in the Pavia scene.

- [15] J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle, *Least Squares Support Vector Machines*, World Scientific, Singapore, 2002.
- [16] G. Cauwenberghs and T. Poggio, "Incremental and decremental support vector machine learning," in *NIPS*, 2000, pp. 409–415.
- [17] J. Kivinen, A.J. Smola, and R.C. Williamson, "Online learning with kernels," *IEEE Trans. on Signal Processing*, vol. 52, no. 8, pp. 2165–2176, 2004.
- [18] P. Laskov, C. Gehl, S. Kruger, and K.-R. Müller, "Incremental support vector learning: Analysis implementation and applications," *Journal of Machine Learning Research*, vol. 7, pp. 1909–1936, 2006.
- [19] L. Csató and M. Opper, "Sparse on-line Gaussian processes," *Neural Computation*, vol. 14, no. 3, pp. 641–668, 2002.
- [20] J. Quiñero-Candela and O. Winther, "Incremental Gaussian Processes," in *NIPS*, 2002, pp. 1001–08.
- [21] B. Settles, "Active learning literature survey," Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.
- [22] J.C. Platt, *Probabilities for SV Machines*, pp. 61–74, MIT Press, 2007.
- [23] S. Tong and D. Koller, "Active learning for parameter estimation in Bayesian networks," in *NIPS*, 2000, pp. 647–653.
- [24] J. Paisley, X. Liao, and L. Carin, "Active learning and basis selection for kernel-based linear models: A Bayesian perspective," *IEEE Trans. on Signal Processing*, vol. 58, pp. 2686–2700, 2010.
- [25] P. Ruiz, J. Mateos, R. Molina, and A.K. Katsaggelos, "A Bayesian Active Learning Framework for a Two-class Classification Problem," in *Lecture Notes in Computer Science series*, 2012, vol. 7252, in press.
- [26] Jun Li, J.M. Bioucas-Dias, and A. Plaza, "Hyperspectral image segmentation using a new Bayesian approach with active learning," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 49, no. 10, pp. 3947–3960, October 2011.
- [27] C.J.C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, pp. 121–167, 1998.
- [28] M. Elad, *Sparse and Redundant Representations - From Theory to Applications in Signal and Image Processing*, Springer, 2010.
- [29] D. Babacan, R. Molina, and A. Katsaggelos, "Bayesian compressive sensing using Laplace priors," *IEEE Trans. on Image Processing*, vol. 19, no. 1, pp. 53–63, 2010.
- [30] D. J. C. MacKay, "Information-based objective functions for active data selection," *Neural Computation*, vol. 4, no. 4, pp. 590–604, 1992.
- [31] M. W. Seeger and H. Nickisch, "Compressed sensing and Bayesian experimental design," in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 912–919.
- [32] P. Mitra, B. Uma Shankar, and S.K. Pal, "Segmentation of multispectral remote sensing images using active support vector machines," *Pattern Recognition Letters*, vol. 25, no. 9, pp. 1067–1074, 2004.
- [33] B. Demir, C. Persello, and L. Bruzzone, "Batch mode active learning methods for the interactive classification of remote sensing images," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 49, no. 3, pp. 1014–1032, 2011.
- [34] E. Pasolli, F. Melgani, and Y. Bazi, "SVM active learning through significance space construction," *IEEE Geoscience and Remote Sensing Letters*, vol. 8, no. 3, pp. 431–435, 2011.
- [35] M. Ferecatu and N. Boujemaa, "Interactive remote sensing image retrieval using active relevance feedback," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 45, no. 4, pp. 818–826, 2007.
- [36] D. Tuia, F. Ratle, F. Pacifici, M.F. Kanevski, and W.J. Emery, "Active learning methods for remote sensing image classification," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 47, no. 7, pp. 2218–2232, July 2009.
- [37] M. Volpi, D. Tuia, and M. Kanevski, "Cluster-based active learning for remote sensing image classification," *IEEE Trans. on Geoscience and Remote Sensing*, 2011.
- [38] Q. Liu, X. Liao, and L. Carin, "Detection of unexploded ordnance via efficient semisupervised and active learning," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 46, no. 9, pp. 2558–2567, September 2008.
- [39] J. Li, J.M. Bioucas-Dias, and A. Plaza, "Semisupervised hyperspectral image segmentation using multinomial logistic regression with active learning," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 48, no. 11, pp. 4085–4098, November 2010.
- [40] D. Tuia, E. Pasolli, and W.J. Emery, "Using active learning to adapt remote sensing image classifiers," *Remote Sensing of Environment*, vol. 115, no. 9, pp. 2232–2242, 2011.
- [41] D. Tuia, M. Volpi, L. Copa, M. Kanevski, and J. Muñoz-Marí, "A survey of active learning algorithms for supervised remote sensing image classification," *IEEE Journal on Selected Topics in Signal Processing*, vol. 4, pp. 606–617, 2011.
- [42] Ryan Rifkin and Aldebaro Klautau, "In defense of one-vs-all classification," *J. Mach. Learn. Res.*, vol. 5, pp. 101–141, Dec. 2004.
- [43] R. Molina, A. K. Katsaggelos, and J. Mateos, "Bayesian and regularization methods for hyperparameter estimation in image restoration," *IEEE Transactions on Image Processing*, vol. 8, no. 2, pp. 231–246, 1999.
- [44] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*, New York: Springer-Verlag, 1985.
- [45] D. J. C. MacKay, "Comparison of approximate methods for handling hyperparameters," *Neural Computation*, vol. 11, no. 5, pp. 1035–1068, 1999.
- [46] L. Gomez-Chova, D. Fernández-Prieto, J. Calpe, E. Soria, J. Vila-Francés, and G. Camps-Valls, "Urban monitoring using multitemporal SAR and multispectral data," *Pattern Recognition Letters, Special Issue on "Pattern Recognition in Remote Sensing"*, vol. 27, no. 4, pp. 234–243, 2006.
- [47] G. Camps-Valls, D. Tuia, L. Gómez-Chova, S. Jiménez, and J. Malo, Eds., *Remote Sensing Image Processing*, Morgan & Claypool Publishers, LaPorte, CO, USA, Sept 2011, Collection "Synthesis Lectures on Image, Video, and Multimedia Processing", Al Bovik, Ed.

5.3 Bayesian Classification and Active Learning Using ℓ_p -Priors. Application to Image Segmentation

- **P. Ruiz**, N. Pérez de la Blanca, R. Molina, and A.K. Katsaggelos, “Bayesian Classification and Active Learning Using ℓ_p -Priors. Application to Image Segmentation” in 22th European Signal Processing Conference (EUSIPCO 2014), 1183-1187, Lisbon (Portugal), September 2014.
 - Status: Published
 - Indexed in CORE Conference Ranking as CORE B
 - H index: 9 (Q3: 755/1201)

BAYESIAN CLASSIFICATION AND ACTIVE LEARNING USING l_p -PRIORS. APPLICATION TO IMAGE SEGMENTATION

Pablo Ruiz^{1*}, Nicolás Pérez de la Blanca¹, Rafael Molina¹ and Aggelos K. Katsaggelos²

¹Dept. Ciencias de la Computación e I.A., Universidad de Granada, Spain.

² Dpt. of Electrical Engineering and Computer Science, Northwestern University, USA.

*e-mail:mataran@decsai.ugr.es

ABSTRACT

In this paper we utilize Bayesian modeling and inference to learn a softmax classification model which performs Supervised Classification and Active Learning. For $p < 1$, l_p -priors are used to impose sparsity on the adaptive parameters. Using variational inference, all model parameters are estimated and the posterior probabilities of the classes given the samples are calculated. A relationship between the prior model used and the independent Gaussian prior model is provided. The posterior probabilities are used to classify new samples and to define two Active Learning methods to improve classifier performance: Minimum Probability and Maximum Entropy. In the experimental section the proposed Bayesian framework is applied to Image Segmentation problems on both synthetic and real datasets, showing higher accuracy than state-of-the-art approaches.

1. INTRODUCTION

The goal of Supervised Classification is to learn a model which automatically assigns samples to a set of predefined categories. Different approximations have been proposed in literature. For example, Support Vector Machines (SVMs) [1, 2] find the boundary decision which maximizes the distance between support vectors, Bayesian approaches such as Relevance vector machine [3] or Gaussian Process Classification [4] attempt to learn the underlying probabilistic model.

The use of Bayesian modeling and inference provides huge benefits: prior distributions are used to introduce information on the adaptive parameters, and hyperparameters are learned from data using a consistent framework. Priors based in l_p -quasinorms, $p \leq 1$, enforce sparsity on the adaptive parameters. The use of sparse priors has already been reported for softmax classification problems, see [5] for the use of the l_1 prior, and [6, 7] for the use of quadratic prior. However, the use of l_p -quasinorms, $p < 1$, is of particular importance when only very few features are relevant to the target output of a large number of features. Current approaches utilizing l_p -regularization treat the logistic regression from a likelihood-based perspective, and employ a cross-validation procedure to estimate the required regularization parameters (see [8] for details). Here we propose a Bayesian modeling and inference approach to sparse softmax classification using l_p -priors with $p < 1$. For a given \mathbf{x} , the output vector $\mathbf{y}(\mathbf{x}) = [y_1(\mathbf{x}), \dots, y_K(\mathbf{x})]^T$ consists of the 1-of- K binary representation of its classification. We have

$$p(\mathbf{y}(\mathbf{x})|\mathbf{W}, \mathbf{x}) = \prod_{k=1}^K \left(\frac{\exp(\mathbf{w}_k^T \phi(\mathbf{x}))}{\sum_{i=1}^K \exp(\mathbf{w}_i^T \phi(\mathbf{x}))} \right)^{y_k(\mathbf{x})} \quad (1)$$

where the function $\phi : \mathcal{X} \rightarrow \mathcal{H}$, maps the observed $\mathbf{x} \in \mathcal{X}$ into a higher dimensional feature space \mathcal{H} of dimension M whose first component is 1 and \mathbf{W} is a matrix whose column vectors are the so called adaptive vectors $\mathbf{w}_1, \dots, \mathbf{w}_K$. The goal in softmax classification is to learn the adaptive matrix \mathbf{W} from a set of samples $\mathbf{x}_i, i = 1, \dots, N$ with known classification $\mathbf{y}(\mathbf{x}_i), i = 1, \dots, N$.

Getting the ground-truth label of each sample is in general a costly task. Active Learning (AL) techniques provide an iterative alternative to minimize such cost (see [9] for a complete survey). These techniques train an initial classifier using a small dataset, then, based on an optimality criterion, iteratively select samples (without knowing their labels). These samples are then classified by an oracle and used to improve the initial classifier.

AL techniques depend on the model the classifier learns, and therefore each classifier has its own AL techniques. For SVM, relevant approaches are: the sampling approach discussed in [9], the binary- and multiclass-level uncertainty [10], and the entropy-query-by-bagging [11]. In [12] a Bayesian framework is proposed and differential entropy is used to select new samples. In [13] A Gaussian process is used to estimate the posterior distribution of the labels, and three AL methods are proposed: maximum variance (equivalent to differential entropy in [12]), minimum distance to decision boundary, and a combination of both minimum normalized distance.

The goal of this paper is twofold. Firstly, using a prior based on l_p -quasinorms, we formulate the softmax classification problem from a Bayesian viewpoint. All required algorithmic parameters are also included in the proposed Bayesian model, and are estimated along with the unknowns. Due to the intractability of the posterior distributions, we employ Variational Bayesian analysis to provide an approximation to the posterior distribution of the unknowns. A relationship between the prior model used and the independent Gaussian prior model is also provided. Secondly, we tackle AL by utilizing the posterior distribution of the classes.

The paper is organized as follows. In Section 2 we use Bayesian modeling to define probability distributions on the unknowns. Variational inference is used to develop a training algorithm and a classification rule in Section 3. A study on the relationship between the proposed classification model and the use of Gaussian independent prior models is presented in Section 4. AL techniques are proposed in Section 5. In Section 6, the proposed methods are applied to Image Segmentation on a synthetic example and a real dataset. Conclusions are presented in Section 7.

This work has been supported in part by the Comisión Nacional de Ciencia y Tecnología under contract TIN2010-15137, CEI BioTic at the University of Granada, and the Department of Energy grant DE-NA0000457.

2. BAYESIAN MODEL

To perform Bayesian inference we assume that we already have the K -dimensional classification vectors $\mathbf{y}_i = \mathbf{y}(\mathbf{x}_i)$ associated to the feature samples $\phi(\mathbf{x}_i)$, $i = 1, \dots, N$. Then we can write

$$p(\mathbf{Y}|\mathbf{W}) = \prod_{i=1}^N p(\mathbf{y}_i|\mathbf{W}) \quad (2)$$

where \mathbf{Y} is a $N \times K$ matrix with i^{th} row \mathbf{y}_i^{T} whose components are y_{ik} , $k = 1, \dots, K$, $p(\mathbf{y}_i|\mathbf{W})$ has been defined in Eq. (1) and the set \mathbf{X} containing all the used samples, has been omitted for simplicity.

To estimate \mathbf{W} we use, for each of its columns, the prior distribution $p(\mathbf{w}_k|\alpha_k)$ based on l_p -quasinorms

$$p(\mathbf{w}_k|\alpha_k) \propto \alpha_k^{M/p} \exp \left[-\alpha_k \sum_{i=1}^M |w_{ki}|^p \right], \quad (3)$$

where $\alpha_k > 0$ and $0 < p \leq 1$, $\mathbf{w}_k = (w_{k1}, \dots, w_{kM})^{\text{T}}$, $k = 1, \dots, K$. This type of prior has been shown to enforce sparsity in estimation problems like logistic regression (see [14] and [15] for a regularization point of view) and in areas like image restoration and compressive sensing (see, for instance [16]).

Then, given $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_K)^{\text{T}}$, we have

$$p(\mathbf{W}|\boldsymbol{\alpha}) = \prod_{k=1}^K p(\mathbf{w}_k|\alpha_k). \quad (4)$$

Finally, we assume that each α_k , $k = 1, \dots, K$ has as hyperprior, $p(\alpha_k)$, the Gamma distribution, $p(\alpha_k) = \Gamma(\alpha_k|a_{\alpha_k}^o, b_{\alpha_k}^o)$, where $b_{\alpha_k}^o > 0$ and $a_{\alpha_k}^o > 0$, and have the following global model

$$p(\boldsymbol{\alpha}, \mathbf{W}, \mathbf{Y}) = p(\boldsymbol{\alpha})p(\mathbf{W}|\boldsymbol{\alpha})p(\mathbf{Y}|\mathbf{W}). \quad (5)$$

3. VARIATIONAL BAYESIAN INFERENCE

The Bayesian paradigm dictates that inference on $(\boldsymbol{\alpha}, \mathbf{W})$ should be based on $p(\boldsymbol{\alpha}, \mathbf{W}|\mathbf{Y})$. However, $p(\boldsymbol{\alpha}, \mathbf{W}|\mathbf{Y})$ cannot be found in closed form. Therefore, we apply variational methods to approximate this distribution by a distribution $q(\boldsymbol{\alpha}, \mathbf{W})$. The variational criterion used to find $q(\boldsymbol{\alpha}, \mathbf{W})$ is the minimization of the Kullback-Leibler (KL) divergence, given by

$$\text{KL}(q(\boldsymbol{\alpha}, \mathbf{W})||p(\boldsymbol{\alpha}, \mathbf{W}|\mathbf{Y})) = \text{const} \quad (6)$$

$$+ \int \int q(\boldsymbol{\alpha}, \mathbf{W}) \log \left(\frac{q(\boldsymbol{\alpha}, \mathbf{W})}{p(\boldsymbol{\alpha}, \mathbf{W}, \mathbf{Y})} \right) d\boldsymbol{\alpha}d\mathbf{W}.$$

Unfortunately, due to the form of the prior and the observation models defined in (4) and (2) respectively, the integral above cannot be calculated. To solve this problem we proceed to bound below the distribution $p(\boldsymbol{\alpha}, \mathbf{W}, \mathbf{Y})$ by a function which renders the calculation of $\text{KL}(q(\boldsymbol{\alpha}, \mathbf{W}) || p(\boldsymbol{\alpha}, \mathbf{W}|\mathbf{Y}))$ possible when $p(\boldsymbol{\alpha}, \mathbf{W}, \mathbf{Y})$ is replaced by such a function. A lower bound on $p(\mathbf{w}_k|\alpha_k)$, $k = 1, \dots, K$ is found by using the following inequality (see [17], and [18] based on [19])

$$a^{\frac{p}{2}} \leq \frac{p}{2} \frac{a + \frac{2-p}{p}b}{b^{1-p/2}}, \quad (7)$$

for $a \geq 0$, $b > 0$, and $0 \leq p \leq 2$, which applied to the energy of the prior produces

$$\alpha_k \sum_{i=1}^M |w_{ki}|^p \leq \frac{1}{2} \alpha_k p \sum_{i=1}^M \frac{w_{ki}^2 + \frac{2-p}{p} \theta_{ki}}{\theta_{ki}^{1-p/2}}, \quad (8)$$

where $\theta_i > 0$. Consequently, for the prior in Eq. (3) we have

$$p(\mathbf{w}_k|\alpha_k) \geq \mathbf{M}(\alpha_k, \mathbf{w}_k, \boldsymbol{\theta}_k) = \quad (9)$$

$$= \text{const} \times \alpha_k^{M/p} \exp \left(-\frac{1}{2} \alpha_k p \sum_{i=1}^M \frac{w_{ki}^2 + \frac{2-p}{p} \theta_{ki}}{\theta_{ki}^{1-p/2}} \right),$$

where $\boldsymbol{\theta}_k = (\theta_{k1}, \dots, \theta_{kM})^{\text{T}}$, and we can write

$$p(\mathbf{W}|\boldsymbol{\alpha}) \geq \prod_{k=1}^K \mathbf{M}(\alpha_k, \mathbf{w}_k, \boldsymbol{\theta}_k) = \mathbf{M}(\boldsymbol{\alpha}, \mathbf{W}, \boldsymbol{\Theta}). \quad (10)$$

where $\boldsymbol{\Theta}$ is a matrix with column vectors $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K$. In order to obtain a lower bound on $p(\mathbf{Y}|\mathbf{W})$ we follow [6] and notice that for any $\mathbf{u} \in \mathbb{R}^K$ and $\beta \in \mathbb{R}$ we have

$$\ln \sum_{k=1}^K e^{u_k} \leq \beta + \sum_{k=1}^K \frac{u_k - \beta - \xi_k}{2}$$

$$+ \sum_{k=1}^K (\lambda(\xi_k)((u_k - \beta)^2 - \xi_k^2) + \ln(1 + e^{\xi_k})) \quad (11)$$

for all $\xi_k \in \mathbb{R}_0^+$ with $\lambda(\xi_k) = \frac{1}{2\xi_k} \left(\frac{1}{1+e^{-\xi_k}} - \frac{1}{2} \right)$. Applying (11) to Eq. (2) we obtain

$$\ln p(\mathbf{Y}|\mathbf{W}) = \sum_{i=1}^N \ln p(\mathbf{y}_i|\mathbf{W}) \geq \sum_{i=1}^N \sum_{k=1}^K y_{ik} \mathbf{w}_k^{\text{T}} \phi(\mathbf{x}_i)$$

$$- \sum_{i=1}^N \sum_{k=1}^K \left(\frac{\mathbf{w}_k^{\text{T}} \phi(\mathbf{x}_i) - \beta_i - \xi_{ik}}{2} + \ln(1 + e^{\xi_{ik}}) \right)$$

$$- \sum_{i=1}^N \sum_{k=1}^K \lambda(\xi_{ik}) ((\mathbf{w}_k^{\text{T}} \phi(\mathbf{x}_i) - \beta_i)^2 - \xi_{ik}^2)$$

$$- \sum_{i=1}^N \beta_i = \ln \mathbf{H}(\mathbf{W}, \boldsymbol{\Xi}, \boldsymbol{\beta}, \mathbf{Y}), \quad (12)$$

where $\boldsymbol{\Xi}$ is a matrix with row vectors $\boldsymbol{\xi}_i^{\text{T}}$, $i = 1 \dots N$, each of these vectors has the form $\boldsymbol{\xi}_i = (\xi_{i1}, \dots, \xi_{iK})^{\text{T}}$ and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_N)^{\text{T}}$.

Notice that in [6] the same parameter β is used for all the samples.

Using the lower bounds in (10) and (12), the joint distribution is bounded below by

$$p(\boldsymbol{\alpha}, \mathbf{W}, \mathbf{Y}) \geq p(\boldsymbol{\alpha}) \mathbf{M}(\boldsymbol{\alpha}, \mathbf{W}, \boldsymbol{\Theta}) \mathbf{H}(\mathbf{W}, \boldsymbol{\Xi}, \boldsymbol{\beta}, \mathbf{Y})$$

$$= \mathbf{F}(\boldsymbol{\alpha}, \mathbf{W}, \boldsymbol{\Theta}, \boldsymbol{\Xi}, \boldsymbol{\beta}, \mathbf{Y}). \quad (13)$$

We replace $p(\boldsymbol{\alpha}, \mathbf{W}, \mathbf{Y})$ by this lower bound in (6) and use the factorization $q(\boldsymbol{\alpha}, \mathbf{W}) = q(\boldsymbol{\alpha})q(\mathbf{W})$.

Then the posterior distribution $q(\mathbf{w}_k)$, $k = 1, \dots, K$ is the multivariate normal distribution $\mathcal{N}(\langle \mathbf{w}_k \rangle, \Sigma_{\mathbf{w}_k})$ where

$$\Sigma_{\mathbf{w}_k}^{-1} = \Lambda_k + 2 \sum_{i=1}^N \lambda(\xi_{ik}) \phi(\mathbf{x}_i) \phi^{\text{T}}(\mathbf{x}_i), \quad (14)$$

$$\langle \mathbf{w}_k \rangle = \Sigma_{\mathbf{w}_k} \sum_{i=1}^N ((y_{ik} - \frac{1}{2}) \phi(\mathbf{x}_i) + 2\beta_i \lambda(\xi_{ik}) \phi(\mathbf{x}_i))$$

with $\Lambda_k = \text{diag} \left(\langle \alpha_k \rangle p \theta_{ki}^{p/2-1} \right)$, $i = 1, \dots, M$.

Furthermore we have

$$\theta_{ki} = \langle w_{ki}^2 \rangle = (\Sigma_{\mathbf{w}_k})_{ii} + (\langle w_{ki} \rangle)^2. \quad (15)$$

Furthermore $q(\alpha_k) = \Gamma(\alpha_k | a_{\alpha_k}^o + \frac{M}{p}, b_{\alpha_k}^o + \sum_{i=1}^M \theta_{ki}^{p/2})$ with mean

$$\langle \alpha_k \rangle = \frac{1}{p} \frac{a_{\alpha_k}^o p + M}{b_{\alpha_k}^o + \sum_{i=1}^M (\theta_{ki})^{p/2}}. \quad (16)$$

Finally we have

$$\xi_{ik} = \sqrt{\phi^T(\mathbf{x}_i) \Sigma_{\mathbf{w}_k} \phi(\mathbf{x}_i) + (\langle \mathbf{w}_k \rangle^T \phi(\mathbf{x}_i) - \beta_i)^2}, \quad (17)$$

and

$$\beta_i = \frac{K - 2 + 4 \sum_{k=1}^K \lambda(\xi_{ik}) \langle \mathbf{w}_k \rangle^T \phi(\mathbf{x}_i)}{4 \sum_{k=1}^K \lambda(\xi_{ik})}. \quad (18)$$

Notice that the uncertainty of the estimate of \mathbf{w}_k is incorporated into the estimation procedure of the other unknowns by the use of the covariance matrix $\Sigma_{\mathbf{w}_k}$ in (15), (16) and (17).

The above inference leads to a learning procedure which is summarized in Algorithm 1. At convergence this algorithm estimates all the parameters, including the distribution of the adaptive vectors \mathbf{w}_k . The point estimates of the adaptive vectors are $\langle \mathbf{w}_k \rangle$ in Eq. (14). Given a new sample \mathbf{x}^* , we utilize as predictive distribution of the classes

$$p(C_k | \mathbf{x}^*) = \frac{\exp(\langle \mathbf{w}_k \rangle^T \phi(\mathbf{x}^*))}{\sum_{i=1}^K \exp(\langle \mathbf{w}_i \rangle^T \phi(\mathbf{x}^*))} \quad (19)$$

and assign \mathbf{x}^* to the class with maximum probability.

Algorithm 1 Learning Procedure

Require: $\alpha^0 = (1, \dots, 1)^T$, $\theta_{ki}^0 = 1$, $\xi_{ik}^0 = 1$ and $\beta_i = 1$.

- 1: **repeat**
 - 2: Calculate $q(\mathbf{W})^{n+1}$ using Eq. (14).
 - 3: Calculate $q(\alpha)^{n+1}$ using Eq. (16).
 - 4: Parameters θ_{ki}^{n+1} , ξ_{ik}^{n+1} , and β_i^{n+1} are updated using Eq. (15), Eq. (17) and Eq. (18) respectively.
 - 5: **until** convergence
-

4. RELATION TO INDEPENDENT GAUSSIAN PRIOR MODEL

Let us study here the relationship between the proposed classification model and the use of Gaussian independent prior models on the components of \mathbf{w}_k , $k = 1, \dots, K$. Let us assume that

$$p_G(\mathbf{w}_k | \mathbf{v}_k) \propto \prod_{i=1}^M v_{ki}^{1/2} \exp\left[-\frac{1}{2} v_{ki} w_{ki}^2\right], \quad (20)$$

$$p(\mathbf{v}_k) = \prod_{i=1}^M p(v_{ki}) = \prod_{i=1}^M \Gamma(v_{ki} | a_{\alpha_k}^o, b_{\alpha_k}^o), \quad (21)$$

where $\mathbf{v}_k = (v_{k1}, \dots, v_{kM})^T$, $k = 1, \dots, K$ and the parameters $a_{\alpha_k}^o, b_{\alpha_k}^o$ are the ones defined for the l_p -quasinnorms.

Utilizing the same observation bound in (12), we obtain

$$\begin{aligned} p_G(\mathbf{Y}, \mathbf{W}, \mathbf{Y}) &= p(\mathbf{Y} | \mathbf{W}) \prod_{k=1}^K p(\mathbf{v}_k) p_G(\mathbf{w}_k | \mathbf{v}_k) \\ &\geq H(\mathbf{W}, \mathbf{\Xi}, \mathbf{\beta}, \mathbf{Y}) \prod_{k=1}^K p(\mathbf{v}_k) p_G(\mathbf{w}_k | \mathbf{v}_k). \end{aligned} \quad (22)$$

where \mathbf{Y} is a matrix with row vectors \mathbf{v}_k^T , $k = 1 \dots K$, each of these vectors has the form $\mathbf{v}_k = (v_{k1}, \dots, v_{kM})^T$

Utilizing $q_G(\mathbf{W}) = \prod_{k=1}^K q_G(\mathbf{w}_k)$, the variational posterior distribution $q_G(\mathbf{w}_k)$ is $\mathcal{N}(\langle \mathbf{w}_k \rangle_G, \Sigma_{\mathbf{w}_k, G})$ with parameters

$$(\Sigma_{\mathbf{w}_k, G})^{-1} = \Lambda_{k, G} + 2 \sum_{i=1}^N \lambda(\xi_{ik}) \phi(\mathbf{x}_i) \phi^T(\mathbf{x}_i), \quad (23)$$

$$\begin{aligned} \langle \mathbf{w}_k \rangle_G &= \Sigma_{\mathbf{w}_k, G} \sum_{i=1}^N \left((y_{ik} - \frac{1}{2}) \phi(\mathbf{x}_i) + 2\beta_i \lambda(\xi_{ik}) \phi(\mathbf{x}_i) \right), \\ \Lambda_{k, G} &= \text{diag}(\langle v_{ki} \rangle). \end{aligned} \quad (24)$$

The mean of the posterior distribution approximation of v_{ki} is

$$\langle v_{ki} \rangle = \frac{a_{\alpha_k}^o + \frac{1}{2}}{b_{\alpha_k}^o + \frac{\langle w_{ki}^2 \rangle}{2}}. \quad (25)$$

Let us assume that $a_{\alpha_k}^o = b_{\alpha_k}^o = 0$ and rewrite (14) making explicit its dependency on p . Utilizing (16) we have

$$\Lambda_{k, p} = \text{diag} \left(\frac{a_{\alpha_k}^o p + M}{b_{\alpha_k}^o + \sum_{i=1}^M \theta_{ki}^{p/2}} \right) \quad (26)$$

Taking the limit $p \rightarrow 0$ and using (15), we obtain

$$\lim_{p \rightarrow 0} \Lambda_{k, p} = \text{diag}(\theta_{ki}^{-1}) = \text{diag}(\langle w_{ki}^2 \rangle^{-1}). \quad (27)$$

Let us now examine the Gaussian model. When $a_{\alpha_k}^o = b_{\alpha_k}^o = 0$, we have from (24) and (25)

$$\Lambda_{k, G} = \text{diag}(\langle v_{ki} \rangle) = \text{diag}(\langle w_{ki}^2 \rangle^{-1}). \quad (28)$$

Consequently, when the starting distributions of the variational algorithms are the same we have $\lim_{p \rightarrow 0} \Lambda_{k, p} = \Lambda_{k, G}$. Therefore, in the limiting case $p \rightarrow 0$, the posterior distributions associated with the l_p -prior and the independent Gaussian priors for each component of \mathbf{w}_k coincide.

5. INCREMENTAL AND ACTIVE LEARNING

Let us now assume that we want to add a new observation \mathbf{x}_{N+1} to the training set, whose corresponding $\mathbf{y}(\mathbf{x}_{N+1})$ will be provided by an oracle. To select \mathbf{x}_{N+1} we propose two active learning methods which are based on the posterior probabilities of the classes.

In the first method, called *Minimum Probability Criteria*, we select the next sample to be used to improve the classifier as

$$\mathbf{x}_{N+1} = \arg \min_{\mathbf{x}^*} (\max_k (p(C_k | \mathbf{x}^*))). \quad (29)$$

In the second method, named *Maximum Entropy Criteria*, we select the sample whose posterior distribution of the classes is less informative. Formally

$$\mathbf{x}_{N+1} = \arg \max_{\mathbf{x}^*} - \sum_{k=1}^K p(C_k | \mathbf{x}^*) \ln p(C_k | \mathbf{x}^*). \quad (30)$$

6. EXPERIMENTAL RESULTS

Due to space limitations, in this section we provide a limited number of experiments to analyze the performance of the proposed model for classification and AL.

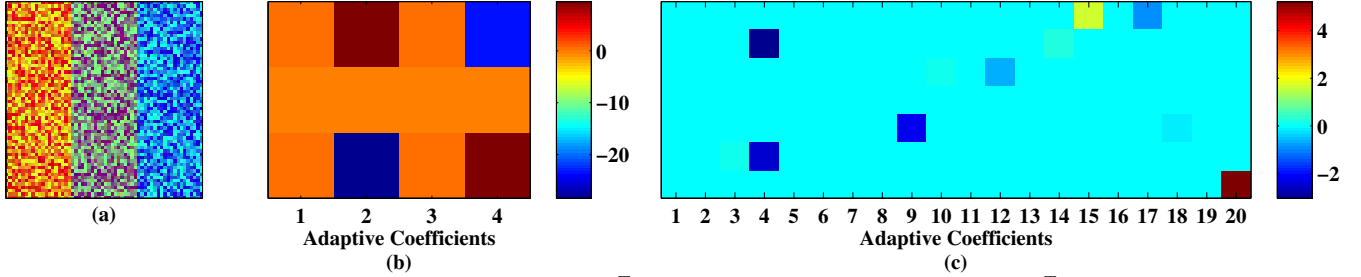


Fig. 1. (a) Original synthetic image. (b) Estimated \mathbf{W}^T for the synthetic dataset. (c) Estimated \mathbf{W}^T for the real dataset.

6.1. Supervised Classification results

Figure 1(a) shows a synthetically generated 60×60 image. The goal is to segment the three vertical rectangles in the image. Each rectangle represents one class in our segmentation problem. The pixels in each class are drawn from Gaussian distributions with mean vectors $\mu_1 = (0.9, 0.5, 0.1)^T$, $\mu_2 = (0.5, 0.5, 0.5)^T$ and $\mu_3 = (0.1, 0.5, 0.9)^T$, respectively. The three components of each pixel are normalized RGB values, each component is corrupted with noise of standard deviations 0.05, 0.5 and 0.05 respectively. Notice that the G band does not provide information to the classifier.

The experiment is repeated 10 times with 10 different training sets, each with 12 samples (4 from each class). As accuracy measure, the Cohen’s Kappa statistic (κ -index) is calculated on a test set of 1500 samples (500 from each class).

The values $p \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$ were tested. For values $p = 0.1, \dots, 0.6$ the obtained κ -index was 1, and 0.99 for the other values. Therefore, for $p = 0.1, \dots, 0.6$, the proposed method segments the synthetic image correctly on the test set. Fig. 1(b) shows the coefficients of the estimated adaptive matrix \mathbf{W}^T for $p = 0.1$. Non zero entries represent components relevant to classification. The proposed model does not use the G band and assigns zero to the corresponding adaptive coefficients.

We compare the proposed method with an SVM classifier. To perform a fair comparison we use a Gaussian kernel whose parameter is manually tuned to obtain the best performance. The SVM cost parameter is estimated using cross-validation. The obtained mean κ -index was 0.95, and therefore, the SVM classifier does not segment correctly the whole test sets from the synthetic image.

In our second classification experiment we evaluate the proposed Bayesian classifier on the real data set “Image Segmentation”, available on-line at the “UCI Machine Learning Repository” [20]. The goal is to classify a set of pixels in 7 classes: “BRICKFACE”, “SKY”, “FOLIAGE”, “CEMENT”, “WINDOW”, “PATH” and “GRASS”. The data set has 2310 samples (330 from each class). Each sample is a 19 component vector representing different attributes measured on a 3×3 neighborhood of the pixel of interest.

The experiment is repeated 10 times on 10 different training sets, each with 126 samples (18 from each class). The κ -index is calculated on a test set with 1050 samples (150 from each class). For $p = 1$, the obtained κ -index was 0.86. The best κ -index, 0.88, was obtained at $p = 0.02$, this implies that l_p -quasinorms with $p < 1$ can outperform the l_1 -norm.

Fig. 1(c) shows the absolute value of the estimated adaptive coefficients in \mathbf{W}^T . Components 9, 12, 14, 15, 17, 18, 20 correspond to attributes “horizontal edge mean”, “rawred-mean”, “rawgreen-mean”, “excess red”, “excess green”, “value-mean” and “hue-mean”, respectively. Attributes like “row” or “column”, which correspond to pixel position in the image, have no discriminative information. In those components, the estimated values of \mathbf{W} were

0 (second and third columns in the figure). The fourth component is “number of pixel where attributes were measured”, this component is equal to 9 for all samples, consequently the fourth component acts as the bias for each class while the first component, which was introduced for this purpose, takes the value zero. Interestingly, and as expected, if we remove the fourth component, the estimated values of the first components are the values of the fourth components multiplied by 9. Notice that because of the prior used, the classifier prefers to make zero the first component and assign small values to the adaptive coefficients of the fourth feature.

Finally we compare again the proposed method with an SVM classifier. Its mean κ -index was 0.84. Its performance is 0.02 and 0.04 lower than the proposed classifier for $p = 1$ and $p = 0.02$, respectively. Additionally we note that our proposed method does not need parameter tuning.

6.2. Active Learning results

To evaluate the performance of the proposed AL methods, we utilize learning curves. We start by training the classifier using Algorithm 1 on a reduced subset from the training set. The estimated adaptive matrix \mathbf{W} is then used to classify the test set, the κ -index is utilized as accuracy measure in the learning curves. Next, the AL methods proposed in Section 5 are used to select a new sample from the training set and the classifier updated.

The proposed AL methods in Sections 5 are noted MIN PRO (minimum probability) and MAX ENTRO (maximum entropy). They are compared to the following AL methods: margin sampling (SVM-MS) [9], entropy-query-by-bagging (SVM-EQB) [11] and multiclass-level uncertainty (SVM-MCLU) [10]. All of them use SVM as classifier. The cost parameter is estimated by cross-validation.

For the synthetic dataset, the experiment is repeated 10 times with 10 different initial training sets. The starting training set has 6 samples (2 from each class) and the whole training and test sets have 1500 samples (500 from each class). We use $p = 0.1$.

Figure 2(a) shows the mean κ -index learning curves. The proposed methods start at κ -index=0.91. Their learning rates are very fast, reaching κ -index=1 after adding only 2 samples to the initial training set. Both methods have the same behavior and perform better than randomly selecting the new samples from the training set and using the proposed classifier. The random approach does not reach κ -index=1 even after 20 samples have been added. Methods that use a SVM classifier start at κ -index=0.78, so they initially perform worse than our classification method. SVM-MCLU needs 5 to reach κ -index=1. SVM-EQB obtains a κ -index=1 when 11 samples have been added. Furthermore SVM-MS does not achieve κ -index=1 even when 20 samples have been added.

For the real dataset we use a test set with 1050 samples (150 from each class), the whole training set also contains 1050 samples

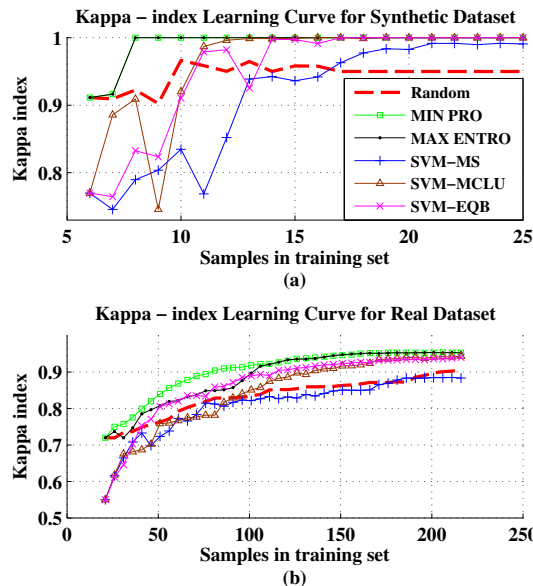


Fig. 2. (a) Learning curves for synthetic dataset. (b) Learning curves for real dataset.

(150 from each class). 10 initial training sets with 21 samples (3 from each class), are used. We use $p = 0.02$.

Figure 2(b) depicts the mean κ -index. The proposed methods start at 0.72 and reach κ -index=0.98 when the training set has 150 samples. After that the corresponding learning curves become flat. In this experiment MIN PRO outperforms MAX ENTRO, in particular notice the difference between both methods when we have less 100 samples. Both methods outperform random sampling which reaches κ -index=0.9 when 200 samples have been added.

The SVM classifiers utilize a Gaussian kernel whose parameters are manually tuned to obtain the performance. They start almost 0.15 below the proposed methods. SVM-MS does not perform well and its learning curve is similar to random sampling. SVM-MCLU and SVM-EQB performs similarly when 150 samples have been added and reach κ -index = 0.96. However SVM-EQB is better than SVM-MCLU for less than 150 samples. None of these methods outperformed the proposed ones.

7. CONCLUSIONS

In this work Bayesian modeling and inference have been used to address Supervised Classification and AL problems. The l_p -prior models utilized on the adaptive coefficients have promoted sparsity on the estimated adaptive parameters. Variational inference has been used to estimate all the model parameters and connections with independent Gaussian priors established. The predictive distribution of the classes has been calculated. This distribution has been used to define two AL methods. In the experimental section the proposed approach has been applied to Image Segmentation problems. Experimental results have shown that the use of l_p -priors allows the classifier to select discriminative features and discard non-relevance components. The proposed approach has shown higher accuracy than SVM methods in both classification and AL problems.

REFERENCES

[1] B. Schölkopf and A. Smola, *Learning with Kernels*, MIT Press, Cambridge, MA, 2002.

[2] C.M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer, 2007.

[3] M. E. Tipping, “The relevance vector machine,” *Journal of Machine Learning Research*, vol. 1, pp. 211–244, 2001.

[4] C. E. Rasmussen and C. K. Williams, *Gaussian Processes for Machine Learning*, MIT Press, NY, 2006.

[5] G. C. Cawley, N. L. C. Talbot, and M. Girolami, “Sparse multinomial logistic regression via bayesian l1 regularisation,” in *Neural Information Processing Systems*, 2006, pp. 209–216.

[6] G. Bouchard, “Efficient bounds for the softmax function and applications to approximate inference in hybrid models,” in *NIPS 2007*, 2007.

[7] N. Ahmed and M. Campbell, “Variational bayesian learning of probabilistic discriminative models with latent softmax variables,” *IEEE Trans. on Sig. Proc.*, vol. 59, no. 7, pp. 3143–3154, July 2011.

[8] S. Ryali, K. Supekar, D.A. Abrams, and V. Menon, “Sparse logistic regression for whole-brain classification of fmri data,” *NeuroImage*, vol. 51, no. 2, pp. 752–764, 2010.

[9] B. Settles, *Active Learning*, Morgan & Claypool, 2012.

[10] B. Demir, C. Persello, and L. Bruzzone, “Batch-mode active-learning methods for the interactive classification of remote sensing images,” *IEEE Trans. on Geoscience and Remote Sensing*, vol. 49, no. 3, pp. 1014–1031, March 2011.

[11] D. Tuia, F. Ratle, F. Pacifici, M.F. Kanevski, and W.J. Emery, “Active learning methods for remote sensing image classification,” *IEEE Trans. on Geoscience and Remote Sensing*, vol. 47, no. 7, pp. 2218–2232, July 2009.

[12] J. Paisley, X. Liao, and L. Carin, “Active learning and basis selection for kernel-based linear models: A Bayesian perspective,” *IEEE Trans. on Sig. Proc.*, vol. 58, pp. 2686–2700, 2010.

[13] P. Ruiz, J. Mateos, G. Camps-Valls, R. Molina, and A.K. Katsaggelos, “Bayesian active remote sensing image classification,” *IEEE Trans. on Geoscience and Remote Sensing*, vol. 52, no. 4, pp. 2186–2196, April 2014.

[14] A. Kabán and R.J. Durrant, “Learning with $l_{q<1}$ vs l_1 -norm regularization with exponentially many irrelevant features,” in *Proc. of ECML PKDD*, 2008, pp. 580–596, Springer-Verlag.

[15] Z. Liu, F. Jiang, G. Tian, S. Wang, F. Sato, S.J. Meltzer, and M. Tan, “Sparse logistic regression with L_p penalty for biomarker identification,” *Statistical Applications in Genetics and Molecular Biology*, 2007.

[16] S. D. Babacan, R. Molina, and A. K. Katsaggelos, “Bayesian compressive sensing using Laplace priors,” *IEEE Trans. on Image Processing*, vol. 19, no. 2, pp. 53–63, Jan. 2010.

[17] S. D. Babacan, R. Molina, and A. K. Katsaggelos, “Parameter estimation in TV image restoration using variational distribution approximation,” *IEEE Trans. on Image Processing*, vol. 17, no. 3, pp. 326–339, March 2008.

[18] J. Palmer, D. Wipf, K. Kreutz-Delgado, and B. Rao, “Variational EM algorithms for non-Gaussian latent variable models,” in *NIPS 2006*.

[19] R. T. Rockafellar, *Convex Analysis (Princeton Mathematical Series)*, Princeton University Press, 1970.

[20] K. Bache and M. Lichman, “UCI machine learning repository,” 2013.

Chapter 6

Other Related Problems

6.1 Light field Acquisition

6.1.1 Compressive Light Field Sensing

- S.D. Babacan, R. Ansorge, M. Luessi, **P. Ruiz**, R. Molina, and A. K. Kat-saggelos, “Compressive Light Field Sensing”, IEEE Transaction on Image Pro-cessing, vol. 21, no. 12, 4746-4757, December 2012.
 - Status: Published
 - Impact Factor (JCR 2013): 3.111
 - Subject Category: Computer Science, Artificial Intelligence (Q1: 14/121), Engineering, Electrical & Electronic (Q1: 27/248)

Compressive Light Field Sensing

S. Derin Babacan, *Member, IEEE*, Reto Ansorge, Martin Luessi, *Member, IEEE*,
Pablo Ruiz, *Student Member, IEEE*, Rafael Molina, *Member, IEEE*,
Aggelos K. Katsaggelos, *Fellow, IEEE*

Abstract

We propose a novel design for light field image acquisition based on compressive sensing principles. By placing a randomly coded mask at the aperture of a camera, incoherent measurements of the light passing through different parts of the lens are encoded in the captured images. Each captured image is a random linear combination of different angular views of a scene. The encoded images are then used to recover the original light field image via a novel Bayesian reconstruction algorithm. Using the principles of compressive sensing, we show that light field images with large number of angular views can be recovered from only a few acquisitions. Moreover, the proposed acquisition and recovery method provides light field images with high spatial resolution and signal-to-noise-ratio (SNR), and therefore does not suffer from limitations common to existing light field camera designs. We present a prototype camera design based on the proposed framework by modifying a regular digital camera. Finally, we demonstrate the effectiveness of the proposed system using experimental results with both synthetic and real images.

S. Derin Babacan is with the Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, IL, USA. e-mail: dbabacan@illinois.edu

Reto Ansorge is with Varian Medical Systems, Baden, Switzerland. e-mail: reto.ansorge@gmx.ch

Martin Luessi is with the Department of Radiology, Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Harvard Medical School, Boston, MA. e-mail: mluessi@nmr.mgh.harvard.edu

Pablo Ruiz Matarán and Rafael Molina are with the Departamento de Ciencias de la Computación e I.A. Universidad de Granada, Spain. e-mail: mataran@decsai.ugr.es, rms@decsai.ugr.es

Aggelos K. Katsaggelos is with the Department of Electrical Engineering and Computer Science, Northwestern University, IL, USA. e-mail: aggk@eecs.northwestern.edu

This work has been supported in part by the Beckman Institute postdoctoral fellowship, "Ministerio de Ciencia e Innovación" under contract TIN2010-15137, the Spanish research program Consolider Ingenio 2010: MIPRCV (CSD2007-00018), and a grant from the Department of Energy (DE-NA0000457).

Preliminary results of this work appeared at the IEEE Conference on Image Processing, Cairo, Egypt, 2009. [1]

I. INTRODUCTION

Recent advances in computational photography [2] provided effective solutions to a number of photographic problems, and also resulted in novel methods to acquire and process images. Novel camera designs allow for the capturing of information of the scene which is not possible to obtain using traditional cameras. This information can then be used for example to generate the three-dimensional scene geometry, or for novel applications, such as digital refocusing or synthetic aperture [2].

Light field cameras are one of the most widely used class of computational cameras. The light field expresses the radiance density function on the camera sensor, or the light energy of all rays in 3D space passing through the camera. For instance, a four-dimensional (4D) discrete light field image $\mathbf{x}(i, k, m, n)$ with spatial dimensions i, k and angular dimensions m, n contains images of a scene from a number of angles, which provide information about the 3D structure of the scene. Each 2D image $\mathbf{x}(i, k, m_0, n_0)$ with fixed angular coordinates m_0, n_0 is called an *angular image*. Traditional cameras integrate these angular images (or equivalently, light rays) over their 2D aperture to obtain the image, which results in the loss of valuable depth information about the scene. On the other hand, light field cameras capture the angular data and provide means to work directly with the light-rays instead of pixels, allowing one to produce many views of the scene, or perform many photographic tasks after the acquisition is made. This provides a clear advantage for light field imaging over traditional photography and makes many novel applications possible.

Compressive sensing (CS) [3], [4] has recently become very popular due to its interesting theoretical nature and wide area of applications. The theory of compressive sensing dictates that a signal can be recovered very accurately from a much smaller number of measurements than required by traditional methods, provided that it is *compressible* (or *sparse*) in some transform basis, i.e., only a few basis coefficients contain the major part of the signal energy. Besides sparsity, compressive sensing makes use of the incoherent measurement principle¹, and has led to many interesting theoretical results and novel applications (see, for instance, [5], [6]).

In this paper, we present a novel application of compressive sensing, namely, a novel framework to acquire light field images. We show that light field acquisition can be formulated using a incoherent measurement principle. We then demonstrate that light field images have a highly sparse nature, which, in combination with incoherent measurements, can be exploited to reconstruct the light field images with

¹Loosely speaking, incoherent measurements refer to non-adaptive and uncorrelated with the signal of interest. See, for instance, [3], [5] for technical definition and interpretations.

much fewer image acquisitions than traditionally required. By exploiting this sparsity in light field images, we develop a novel reconstruction algorithm that recovers the original images from few compressive measurements with a very high degree of fidelity.

In addition, we propose a novel camera design based on the developed acquisition framework. We build our design on ideas from coded aperture imaging, computational photography and compressive sensing. By exploiting the fact that different regions of the aperture of a camera correspond to images of the scene from different angles, we incorporate a compressively coded mask placed at the aperture to obtain incoherent measurements of the incident light field. These measurements are then decoded using the proposed reconstruction algorithm to recover the original light field image. We exploit the highly sparse nature of the light field images to obtain accurate reconstructions with a small number of measurements compared to the high angular dimension of the light field image. The proposed camera design provides images with high signal-to-noise ratio and does not suffer from the spatio-angular resolution trade-off in most existing light field camera designs. Finally, we demonstrate the efficiency of the proposed framework with both synthetic experiments and real images captured by a prototype camera.

The paper is organized as follows. First we review related prior work in light field and coded aperture imaging in Sec. II. In Sec. III we present the proposed acquisition method to obtain incoherent measurements of the light field image. We model the acquisition system and the light field images using a Bayesian framework, which is described in Sec. IV. The Bayesian inference procedure used to develop the reconstruction algorithm is presented in Sec. V. We present a prototype light field camera based on the proposed framework in Sec. VI. The effectiveness of the proposed system is demonstrated with both synthetic and real light field images in Sec. VII and conclusions are drawn in Sec. VIII.

II. RELATED PRIOR WORK

A. Light Field Acquisition

Light field acquisition, based on the principles of integral photography, was first proposed over a century ago [7], [8]. The same ideas appeared in the computer vision literature first as the *plenoptic camera* [9], and then the potential of light field imaging was demonstrated in [10] and [11]. The original design in [9] is implemented in a hand-held camera in [12], where a microlens (lenticular) array is placed between the main lens and the camera sensor. A similar approach is proposed in [13], where instead of using microlenses, a lens array is placed in front of the camera main lens. In both approaches, the light field image is captured using one acquisition. The additional lens array is used to capture the angular information, and reordering the captured image results in images of different views of the scene. Other

proposed light field camera designs include multi-camera systems [14] and mask-based designs [15], [16], which encode the angular information using frequency-multiplexing.

Many of these designs suffer from the spatio-angular resolution² trade-off [13], that is, one cannot obtain light field images with both high spatial- and high angular resolution. This problem is inherent in designs with one recording sensor (or film) and where only one acquisition is made. If the captured light field image has an angular resolution of $N_h \times N_v$, and a spatial resolution of $P_h \times P_v$, then $N_h \times N_v \times P_h \times P_v$ can only be less than or equal to the number of pixels in the camera sensor. For instance, a typical light field image captured using the plenoptic camera in [12] provides 14x14 angular images of size approximately 300x300 in a 16 megapixel camera. Multi-camera systems [14] are not affected from the spatio-angular resolution trade-off, but they are very costly to implement and cumbersome for practical usage.

Recently, a *programmable aperture camera* is proposed [17], where a binary mask is used to code the aperture. Angular images are *multiplexed* into single 2D images similarly to the principle of coded aperture imaging. After multiple acquisitions are made, a linear estimation procedure is employed to recover the full light field image. Although this design captures images with both high spatial and angular resolution, the number of acquisitions are equal to the number of angular dimensions. Therefore, obtaining a light field image with a high angular resolution is not practical.

During the development of this work, we became aware of [18], which appeared after [1], and independently considered the application of compressive sensing to light field acquisition. The work in [18] devises a linear recovery procedure from compressive measurements incorporating statistical correlations among the angular images via their autocorrelation matrix. In contrast, in this work we exploit the structure within the light field more explicitly using nonlinear relationships among the angular images.

B. Coded Aperture Imaging

Coded aperture imaging is developed in order to collect more light in situations where a lens system cannot be used, due to the measured wavelengths. Imaging systems with coded apertures are currently widely used in astronomy and medicine. The technique is based on the principle of pinhole cameras, but instead of only one pinhole which suffers from low SNR, a specially designed array of pinholes is used.

²The spatial and angular resolution here only refer to the number of digitally acquired elements such as pixels and images. Certain optical effects such as diffraction due to aperture size are not included in this analysis.

This array of pinholes provides images that are overlapping copies of the original scene, which can then be decoded using computational algorithms to provide a sharp image. There is a vast literature on coded aperture methods in astronomy and medicine (see, for example, [19], [20]).

Recent works considered coded aperture methods for developing novel image acquisition methods. In [21], the aperture is coded in the time-domain to modify the exposure for motion deblurring. Spatially modifying the aperture has been used for a range of applications: Levin *et. al.* [22] proposed utilizing an aperture mask to reconstruct both the original image and the depth of the scene from a single snapshot. A lensless imaging system is proposed in [23] that allows for the manipulation of the captured scene in ways not possible by traditional cameras, such as splitting field of view. Nayar *et. al.* [24] used a spatial light modulator to control the exposure per pixel, which can be used to obtain high-dynamic range images. Other uses of coded apertures include super resolution [25] and range estimation [26].

Compressive sensing methods have also been applied in conjunction with coded apertures or compressively coded blocking masks. Novel imaging methods have been proposed for spectral imaging [27], dual-photography [28], and the design of structured light for recovering inhomogeneous participating media [29]. Most recently, compressively coded aperture masks are used for single-image super-resolution and shown to provide higher quality images than traditional coded apertures [30], [31].

A related approach to coded aperture imaging is *wavefront coding* [32], where the image is intentionally defocused using phase plates so that the defocus is uniform throughout the image. The captured image can then be deconvolved to obtain an image with an enlarged depth of field.

III. COMPRESSIVE SENSING OF LIGHT-FIELDS

In this section, we will show that light field image acquisition can be formulated within the compressive sensing framework. We first show that light field images can be acquired by coding the camera aperture, and then present the proposed compressive acquisition system. In the following, a 4D light field image is denoted by \mathbf{x} , which is the collection of N angular images \mathbf{x}^j , such that $\mathbf{x} = \{\mathbf{x}^j\}, j = 1, \dots, N$.

A. Light-Field Acquisition by Coded Apertures

A fundamental principle used in this work is that different regions of the aperture capture images of the scene from different angles³ [17], [23], [34], [35]. Specifically, the main camera lens can be interpreted

³This is a widely used model employing a geometric optics perspective. A more recent analysis of light fields based on wave optics provide additional views on the transformation of light fields through lenses [33].

as an array of multiple virtual lenses (or cameras). This concept is illustrated in Fig. 1(a)-(c), where only certain parts (white blocks) of the aperture are left open. As can be seen from Fig. 1(a)-(c), the acquired images exhibit vertical and horizontal parallax. By separately opening one region of the aperture and blocking light in the others, the complete light field with an angular dimension of N can be captured with N exposures. However, obtaining the light field image in this fashion has two disadvantages: First, due to the very small amount of light arriving to the sensor at each exposure, the captured angular images have very low signal-to-noise ratios (SNR). Second, a large number of acquisitions have to be made in order to obtain high angular resolution. The programmable aperture camera design in [17] addressed the first problem by incorporating a multiplexing scheme, but the second problem remains a serious drawback.

We address both of these issues by using a randomly coded non-refractive mask in front of the aperture. Each image acquired in this fashion is a random linear combination (and therefore an incoherent measurement) of the angular images. An example image captured in this fashion is illustrated in Fig. 1(d), where the amount of light passing through different regions of the aperture are randomly selected (shown at the bottom of Fig. 1(d)). As shown in the following, using such a random mask overcomes both of the problems described above.

The mathematical principle behind this idea is formulated as follows. Let us assume that the aperture of the main camera lens is divided into N blocks, with $N = N_h \times N_v$ where N_h and N_v represent the number of horizontal and vertical divisions. During each acquisition i , each block j is assigned a weight $0 \leq a^{ij} \leq 1$ which controls the amount of light passing through this block. Therefore, a^{ij} represents the transmittance of the block j , i.e., it is the fraction of incident light that passes through the block. As mentioned above, each block j captures an angular image \mathbf{x}^j in the light field image, and therefore the acquired image \mathbf{y}^i at the i^{th} acquisition can be represented as a linear combination of the N angular images as

$$\mathbf{y}^i = \sum_{j=1}^N a^{ij} \mathbf{x}^j, \quad i = 1, \dots, M, \quad (1)$$

where we use the vector notation such that \mathbf{y}^i and \mathbf{x}^j are both $P \times 1$ vectors, with P the number of pixels in each image. Note that in a traditional camera, the acquired image is the average of all angular images, i.e., $a^{ij} = \frac{1}{N}$, since the aperture integrates all light rays coming from different directions.

After M acquisitions ($M \leq N$), the complete set of observed images $\{\mathbf{y}^i\}$ can be expressed in

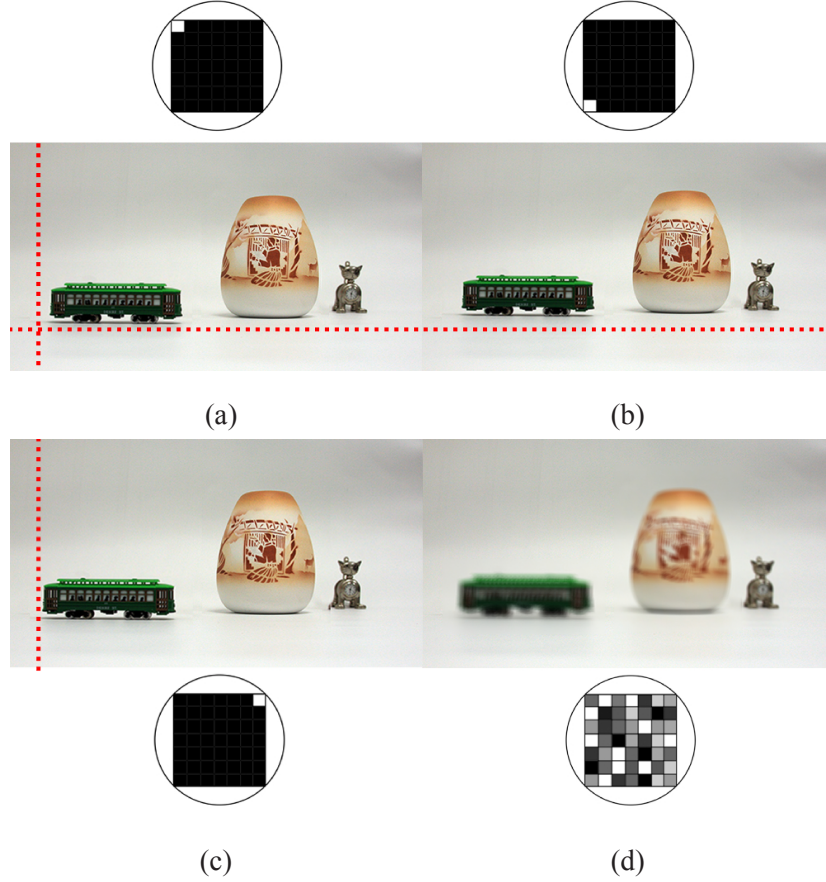


Fig. 1. The basic principle of utilizing a coded aperture to obtain light field images. The angular images are shown in (a), (b) and (c) when only corner blocks of the aperture are left open. Both horizontal and vertical parallax can be observed between these images (horizontal and vertical dashed lines are shown to denote the vertical and horizontal parallax, respectively). Figure (d) shows a captured image with the randomly coded aperture used in the proposed compressive sensing light field camera. All images are from a synthetic light field image (see Sec. VII).

matrix-vector form as

$$\begin{pmatrix} \mathbf{y}^1 \\ \mathbf{y}^2 \\ \cdot \\ \cdot \\ \mathbf{y}^M \end{pmatrix} = \begin{pmatrix} a^{11}\mathbf{I} & a^{12}\mathbf{I} & \cdot & \cdot & a^{1N}\mathbf{I} \\ a^{21}\mathbf{I} & a^{22}\mathbf{I} & \cdot & \cdot & a^{2N}\mathbf{I} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ a^{M1}\mathbf{I} & a^{M2}\mathbf{I} & \cdot & \cdot & a^{MN}\mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{x}^1 \\ \mathbf{x}^2 \\ \cdot \\ \cdot \\ \mathbf{x}^N \end{pmatrix}, \quad (2)$$

with \mathbf{I} the $P \times P$ identity matrix. The system in (2) is expressed in a more compact form as

$$\mathbf{y} = \mathbf{A}\mathbf{x}, \quad (3)$$

with

$$\mathbf{A} = \begin{pmatrix} a^{11} & a^{12} & \dots & a^{1N} \\ a^{21} & a^{22} & \dots & a^{2N} \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ a^{M1} & a^{M2} & \dots & a^{MN} \end{pmatrix} \otimes \mathbf{I} = \hat{\mathbf{A}} \otimes \mathbf{I}, \quad (4)$$

where \otimes is the Kronecker product. Taking also the acquisition noise into account, the final observation model can be expressed as

$$\mathbf{y} = \left(\hat{\mathbf{A}} \otimes \mathbf{I} \right) \mathbf{x} + \mathbf{n} = \mathbf{A} \mathbf{x} + \mathbf{n}, \quad (5)$$

with \mathbf{n} the $PM \times 1$ noise vector.

B. Compressively Coded Apertures

If the linear measurement matrix \mathbf{A} satisfies certain properties dictated by the theory of compressive sensing [3], the light field acquisition system in (5) can be seen as a noisy incoherent measurement system. A sufficient condition for a matrix to be a compressive sensing matrix is the *restricted isometry property* (RIP) [3], [36], which is proven to hold with a very high probability for a general class of matrices with their entries drawn from certain random probability distributions. For instance, if $\hat{\mathbf{A}}$ in (5) is constructed by independently drawing its entries from a Gaussian distribution, then $\hat{\mathbf{A}}$ satisfies RIP with an overwhelming probability.

It is straightforward to show that if $\hat{\mathbf{A}}$ is a valid compressive sensing matrix, then \mathbf{A} is a valid compressive sensing matrix as well. A simple proof is as follows. The mutual coherence of matrix $\hat{\mathbf{A}}$ is given by [37]

$$\mu(\hat{\mathbf{A}}) = \max_{i \neq j} \frac{|\hat{A}_i^T \hat{A}_j|}{\|\hat{A}_i\| \|\hat{A}_j\|}, \quad (6)$$

where \hat{A}_i is the i^{th} column of $\hat{\mathbf{A}}$. The mutual coherence characterizes the correlation between the columns of matrix $\hat{\mathbf{A}}$, and it is always positive for matrices with more columns than rows. It is shown that the mutual coherence provides a bound for the RIP constants [38], and therefore RIP-based guarantees can be applied using mutual coherence. Using properties of the Kronecker product, it can be seen that

$$\mathbf{A}^T \mathbf{A} = \left(\hat{\mathbf{A}} \otimes \mathbf{I} \right)^T \left(\hat{\mathbf{A}} \otimes \mathbf{I} \right) \quad (7)$$

$$= \hat{\mathbf{A}}^T \hat{\mathbf{A}} \otimes \mathbf{I}. \quad (8)$$

Thus, the inner products of columns of \mathbf{A} have the exact same values as the columns of $\hat{\mathbf{A}}$, and therefore they have the same mutual incoherence. If the mutual coherence of $\hat{\mathbf{A}}$ is sufficiently small so as to satisfy RIP [38], \mathbf{A} also satisfies the restricted isometry property and it is therefore a valid compressive sensing matrix.

Based on this, the acquisition system in (5) is an incoherent measurement system of angular images \mathbf{x}^j , where each acquired image is a random linear combination of the angular images. The theory of compressive sensing then dictates that if the unknown image \mathbf{x} can be represented sparsely in some transform domain, then it can be recovered with much fewer measurements than traditionally required ($M \ll N$). Due to the nature of multi-view images and especially in the specific case considered in this work where the angular images are aligned on a small-baseline, the redundancy within the light field images is very high. In fact, there are multiple sources of sparsity inherent in light field images, due to correlations both within and in between the angular images (see Sec. IV-B for details). Therefore, light field images can be very accurately reconstructed with very few acquisitions by utilizing the compressive acquisition system in (5) and by exploiting their sparse nature within nonlinear reconstruction frameworks.

An important design issue is the selection of the measurement matrix \mathbf{A} , which determines the level of incoherence of the measurements and therefore the reconstruction performance. The design of measurement matrices for compressive sensing is an active area of research, and many of the existing designs can be used for the proposed aperture mask. In this work, we specifically experimented with two different types of measurement matrices, namely, uniform spherical and scrambled Hadamard ensembles [39]. If fractional values of the block transmittances are permitted, a general class of matrices can be utilized, with positivity of the matrix entries as the only constraint. In this case, uniform spherical ensembles (with values ranging between 0 and 1) are very suitable as measurement matrices. If the mask is limited to binary codes, scrambled Hadamard ensembles can be used to code the aperture. Moreover, the measurement matrices can also be selected depending on specific requirements of the optical systems, e.g., the expected amount of transmitted light can be varied by varying the mean value of the corresponding probability distribution, or by choosing a specific construction of the random measurement matrix.

It should be noted that since many (or possibly all) blocks are open in each exposure, each captured image has a high SNR due to the small amount of loss of light. In fact, the measurement matrices can be designed to optimize the amount of passing light while maintaining the random structure. Moreover, as shown in the experimental results section, incorporating a nonlinear reconstruction mechanism provides images with much higher SNRs than those of linear reconstruction methods, such as demultiplexing.

Finally, it should be noted that the coded aperture setup used in this work is a specific application of

the acquisition system in (5). The proposed compressive sensing formulation for light field acquisition can be applied to a wider range of light field imaging applications. For instance, multiple camera or multiple lens imaging systems such as camera arrays and stereo cameras can equally well incorporate the incoherent measurement system in (5) and significantly reduce the number of acquisitions without sacrificing spatial or angular resolution.

IV. HIERARCHICAL BAYESIAN MODEL FOR RECONSTRUCTION

In order to reconstruct the angular images $\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^N$ from the incoherent measurements $\mathbf{y}^1, \mathbf{y}^2, \dots, \mathbf{y}^M$ and \mathbf{A} , both the observation process (5) and the unknown light field image \mathbf{x} have to be modeled. For this modeling, we use the hierarchical Bayesian framework by employing a conditional distribution $p(\mathbf{y}|\mathbf{x}, \beta)$ for the observation model in (5) and a *prior* distribution $p(\mathbf{x}|\alpha_{\text{TV}}, \alpha_c)$ on the unknown light field image \mathbf{x} . These distributions depend on the model parameters β , α_{TV} and α_c , which are called *hyperparameters*. In the second stage of the hierarchical model we use additional prior distributions, called *hyperpriors*, to model them. In the following subsections, we present the specific forms of each of these distributions.

A. Observation (Noise) Model

The observation noise is assumed to be independent and Gaussian with zero mean and variance equal to β^{-1} , that is, using (5),

$$p(\mathbf{y}|\mathbf{x}, \beta) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x}, \beta^{-1}). \quad (9)$$

B. Light-Field Image Model

The choice of randomly programmed coded apertures makes the exact/approximate recovery of the angular images possible through the use of sparsity inherent in light field images. There are two sources of sparsity within a light field image that can be exploited. The first one is sparsity within each angular image. It is already well known that two-dimensional images can be very accurately represented by only a small number of coefficients of a *sparsifying* transform, such as wavelet transforms or total-variation (TV) function applied on the image. In the case of light field images, there is another fundamental source of sparsity, that is, the angular images are very closely related to each other. Specifically, each angular image can be accurately estimated from another one using dense warping (or correspondence) fields, as shown below.

Based on the above, we use the following factorized form of the prior distribution

$$p(\mathbf{x}|\alpha_{\text{TV}}, \alpha_c) = p(\mathbf{x}|\alpha_{\text{TV}}) p(\mathbf{x}|\alpha_c) C(\alpha_{\text{TV}}, \alpha_c), \quad (10)$$

where $p(\mathbf{x}|\alpha_{\text{TV}})$ is the TV image prior employed on each angular image separately, $p(\mathbf{x}|\alpha_c)$ is the prior that models the sparsity arising from the strong dependency between angular images and $C(\alpha_{\text{TV}}, \alpha_c)$ is a function of the unknown hyperparameters needed for the image prior model to integrate to one. In this work, we assume $C(\alpha_{\text{TV}}, \alpha_c)$ is constant.

Next we describe the specific models used for each of the prior distributions in this factorization.

1) *Total Variation Image Prior*: The angular images \mathbf{x}^i are natural images, hence they are expected to be mostly smooth except at a number of discontinuities (e.g., spatial edges). As spatial domain image priors, we employ the total variation function which, due to its edge-preserving property, does not over-penalize discontinuities in the image while imposing smoothness [40]. Specifically, $p(\mathbf{x}|\alpha_{\text{TV}})$ is expressed as

$$p(\mathbf{x}|\alpha_{\text{TV}}) \propto \prod_{i=1}^N (\alpha_{\text{TV}}^i)^{P/2} \exp \left[-\frac{1}{2} \alpha_{\text{TV}}^i \text{TV}(\mathbf{x}^i) \right], \quad (11)$$

with

$$\text{TV}(\mathbf{x}^i) = \sum_k \sqrt{(\Delta_k^h(\mathbf{x}^i))^2 + (\Delta_k^v(\mathbf{x}^i))^2}, \quad (12)$$

where Δ_k^h and Δ_k^v correspond to, respectively, horizontal and vertical first order differences, at pixel k , that is, $\Delta_k^h(\mathbf{x}^i) = (\mathbf{x}^i)_k - (\mathbf{x}^i)_{l(k)}$ and $\Delta_k^v(\mathbf{x}^i) = (\mathbf{x}^i)_k - (\mathbf{x}^i)_{a(k)}$, where $l(k)$ and $a(k)$ denote the nearest neighbors of pixel k , to the left and above, respectively.

2) *Cross-image prior*: As mentioned above, there is a high correlation between angular images in the light field image. Specifically, disregarding occlusions, each angular image \mathbf{x}^i can be very closely approximated by another angular image \mathbf{x}^j using the dense warping field \mathbf{M}^{ij} between i and j , that is, $\mathbf{x}^i \approx \mathbf{M}^{ij} \mathbf{x}^j$. Therefore, the dependency of each angular image on another one is very strong and can be exploited while modeling \mathbf{x} . Based on this, we use the following cross-image prior between angular images

$$p(\mathbf{x}|\alpha_c) \propto \exp \left(\sum_{i=1}^N \sum_{j \in \Omega(i)} -\frac{\alpha_c^{ij}}{2} \|\mathbf{x}^i - \mathbf{M}^{ij} \mathbf{x}^j\|_{\mathbf{O}^{ij}}^2 \right), \quad (13)$$

$$= \exp \left(\sum_{i=1}^N \sum_{j \in \Omega(i)} -\frac{\alpha_c^{ij}}{2} (\mathbf{x}^i - \mathbf{M}^{ij} \mathbf{x}^j)^T \mathbf{O}^{ij} (\mathbf{x}^i - \mathbf{M}^{ij} \mathbf{x}^j) \right), \quad (14)$$

where α_c^{ij} is the precision of the registration error, and \mathbf{O}^{ij} is a diagonal matrix with 0 and 1's on the diagonal to account for occlusions. In the occluded areas, the corresponding entries are set equal to zero, and the remaining entries equal to 1. This usage of the weighted norm is equivalent to the assumption that $\mathbf{O}^{ij} \mathbf{x}^i \approx \mathbf{O}^{ij} \mathbf{M}^{ij} \mathbf{x}^j$, that is, the angular image \mathbf{x}^i can be closely approximated by the warped angular

image $\mathbf{M}^{ij} \mathbf{x}^j$ except at the occluded areas. Notice that the occluded areas (hence matrices \mathbf{O}^{ij}) can easily be extracted if the warping fields \mathbf{M}^{ij} are known.

In (14), $\Omega(i)$ defines a neighborhood of \mathbf{x}^i , which consists of the angular images with closest viewpoints to that of \mathbf{x}^i (a maximum of 8 images on a rectangular grid). In other words, angular images captured by nearby regions in the aperture are treated as neighboring images. This neighborhood is imposed in (14) for several reasons. First, angular images far apart in the aperture can be less accurately related by dense warping fields due to the 3D structure of the scene and increased size of the occluded areas. Second, incorporating a cross-image prior between each pair of angular images in \mathbf{x} largely increases memory requirements and therefore it is computationally not efficient during the reconstruction phase. Finally, since \mathbf{x}^i is part of at least one neighborhood defined on \mathbf{x} , the warping constraint is propagated to all angular images during the reconstruction algorithm.

The cross-image prior in (14) can be written in matrix-vector form as

$$p(\mathbf{x}|\boldsymbol{\alpha}_c) = z_c \exp\left(-\frac{1}{2} \mathbf{x}^T \Pi \mathbf{x}\right), \quad (15)$$

where z_c is the partition function, and Π is a sparse $NP \times NP$ matrix constructed from $N \times N$ blocks of size $P \times P$. Its explicit form is given by

$$\Pi = \begin{pmatrix} \Pi_{11} & \Pi_{12} & \cdot & \cdot & \Pi_{1N} \\ \Pi_{21} & \Pi_{22} & \cdot & \cdot & \Pi_{2N} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \Pi_{N1} & \Pi_{N2} & \cdot & \cdot & \Pi_{NN} \end{pmatrix}. \quad (16)$$

The $P \times P$ block Π_{ij} can be obtained from (14) as

$$\Pi_{ij} = \begin{cases} \sum_{s \in \Omega(i)} \alpha_c^{is} \mathbf{O}^{is} + \alpha_c^{si} (\mathbf{M}^{si})^T \mathbf{O}^{si} \mathbf{M}^{si} & \text{if } i = j \\ -\alpha_c^{ij} \mathbf{O}^{ij} \mathbf{M}^{ij} - \alpha_c^{ji} \mathbf{O}^{ij} (\mathbf{M}^{ji})^T & \text{if } j \neq i, j \in \Omega(i) \\ 0 & \text{else} \end{cases}$$

The form of the matrix Π makes the calculation of the partition function z_c of the distribution in (15) intractable. To overcome this difficulty, we approximate the partition function by a quadratic form, and use the following as the cross-image prior

$$p(\mathbf{x}|\boldsymbol{\alpha}_c) = c \left[\prod_{i,j} (\alpha_c^{ij})^{P/2} \right] \exp\left(-\frac{1}{2} \mathbf{x}^T \Pi \mathbf{x}\right), \quad (17)$$

with c being a constant.

It is clear that incorporating the cross-image prior requires knowledge of the dense warping fields \mathbf{M}^{ij} , which cannot be directly obtained from the compressive measurements. In this work, we overcome this problem by acquiring two additional images from two opposite diagonal sides of the aperture. These images exhibit full horizontal and vertical parallax, and a dense registration algorithm based on graph-cuts [41] is employed to obtain the warping field from them. Due to the uniform partitioning of the aperture, this warping field can be used to obtain approximate intermediate warping fields between all angular images. The disadvantage of this approach is that two additional exposures have to be taken with small apertures (and therefore with low SNR), and combined with the approximate calculation of the intermediate warping fields, the constraints imposed in the cross-image prior might not fully characterize the actual relations within the light field image. However, our experiments have shown that this approach provides accurate enough warping fields so that accurate reconstructions are obtained. Moreover, estimating the precision variables α_c^{ij} along with the image compensates for the inaccuracies in the warping fields during reconstruction.

An alternative method is to use $\mathbf{x}^i \approx \mathbf{x}^j$, which is similar to the approximation used in the compressive video sensing algorithm in [42]. Although this method does not require knowledge of the warping fields, it is a very crude approximation and therefore does not provide reconstruction results comparable to the ones reported here. However, it can be used with relatively high performance in the case of very densely packed angular images, since the variation between two neighboring angular images will be very small.

It should be emphasized that the modeling in (14) is an approximation to the structure within the light field image. It implicitly assumes that the scene is Lambertian, and that the occluded areas between neighboring angular images are relatively small in size. Nevertheless, it provides a close approximation to the light-field structure (especially with small-baseline angular images as considered in this paper), and as shown in the experimental results section, it leads to a high reconstruction performance. Without such an enforcement of the internal structure of the light-field (i.e., without the use of the cross-image priors) and by only using separate image priors on the angular images, accurate reconstructions cannot be obtained. On the other hand, the role of the intra-image priors is to individually impose smoothness on the angular image estimates while preserving the sharp image features, and the advantages of employing them are demonstrated in a number of works in the literature (see, e.g., [43]).

C. Hyperpriors on the Hyperparameters

The form of the hyperprior distributions on the hyperparameters β , α_{TV} and α_c determines the ease of calculation of the posterior distribution $p(\mathbf{x}, \beta, \alpha_{\text{TV}}, \alpha_c | \mathbf{y})$. Since the distributions $p(\mathbf{y} | \mathbf{x}, \beta)$ and $p(\mathbf{x} | \alpha_c)$ are Gaussian distributions, and we will approximate the distribution $p(\mathbf{x} | \alpha_{\text{TV}})$ by a Gaussian distribution (shown in Section V), we chose to utilize Gamma distributions for all hyperparameters, as it is the conjugate prior for the inverse variance (precision) of the Gaussian distribution [44]. Thus, the hyperprior distributions are given by

$$p(\beta) = \Gamma(\beta | a^\circ, b^\circ) = \frac{(b^\circ)^{a^\circ}}{\Gamma(a^\circ)} \beta^{a^\circ-1} \exp[-b^\circ \beta] \quad (18)$$

$$p(\alpha_{\text{TV}}^i) = \Gamma(\alpha_{\text{TV}}^i | a^\circ, b^\circ), \quad i = 1, \dots, N \quad (19)$$

$$p(\alpha_c^{ij}) = \Gamma(\alpha_c^{ij} | a^\circ, b^\circ), \quad i = 1, \dots, N, j \in \Omega(i) \quad (20)$$

with identical shape and inverse scale parameters a° and b° , respectively. These parameters are set equal to small values (e.g., 10^{-5}) to make the hyperpriors vague, which makes the estimation process depend more on the observations than the prior knowledge. Note, however, that if some prior knowledge about the hyperparameters is available (for example, approximate values of the noise variances in the observations), this knowledge can easily be incorporated into the estimation procedure using appropriate values of the shape and inverse scale parameters (see, for example, [43]).

V. RECONSTRUCTION ALGORITHM

Let us denote by $\Theta = \{\beta, \alpha_{\text{TV}}, \alpha_c, \mathbf{x}\}$ the set of all unknowns. The Bayesian inference is based on the posterior distribution

$$p(\Theta | \mathbf{y}) = p(\mathbf{x}, \beta, \alpha_{\text{TV}}, \alpha_c | \mathbf{y}) = \frac{p(\mathbf{x}, \beta, \alpha_{\text{TV}}, \alpha_c, \mathbf{y})}{p(\mathbf{y})}, \quad (21)$$

where $p(\beta, \alpha_{\text{TV}}, \alpha_c, \mathbf{x}, \mathbf{y})$ is given by

$$p(\mathbf{y}, \mathbf{x}, \beta, \alpha_{\text{TV}}, \alpha_c) = p(\mathbf{y} | \mathbf{x}, \beta) p(\mathbf{x} | \alpha_{\text{TV}}, \alpha_c) p(\beta) p(\alpha_{\text{TV}}) p(\alpha_c). \quad (22)$$

Unfortunately, the posterior $p(\Theta | \mathbf{y})$ is intractable (since $p(\mathbf{y})$ is intractable), and therefore approximations are utilized. A common approximation is to represent the posterior by a delta function at its mode. Then, using $p(\mathbf{x} | \mathbf{y}, \Theta) \propto p(\Theta, \mathbf{y})$, the unknowns can be found by

$$\Theta = \arg \max_{\Theta} p(\Theta | \mathbf{y}) = \arg \max_{\Theta} p(\Theta, \mathbf{y})$$

Note that this formulation results in the well-known *maximum a posteriori* (MAP) estimate of Θ . Specifically, assuming uniform hyperpriors on the hyperparameters, the estimates found by this inference procedure are equivalent to the solution of the following regularized inverse problem:

$$\Theta = \arg \min_{\Theta} \left[\beta \|\mathbf{y} - \mathbf{A}\mathbf{x}\|^2 + \sum_{i=1}^N \alpha_{\text{TV}}^i \text{TV}(\mathbf{x}^i) + \sum_{i=1}^N \sum_{j \in \Omega(i)} \alpha_c^{ij} \|\mathbf{x}^i - \mathbf{M}^{ij} \mathbf{x}^j\|_{\mathbf{O}^{ij}}^2 + \log z_{\alpha} \right], \quad (23)$$

where $z_{\alpha} = \prod_{i,j} (\alpha_{\text{TV}}^i)^{P/2} (\alpha_c^{ij})^{P/2}$ represents all (approximate) normalizing terms in $p(\mathbf{x}|\alpha_{\text{TV}}, \alpha_c)$. Therefore, existing methods for TV-regularized optimization can also be employed for solving the recovery problem (see, for example, [45], [46]). However, even with the MAP approximation, the calculation of the hyperparameters is hard due to the use of the TV priors. Therefore, we resort to the majorization-minimization method proposed in [43]. We omit the details of the derivations here, and provide only the form of the updates of each unknown variable.

The estimate for the light field image $\hat{\mathbf{x}}$ can be calculated as

$$\hat{\mathbf{x}} = \Sigma_{\mathbf{x}} \beta \mathbf{A}^T \mathbf{y}, \quad (24)$$

$$\Sigma_{\mathbf{x}}^{-1} = \text{diag} \left(\alpha_{\text{TV}}^i (\Delta^h)^T \mathbf{W}_{\text{TV}}^i (\Delta^h) + \alpha_{\text{TV}}^i (\Delta^v)^T \mathbf{W}_{\text{TV}}^i (\Delta^v) \right) + \Pi + \beta \mathbf{A}^T \mathbf{A}, \quad (25)$$

where the first matrix term in (25) is a $NP \times NP$ block diagonal matrix created by $P \times P$ blocks $\alpha_{\text{TV}}^i (\Delta^h)^T \mathbf{W}_{\text{TV}}^i (\Delta^h) + \alpha_{\text{TV}}^i (\Delta^v)^T \mathbf{W}_{\text{TV}}^i (\Delta^v)$. The matrices \mathbf{W}_{TV}^i are calculated by

$$\mathbf{W}_{\text{TV}}^i = \text{diag} \left(\frac{1}{\sqrt{(\mathbf{w}_{\text{TV}}^i)_k}} \right), \quad k = 1, \dots, P \quad (26)$$

where

$$(\mathbf{w}_{\text{TV}}^i)_k = (\Delta_k^h(\hat{\mathbf{x}}^i))^2 + (\Delta_k^v(\hat{\mathbf{x}}^i))^2. \quad (27)$$

It is clear that the vector \mathbf{w}_{TV}^i (and hence the matrix \mathbf{W}_{TV}^i) represents the local spatial activity in each angular image \mathbf{x}^i using its total variation. The estimates of the hyperparameters are given by

$$\beta = \frac{\frac{1}{2}NP + a^o - 1}{\frac{1}{2} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|^2 + b^o}, \quad (28)$$

$$\alpha_{\text{TV}}^i = \frac{\frac{1}{2}P + a^o - 1}{\sum_k (\mathbf{w}_{\text{TV}}^i)_k + b^o}, \quad (29)$$

$$\alpha_c^{ij} = \frac{\frac{1}{2}P + a^o - 1}{\frac{1}{2} \|\mathbf{x}^i - \mathbf{M}^{ij} \mathbf{x}^j\|_{\mathbf{O}^{ij}}^2 + b^o}. \quad (30)$$

Finally, the algorithm iterates among estimating the light field image using (24), the spatial adaptivity vectors using (27), and the hyperparameters using (28)-(30) until convergence.

VI. PROTOTYPE LIGHT FIELD CAMERA

We have assembled a prototype of the proposed system as shown in Fig. 2(a). A binary LCD array (Electronic Assembly DOGL128S-6), shown in Fig. 2(b), is mounted to the lens of a digital camera. The LCD array consists of 128×64 pixels and we used 8×8 pixel segments as the aperture blocks. To avoid excessive vignetting, we only use the central 56×40 pixel part of the LCD array as the mask; the remaining part outside this area is covered with black carton to block light. Note that since the LCD array is binary, uniform measurement matrices cannot be realized with this mask, which require LCDs that can produce gray-scale values.

Both the LCD array and the digital camera are controlled by a computer. A specifically designed computer program successively changes the LCD image and makes an acquisition using the camera. The delay between the mask display and exposure is negligible, hence the total acquisition time approximately consists of the exposure times of each image.

Since the LCD array is not designed for this purpose, there are multiple sources of imperfections in the acquisitions⁴. For instance, the diffraction due to pixel boundaries in the LCD causes some artifacts in the acquired images. More importantly, a black pixel in the LCD array does not completely block light passing through it, which changes the effective measurement matrix. Similarly, a white pixel does not completely pass the light. Also, the pixels in the LCD array have different responses, which cause inhomogeneity within the images. To compensate for these effects, we have acquired images of white backgrounds and color calibration boards, and used these acquisitions to approximately calculate the pixel responses. These pixel responses are then used to calculate the actual measurement mask. Although this calibration significantly reduced artifacts, this prototype system can be considerably improved with specially designed hardware. For instance, [34] recently reported that Liquid Crystal on Silicon devices are more suitable for aperture coding than LCDs. Our incoherent acquisition and reconstruction framework can be directly applied to this system as well.

VII. EXPERIMENTAL RESULTS

A. Synthetic Experiments

For synthetic experiments, a 4D light field image is constructed using the Blender software [47]. We constructed a toy 3D scene with three objects at different depths, and the camera is moved vertically

⁴Notice also that the LCD array is not placed right at the aperture plane, but in front of the lens, as the former requires extensive modification of the main lens. This also introduces some artifacts in the angular images.

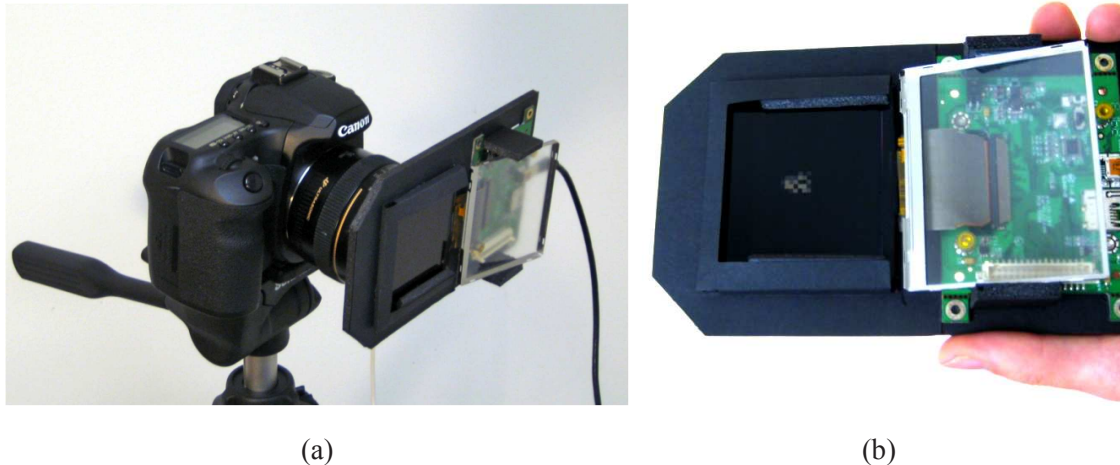


Fig. 2. (a) Our prototype camera where an LCD array is mounted to the lens of a digital camera, (b) the LCD array showing an example mask combination, and (c) zoomed area marked in (b).

and horizontally to acquire angular images that compose the 4D light field image with both horizontal and vertical parallax. One angular image from this set is shown Fig. 4(a). The light field image has a spatial resolution of 200×150 and an angular resolution of 5×7 . The warping fields between the angular images are assumed to be known to test the best-case reconstruction performance. Our current (unoptimized) MATLAB implementation takes about 10 minutes on a 3GHz Core2 Duo CPU to obtain the final reconstructions.

We experimented with two different measurement matrices \mathbf{A} : 1) The uniform spherical ensemble, where the entries of \mathbf{A} are drawn from a uniform distribution and are between 0 and 1, and 2) scrambled Hadamard matrices, where a random subset of rows of a S-matrix [48] is chosen to generate \mathbf{A} . In both cases, the expected mean of the entries in one row of \mathbf{A} is 0.5, as the mean of the distribution is 0.5 in the first case and due to the property of the Hadamard matrices in the second case. Therefore, the expected amount of light passing through the aperture in each acquisition is half of the maximum possible with both measurement matrices. Finally, we add zero-mean Gaussian noise to the measurements to obtain the final observations. We tested the reconstruction performance at two different noise levels with corresponding variances 0.001 and 1, where the intensity interval of the images is $[0, 255]$.

We vary the number of acquired images M from 3 to 35 and apply the proposed reconstruction algorithm using the incoherent observations to obtain estimates of the original light field image. The relative reconstruction error is calculated according to $\| \hat{\mathbf{x}} - \mathbf{x} \|_2^2 / \| \mathbf{x} \|_2^2$, where \mathbf{x} and $\hat{\mathbf{x}}$ are the

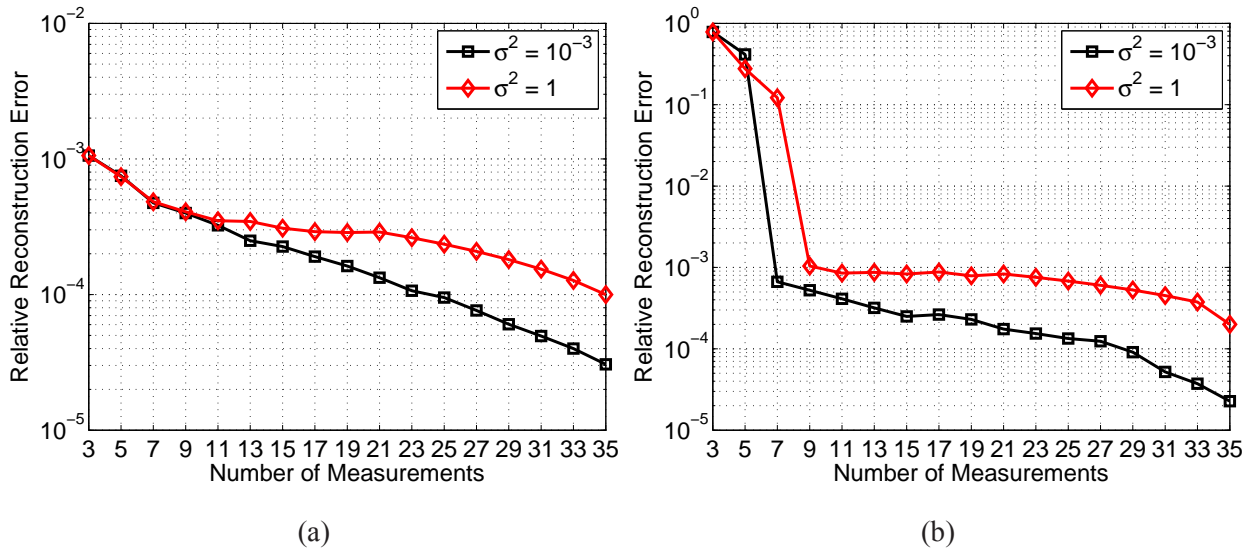


Fig. 3. Number of measurements vs relative reconstruction error (averaged over 20 runs) at two different noise levels (a) with uniform and (b) with scrambled Hadamard measurement matrices.

original and estimated images, respectively.

Average reconstruction errors over 20 runs are shown in Fig. 3. Multiple remarks can be made from this figure: First, using uniform ensembles as measurement matrices generally result in lower reconstruction errors than in the case of scrambled Hadamard matrices. This is an expected result, as the uniform measurement matrices collect information from all angular images at each acquisition whereas when Hadamard matrices are used, only some of the acquired images contain information about a particular angular image. Therefore, more acquisitions are generally required to achieve the same reconstruction error.

Second, note that when the number of acquisitions is very low, e.g., 3-7, in some cases the algorithm is unable to provide successful restorations with Hadamard measurements, whereas we can always obtain some estimate of the light field image with the uniform measurements. Note, however, that although uniform measurements are clearly superior to Hadamard measurements with a low number of measurements, they achieve almost the same reconstruction performance when the number of measurements is higher than 11. This is an important result as the practical application of Hadamard matrices is much easier than employing masks with uniform measurements.

Note that the difference in the reconstruction errors between low- and high-noise cases is not significant. It is clear that the reconstruction method is very successful in reconstructing the light field image when heavy noise is present. This is especially evident in the reconstruction error at full measurement ($M = 35$);

the error level is around the same order as when $M \geq 9$ in the uniform measurement case and $M \geq 11$ in the Hadamard measurement case, and the visual fidelity of the reconstructed light field remains nearly unchanged.

Overall, it is clear that very accurate reconstructions can be obtained using few measurements compared to the angular dimension of the light field image. In the low-noise case, average reconstruction errors of around 6×10^{-4} and 3×10^{-4} from 9 and 15 measurements, respectively, are obtained with uniform measurement matrices. With scrambled Hadamard matrices, same error levels are achieved with about 11 and 17 measurements.

For visual quality assessment, examples of reconstructed images using 9, 13 and 17 measurements are shown in Fig. 4 for uniform and in Fig. 5 for Hadamard matrices, respectively. Note that in both cases, the reconstructed images are very close to the original angular image; the image details and structure of the scene are accurately reconstructed. The visual quality of the reconstructions can also be observed from Fig. 6, where nine angular images from the light field reconstructed from 13 measurements with uniform matrices are shown.

Light field images have a number of applications in image based rendering, with typical ones being novel view synthesis and refocusing. To assess the visual quality of the reconstructed images in such a postprocessing application, we present digital refocusing results in Fig. 7. Notice that although only the reconstructed 35 images are used to obtain the refocused images, the results are of high visual quality without ghosting artifacts. Moreover, the refocused images using the reconstructions are nearly indistinguishable from the refocused images rendered using the original light field image.

In summary, it can be observed that using the proposed design the number of acquisitions can be significantly reduced (by a factor between 4 and 6). Furthermore, the reduction in the number of acquisitions is expected to be much higher with larger light field images, due to the increased level of sparsity.

B. Experiments with Real Images

Using the camera described in Section VI, we have acquired a real light field image of a representative scene. The acquired light field has angular dimensions 5×7 , and each acquired image is around 10 megapixels (3888×2592). To reduce the computational load for demonstration purposes, we cropped and downsampled them to 350×230 images. We have acquired a full set of measurements (total of 35 exposures) with Hadamard measurements to compare the compressive sensing reconstruction with linear reconstruction (such as the method in [17]). The warping fields are obtained by acquiring the single-block

images from the opposite ends of the mask, and using the procedure described in Section IV-B2.

Three of the 35 acquired images are shown in Fig. 8(a). Three angular images reconstructed using linear Hadamard inversion from the full set of 35 images are shown in Fig. 8(b). The amplified noise level is clearly visible, which is not surprising since no postprocessing (such as denoising) is applied to handle the noise during the acquisition and multiplexing phase. Figures 8(c)-(e) show corresponding reconstructed angular images using the proposed scheme with 10, 15 and 20 measurements, respectively. The central parts of the images are shown in Fig. 9 for a closer inspection. Although much fewer acquisitions are used, the quality of the reconstructed images is higher than using linear reconstruction with the full dataset. It is clear that the proposed method successfully controls the trade-off between noise amplification and smoothness of the solution, thus resulting in noise-free images with sharp edges, while correcting the vignetting artifacts and nonuniform lighting to some extent. Notice that no additional postprocessing is applied to the final images to demonstrate the effectiveness of the reconstruction algorithm; the remaining illumination differences between the angular images can be corrected by employing additional postprocessing algorithms.

VIII. CONCLUSIONS

In this paper, we proposed a novel application of compressive sensing to a new camera design to acquire 4D light field images. We have shown that incoherent measurements of angular images can be collected by using a randomly coded mask placed at the aperture of a traditional camera. These measurements are then used to reconstruct the original light field image. We developed a reconstruction algorithm which exploits the high degree of information redundancy (and hence, sparsity) inherent in the light field images, and have shown that the complete light field image can be reconstructed using only a few acquisitions. Moreover, the captured images have high signal-to-noise ratios due to small amount of loss of light. The proposed design provides high spatial and angular resolution light field images, and does not suffer from limitations of many existing light field imaging systems. Finally, the proposed design can be implemented by simple modifications of traditional cameras. Experimental results with both synthetic and real image sets show the effectiveness and potential of this approach for light field acquisition.

The proposed design, although powerful in terms of providing both high spatial- and angular-resolution, also has several limitations. Most importantly, it requires the scene and the camera to be static as a number of acquisitions have to be made. Any object or camera motion will necessarily introduce significant artifacts in the reconstructed images. In addition, our current implementation of the reconstruction method requires simultaneous processing of all angular images and the observations. Although we observed that

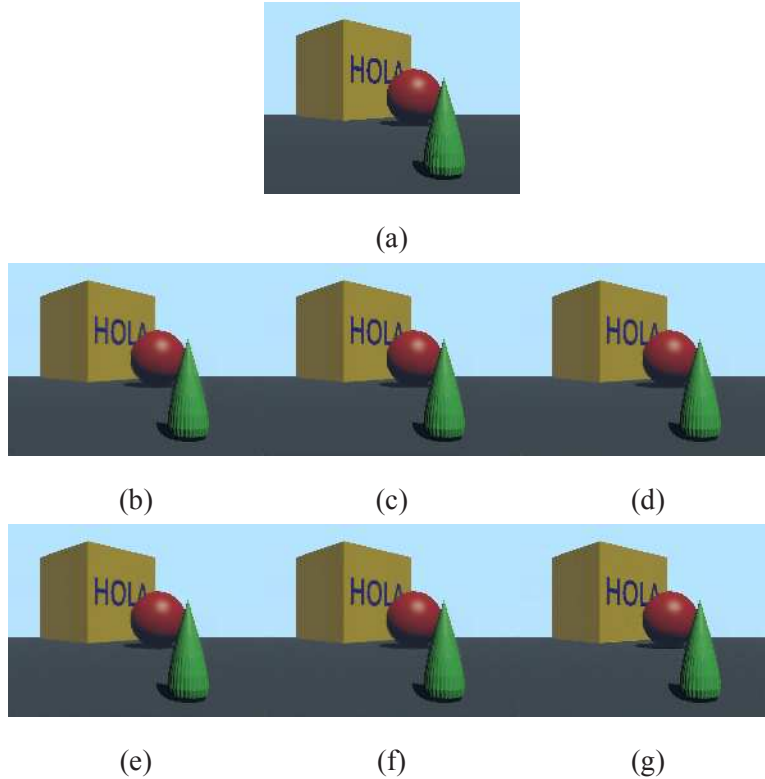


Fig. 4. Reconstruction examples with uniform matrices. (a) Original angular image, reconstructed images from (b,e) 9 measurements, (c,f) 13 measurements, and (d,g) 17 measurements. The middle row corresponds to the low noise case, and the bottom row corresponds to the high noise case.

the convergence is generally very fast, this processing might lead to high computational load if the size of the light field is large. Although not explored in this paper, this problem can potentially be addressed by processing images in patches and by parallel processing.

REFERENCES

- [1] S. D. Babacan, R. Ansorge, M. Luessi, R. Molina, and A. K. Katsaggelos, “Compressive sensing of light fields,” in *IEEE International Conference on Image Processing*, Cairo, Egypt, July 2009.
- [2] F. Durand and R. Szeliski, “Guest editors’ introduction: Computational photography,” *IEEE Computer Graphics and Applications*, vol. 27, no. 2, pp. 21–22, 2007.
- [3] E. Candes, J. Romberg, and T. Tao, “Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information,” *IEEE Trans. Inform. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.
- [4] D. L. Donoho, “Compressed sensing,” *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.
- [5] R. Baraniuk, “Compressive sensing,” *IEEE Signal Processing Magazine*, vol. 24, no. 4, pp. 118–121, July 2007.
- [6] E. Candès, “Compressive sampling,” in *Int. Congress of Mathematics 3*, Madrid, Spain, 2006, pp. 1433–1452.
- [7] F. Ives, “Parallax stereogram and process of making same,” *Patent US 725,567*, 1903.

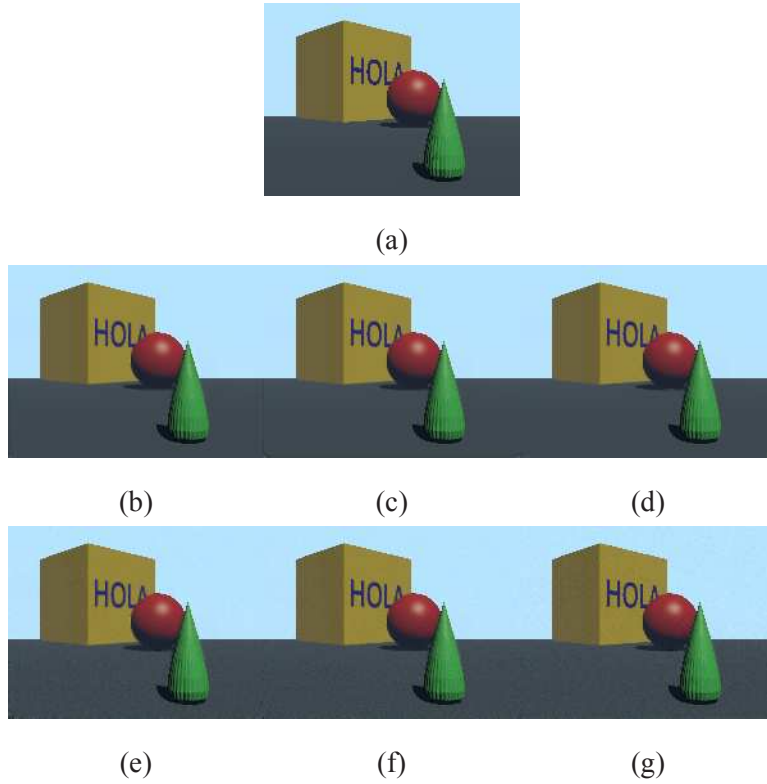


Fig. 5. Reconstruction examples with scrambled Hadamard matrices. (a) Original angular image, reconstructed images from (b,e) 9 measurements, (c,f) 13 measurements, and (d,g) 17 measurements. The middle row corresponds to the low noise case, and the bottom row corresponds to the high noise case.

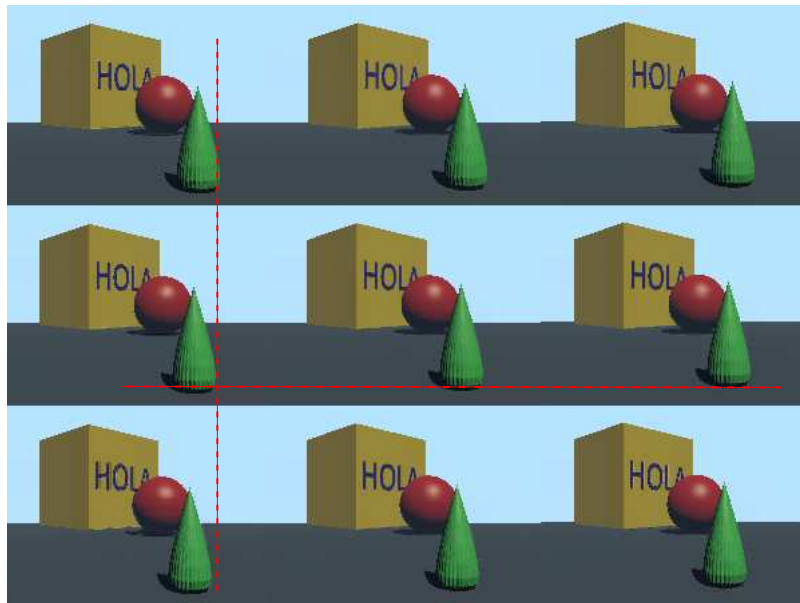


Fig. 6. Nine angular images from the light-field reconstructed with 13 uniform measurements and $\sigma^2 = 10^{-3}$. The displayed images are 80% of their original size.

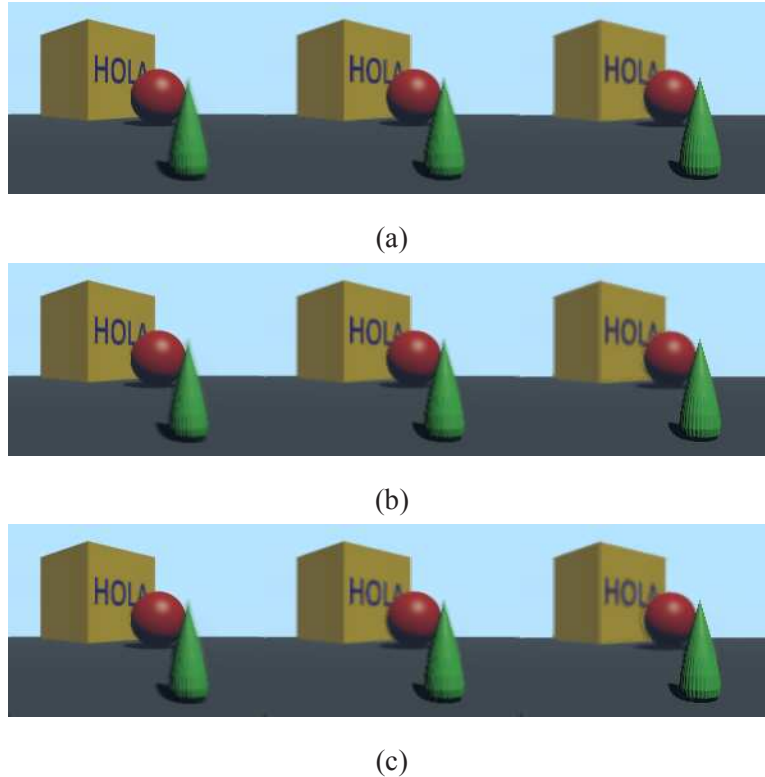


Fig. 7. Digital refocusing examples using (a) the original light field image, (b) the reconstructions using uniform matrices from 9 measurements, and (c) using the scrambled Hadamard matrices from 9 measurements.

- [8] G. Lippmann, “Epreuves reversible donnant la sensation du relief,” *J. Phys.* 7, pp. 821–825, 1908.
- [9] T. Adelson and J. Wang, “Single lens stereo with a plenoptic camera,” *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 14, no. 2, pp. 99–106, Feb 1992.
- [10] M. Levoy and P. Hanrahan, “Light field rendering,” *ACM Trans. Graph.*, pp. 31–42, 1996.
- [11] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen, “The lumigraph,” *ACM Trans. Graph.*, pp. 43–54, 1996.
- [12] “Light field photography with a hand-held plenoptic camera,” *Stanford Tech. Rep.*, 2005.
- [13] T. Georgiev, K. C. Zheng, B. Curless, D. Salesin, S. Nayar, and C. Intwala, “Spatio-angular resolution tradeoffs in integral photography,” in *EGSR*, june 2006, pp. 263–272.
- [14] B. Wilburn, N. Joshi, V. Vaish, E. Talvala, E. Antunez, A. Barth, A. Adams, M. Horowitz, and M. Levoy, “High performance imaging using large camera arrays,” *ACM Trans. Graph.*, vol. 24, no. 3, pp. 765–776, 2005.
- [15] A. Veeraraghavan, R. Raskar, A. Agrawal, A. Mohan, and J. Tumblin, “Dappled photography: Mask enhanced cameras for heterodyned light fields and coded aperture refocusing,” *ACM Trans. Graph.*, vol. 26, no. 3, pp. 69:1–69:12, July 2007.
- [16] T. Georgiev, C. Intwala, S. D. Babacan, and A. Lumsdaine, “Unified frequency domain analysis of lightfield cameras,” in *ECCV*, Marseille, France, December 2008.
- [17] C.-K. Liang, T.-H. Lin, B.-Y. Wong, C. Liu, and H. H. Chen, “Programmable aperture photography: multiplexed light field acquisition,” *ACM Trans. Graph.*, pp. 1–10, 2008.
- [18] A. Ashok and M. Neifeld, “Compressive light field imaging,” in *Proc. SPIE 7690*, 2010, p. 76900Q.

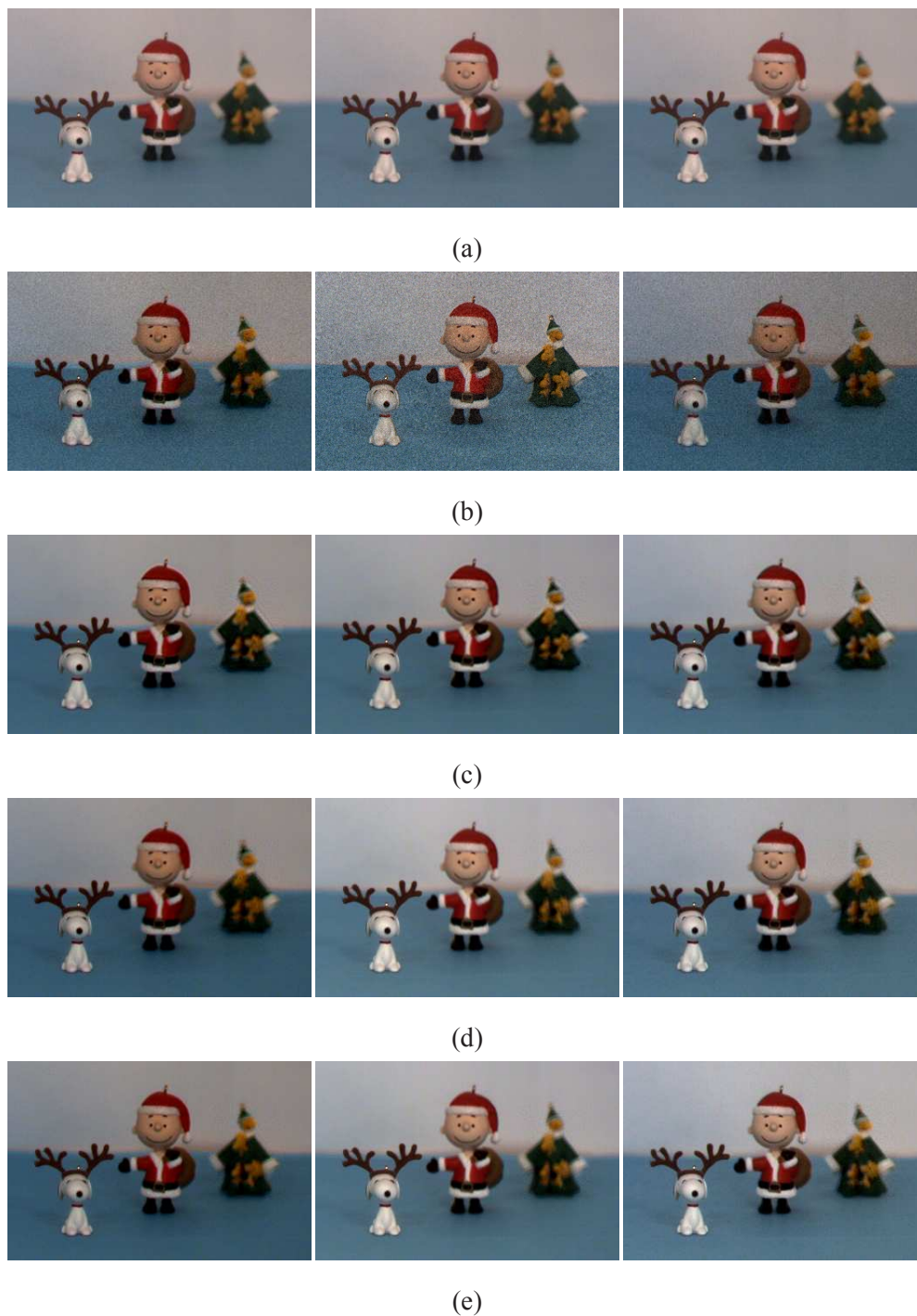


Fig. 8. Reconstruction results from a real dataset. (a) Three of the acquired images, reconstructed images (b) using linear Hadamard inversion from 35 images, and using the proposed scheme from (c) 10, (d) 15 and (e) 20 acquired images.



(a)



(b)



(c)



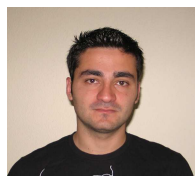
(d)

Fig. 9. Detailed parts from Figure 8. Reconstruction results (a) using linear Hadamard inversion from 35 images, and using the proposed scheme from (b) 10, (c) 15 and (d) 20 acquired images.

- [19] E. E. Fenimore and T. M. Cannon, “Coded aperture imaging with uniformly redundant arrays,” *Appl. Opt.*, vol. 17, no. 3, pp. 337–347, 1978.
- [20] S. R. Gottesman and E. E. Fenimore, “New family of binary arrays for coded aperture imaging,” *Appl. Opt.*, vol. 28, no. 20, pp. 4344–4352, 1989.
- [21] R. Raskar, A. Agrawal, and J. Tumblin, “Coded exposure photography: motion deblurring using fluttered shutter,” in *SIGGRAPH '06: ACM SIGGRAPH 2006 Papers*. New York, NY, USA: ACM, 2006, pp. 795–804.
- [22] A. Levin, R. Fergus, F. Durand, and W. T. Freeman, “Image and depth from a conventional camera with a coded aperture,” in *ACM Transactions on Graphics, SIGGRAPH 2007 Conference Proceedings*. New York, NY, USA: ACM, 2007, p. 70.
- [23] A. Zomet and S. K. Nayar, “Lensless imaging with a controllable aperture,” in *CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Washington, DC, USA: IEEE Computer Society, 2006, pp. 339–346.

- [24] S. K. Nayar and V. Branzoi, "Adaptive dynamic range imaging: Optical control of pixel exposures over space and time," in *ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision*. Washington, DC, USA: IEEE Computer Society, 2003, p. 1168.
- [25] A. Mohan, X. Huang, J. Tumblin, and R. Raskar, "Sensing increased image resolution using aperture masks," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008)*, pp. 1–8, June 2008.
- [26] H. Farid and E. P. Simoncelli, "Range estimation by optical differentiation," *JOSA A*, vol. 15, pp. 1777–1786, 1998.
- [27] M. E. Gehm, R. John, D. J. Brady, R. M. Willett, and T. J. Schulz, "Single-shot compressive spectral imaging with a dual-disperser architecture," *Opt. Express*, vol. 15, no. 21, pp. 14 013–14 027, 2007.
- [28] P. Sen and S. Darabi, "Compressive Dual Photography," *Computer Graphics Forum*, vol. 28, no. 2, pp. 609 – 618, 2009.
- [29] J. Gu, S. K. Nayar, E. Grinspun, P. N. Belhumeur, and R. Ramamoorthi, "Compressive Structured Light for Recovering Inhomogeneous Participating Media," in *European Conference on Computer Vision (ECCV)*, Oct 2008.
- [30] R. Marcia and R. Willett, "Compressive coded aperture superresolution image reconstruction," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2008)*, pp. 833–836, 31 2008–April 4 2008.
- [31] R. F. Marcia, Z. T. Harmany, and R. M. Willett, "Compressive coded aperture imaging," in *Computational Imaging VII*, C. A. Bouman, E. L. Miller, and I. Pollak, Eds., vol. 7246, no. 1. SPIE, 2009, p. 72460G.
- [32] W. T. Cathey and E. R. Dowski, "New paradigm for imaging systems," *Appl. Opt.*, vol. 41, no. 29, pp. 6080–6092, 2002.
- [33] Z. Zhang and M. Levoy, "Wigner distributions and how they relate to the light field," in *Proc. ICCP*, 2009.
- [34] H. Nagahara, C. Zhou, T. Watanabe, H. Ishiguro, and S. Nayar, "Programmable aperture camera using LCoS," in *European Conference on Computer Vision (ECCV)*, 2010.
- [35] C.-K. Liang, Y.-C. Shih, and H. Chen, "Light field analysis for modeling image formation," *IEEE Trans. Image Processing*, vol. 20, no. 2, pp. 446–460, Feb 2011.
- [36] E. J. Candes and T. Tao, "Decoding by linear programming," *IEEE Trans. Inform. Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.
- [37] D. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition," *IEEE Transactions on Information Theory*, vol. 47, no. 7, pp. 2845 –2862, Nov. 2001.
- [38] Z. Ben-Haim, Y. C. Eldar, and M. Elad, "Coherence-based performance guarantees for estimating a sparse vector under random noise," *IEEE Transactions on Signal Processing*, vol. 58, no. 10, pp. 5030 –5043, 2010.
- [39] L. Gan, T. Do, and T. Tran, "Fast compressive imaging using scrambled block Hadamard ensemble," in *EUSIPCO 2008*, Lausanne, Switzerland, August 2008.
- [40] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D*, pp. 259–268, 1992.
- [41] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 23, no. 11, pp. 1222–1239, Nov 2001.
- [42] R. Marcia and R. Willet, "Compressive coded aperture video reconstruction," in *EUSIPCO 2008*, Lausanne, Switzerland, August 2008.
- [43] S. D. Babacan, R. Molina, and A. K. Katsaggelos, "Parameter estimation in TV image restoration using variational distribution approximation," *IEEE Trans. Image Processing*, vol. 17, no. 3, pp. 326–339, March 2008.
- [44] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*. New York, Springer Verlag, 1985, ch. 3 and 4.
- [45] T. Chan, S. Esedoglu, F. Park, and A. Yip, "Recent developments in total variation image restoration," in *Handbook of Mathematical Models in Computer Vision*, N. Paragios, Y. Chen, and O. Faugeras, Eds. Springer Verlag, 2005.

- [46] S. Ma, W. Yin, Y. Zhang, and A. Chakraborty, "An efficient algorithm for compressed MR imaging using total variation and wavelets," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008)*, June 2008, pp. 1–8.
- [47] "Blender software." [Online]. Available: <http://www.blender.org/>
- [48] M. Harwit and N. J. A. Sloane, *Hadamard transform optics*. Academic Press, 1979.



S. Derin Babacan (M'10) received the B.Sc. degree from the Electrical and Electronics Department at Bogazici University, Turkey in 2004 and the M.Sc. and Ph.D. degrees from the Department of Electrical Engineering and Computer Science at Northwestern University, in 2006 and 2009, respectively.

He is currently a Beckman Postdoctoral Fellow at the Beckman Institute for Advanced Science and Technology at the University of Illinois at Urbana-Champaign. His primary research interests are inverse problems in image processing, computer vision and computational photography. He is the recipient of an IEEE International Conference on Image Processing Paper Award (2007).



Reto Ansoerge was born 1979. He received the Dipl Ing FH degree in Elektrotechnik in 2005 from the HSR Hochschule fuer Technik Rapperswil, Switzerland and the M.Sc. degree from the Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL, in 2010. Between 2005 and 2007 he was been with the HSR Medialab, Rapperswil, Switzerland as a research engineer. Since 2009, he has been with Varian Medical Systems, Baden, Switzerland, where he is currently working on kV and MV image acquisition projects.



Martin Luessi (S'04-M'12) received the Ing. FH degree in electrical engineering from the Hochschule fuer Technik Rapperswil (HSR), Rapperswil, St. Gallen, Switzerland, in 2006, and the M.S and Ph.D. degrees in electrical engineering from the Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL, in 2007 and 2011, respectively.

In 2011, he joined the Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital (MGH), Boston, MA, as a Research Fellow. He also holds an appointment as a Research Fellow at Harvard Medical School. His current research focuses on signal processing methods for neuroimaging, in particular on methods for magnetoencephalography (MEG). Other research interests are Bayesian modeling and inference, numerical optimization, sparse signal processing, and machine learning.



Pablo Ruiz Matarán received the MS Degree in Mathematics in 2008 and Master in Multimedia Technologies in 2009, both from University of Granada. Currently, he is Ph.D. student of the Visual Information Processing group, at the Department of Computer Science and Artificial Intelligence of the University of Granada, and he is participating in the Spanish research programme Consolider Ingenio 2010: Multimodal Interaction in Pattern Recognition and Computer Vision (MIPRCV). His research interest include super-resolution and classification of multispectral satellite images, video retrieval from video databases and fusion of visible-infrared images.



Rafael Molina (M'88) was born in 1957. He received the degree in mathematics (statistics) in 1979 and the Ph.D. degree in optimal design in linear models in 1983. He became Professor of computer science and artificial intelligence at the University of Granada, Granada, Spain, in 2000. His areas of research interest are image restoration (applications to astronomy and medicine), parameter estimation in image restoration, super resolution of images and video, and blind deconvolution.



Aggelos K. Katsaggelos (F'98) received the Diploma degree in electrical and mechanical engineering from the Aristotelian University of Thessaloniki, Greece, in 1979, and the M.S. and Ph.D. degrees in Electrical Engineering from the Georgia Institute of Technology, in 1981 and 1985, respectively.

In 1985, he joined the Department of Electrical Engineering and Computer Science at Northwestern University, where he is currently a Professor holder of the AT&T chair. He was previously the holder of the Ameritech Chair of Information Technology (1997-2003). He is also the Director of the Motorola Center for Seamless Communications, a member of the Academic Staff, NorthShore University Health System, an affiliated faculty at the Department of Linguistics and he has an appointment with the Argonne National Laboratory.

He has published extensively in the areas of multimedia signal processing and communications (over 180 journal papers, 400 conference papers and 40 book chapters) and he is the holder of 19 international patents. He is the co-author of Rate-Distortion Based Video Compression (Kluwer, 1997), Super-Resolution for Images and Video (Claypool, 2007) and Joint Source-Channel Video Transmission (Claypool, 2007).

Among his many professional activities, Prof. Katsaggelos was Editor-in-Chief of the IEEE Signal Processing Magazine (1997-2002), a BOG Member of the IEEE Signal Processing Society (1999-2001), and a member of the Publication Board of the IEEE Proceedings (2003-2007). He is a Fellow of the IEEE (1998) and SPIE (2009) and the recipient of the IEEE Third Millennium Medal (2000), the IEEE Signal Processing Society Meritorious Service Award (2001), the IEEE Signal Processing Society Technical Achievement Award (2010), an IEEE Signal Processing Society Best Paper Award (2001), an IEEE ICME Paper Award (2006), an IEEE ICIP Paper Award (2007) and an ISPA Paper Award (2009). He was a Distinguished Lecturer of the IEEE Signal Processing Society (2007-2008).

6.1.2 Light field acquisition from blurred observations using a programmable coded aperture camera

- **P. Ruiz**, J. Mateos, C. Cárdenas, S. Nakajima, R. Molina, and A.K. Katsaggelos, “Light field acquisition from blurred observations using a programmable coded aperture camera” in *21th European Signal Processing Conference (EUSIPCO 2013)*, 1569743131, Marrakech (Morocco), September 2013.
 - Status: Published
 - Indexed in CORE Conference Ranking as CORE B
 - H index: 9 (Q3:755/1201)

LIGHT FIELD ACQUISITION FROM BLURRED OBSERVATIONS USING A PROGRAMMABLE CODED APERTURE CAMERA

P. Ruiz^a, J. Mateos^a, M. C. Cárdenas^b, S. Nakajima^c, R. Molina^a and A. K. Katsaggelos^d

(a) Dept. de Ciencias de la Computación e I.A., Universidad de Granada, Granada, Spain

(b) Instituto de Astrofísica de Andalucía, Granada, Spain*

(c) Optical Research Laboratory, Nikon Corporation, Japan

(d) Dept. of Electrical Engineering & Computer Science, Northwestern University, USA

ABSTRACT

In this paper we deal with the problem of acquiring a scene light field using a programmable coded aperture camera when the angular observations are out-of-focus. We describe a portable programmable coded aperture prototype that can be attached to any DSLR camera lens and propose a blind deconvolution method to deblur light fields. The performance of the proposed method is evaluated on synthetic and real images.

Index Terms— Computational photography, light field, blurred observations, programmable coded aperture camera.

1. INTRODUCTION

Moving from analog to digital has been a major advance in the world of photography. Besides the cost reduction, digital images can be edited and post-processed in countless ways by using a computer. In computational photography (CP), the postprocessing does most of the work, considering the image captured by the sensor as an intermediate data [1].

In the present work, we will use CP techniques to capture the light field of a scene. In recent years a number of light-field cameras have been developed. Plenoptic cameras, like Lytro [2] or Raytrix [3], introduce an array of microlenses in front of the sensor. This allows the sensor to record different angular views of the scene. Depending on the number of microlenses used, the resolution of the captured images can be greatly reduced. That is, there is a trade-off between angular resolution and spatial resolution of the light field; the more angular views are generated, the smaller the spatial resolution of each view.

To deal with this problem, systems using a coded aperture have been designed. In coded aperture acquisition systems, a

pattern mask is introduced to modify the lens aperture and to capture images that, once processed, allow to reconstruct the light field. Coded aperture began to be used for light field acquisition only a few years ago. In [4], the N angular views are obtained from N scrambled images captured with different masks and then solving a determined system of linear equations. The masks are loaded into a programmable LCD that is placed into the lens. Babacan et al. [5] reduce the number of observations required using Compressive Sensing theory in a system that uses an LCD to place the masks in front of the lens. The design by Nagahara [6] uses Liquid Crystal on Silicon (LCoS) to create the masks. This reduces the loss of light and improves the brightness and contrast but makes the lens bulkier than the LCD design.

None of the proposed models has dealt with the problem of defocused light fields. In spite of the small size of the individual blocks composing the coded aperture, the depth of field is limited and objects outside it will appear defocused in the reconstructed views. In this paper, we deal with the problem of blurred light field captured by the new coded aperture LCD based prototype, described in section 2, based on the design in [5], that can be mounted as a filter on any DSLR camera. To recover the light field from a set of blurred multiplexed observations, in section 3, we propose a new blind light field deconvolution method that adapts the model in [4] and the blind deconvolution method in [7] to our problem. The proposed method is evaluated on synthetic and real images and its performance is analyzed in section 4. Finally, section 5 concludes the paper.

2. PROTOTYPE DESCRIPTION

The coded aperture LCD based prototype we have constructed, see Fig. 1, can be mounted in front of the lens and has a small battery and controls so that it is portable and can be used autonomously. It uses an LCD array (Electronic Assembly DOGXL160S-7) consisting of 160×104 pixels of 0.418×0.397 mm with an active area of 70.0 mm \times 43.5 mm. In the prototype we have used a central part of 42

This research was supported by the Spanish Ministry of Economy and Competitiveness under project TIN2010-15137, the European Regional Development Fund (FEDER), and in part by the US Department of Energy grant DE-NA0000457.

Work in collaboration with the CP team members at IAA: J. Rodríguez Gómez, I. Bustamante Díaz, G. P. Candini, L. Costillo Iciarra, J. M. Jerónimo Zafra and M. R. Sanz Mesa.

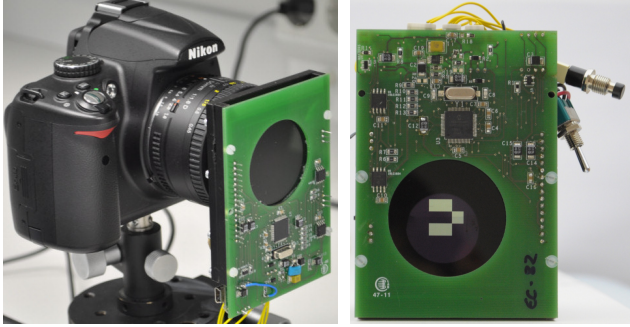


Fig. 1. (a) Mechanical interface, electronic control board and LCD, mounted in the prototype. It is equipped on a Nikon D5000 camera with a Nikkor 50mm f/1.8 lens. (b) The LCD showing one of the coded apertures.

mm in diameter and baffled the remaining area of the LCD to minimize the stray light. A high level software has been developed in the Labview environment that automatically detects the connection of the prototype to the computer USB port. It also allows to create masks, load them from disk, store them locally or in the prototype, set the LCD contrast, and display a given mask stored in the prototype EEPROM. Also, a low level interface has been programmed in Matlab so that the prototype and the camera can be directly controlled from a PC. This simplifies the capture of pictures in batch mode.

The LCD allows four different transmission levels for each pixels: transparent, opaque and two intermediate gray levels. The transmission of the LCD has been measured in the visible spectral range, from 400 nm to 800 nm, in the four states (see Fig. 2) and the contrast of the LCD has been set to 95% in order to maximize the transmission when the pixel is “transparent” and to provide a good separation between the 2 gray states. Unfortunately, the transmission in the “opaque” state is not negligible, and the images have to be properly corrected. Furthermore, the images captured by the prototype suffer from a set of aberrations. Firstly, the LCD spectral transmittance is not uniform and, also, it is not the same at all spatial locations. Secondly, the location of the prototype with respect to the lens creates a mechanical vignetting effect that heavily affects apertures with diameter smaller than half the LCD size.

To ameliorate these problems, we concentrate on a small 30×30 pixels central part of the LCD where the transmittance of the LCD can be considered as spatially invariant. Also, we take into account only the central part of the images where no vignetting is present. This allows us to simplify the pre-processing of the captured images that, in fact, reduces to camera calibration. We only need to perform white balance using a white surface and take two calibration pictures of this surface; one with the LCD set to opaque and another to transparent. These images will allow us to recover the original luminance

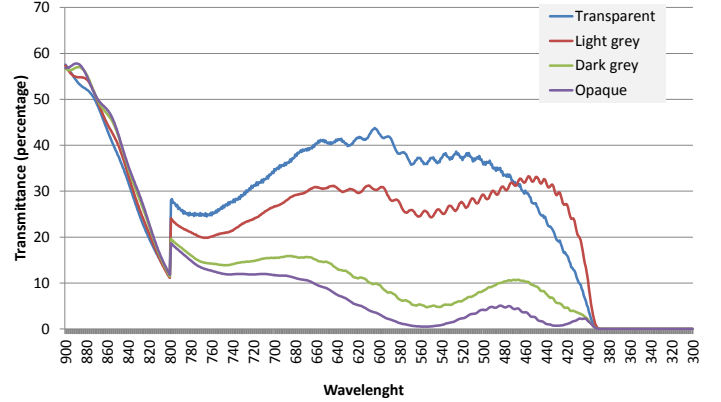


Fig. 2. LCD transmittance for the different wavelengths with a contrast of 95%.

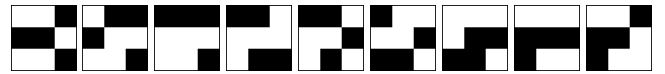


Fig. 3. Set of coded apertures used in the experiments. White means transparent and black corresponds to opaque.

of the scene despite the lower transmittance of the LCD.

3. IMAGE MODEL AND RECONSTRUCTION

By setting different blocks of the LCD to opaque or transparent we can capture different angular views of the same scene [4, 5, 6]. Opening only one block at a time allows to capture the light field sequentially by acquiring N angular views in N exposures. However, better results are obtained [4] using a multiplexed strategy where several blocks are set to transparent at the same time using a so called coded aperture.

To recover N views of the light field, we consider capturing M different pictures with coded apertures like the ones shown in Fig. 3. Then, each acquired image, \mathbf{y}_i , $i = 1, \dots, M$, is modeled as a linear combination of the different N , possibly blurred, angular views, as

$$\mathbf{y}_i = \sum_{j=1}^N a_{ij} \mathbf{H} \mathbf{x}_j + \mathbf{r}_i, \quad i = 1, \dots, M, \quad (1)$$

where \mathbf{x}_j is the j -th original (unknown) angular view of size $P = P_x \times P_y$ pixels, represented as a column vector. We assume that all the angular views share the same blur, \mathbf{H} , that is a $P \times P$ blurring matrix obtained from the unknown blur kernel \mathbf{h} of support $K = K_x \times K_y$, and \mathbf{r}_i is the capture noise. The a_{ij} coefficients indicate the contribution of the light field angular view j to picture i . Notice that, if the LCD behaved ideally, those coefficients would be 0 if the corresponding block in the LCD is set to opaque, or 1 if it is set to

transparent. Unfortunately this is not the case on real LCDs but the values for the a_{ij} coefficients, with values between 0 and 1, can be estimated from the calibration pictures.

Our goal is to estimate the light field angular views \mathbf{x}_i , $i = 1, \dots, N$, and the blurring kernel, \mathbf{h} , from the set of $M = N$ multiplexed observed images \mathbf{y}_i , $i = 1, \dots, N$.

We first recover each pixel k of the different blurred angular views, represented by \mathbf{z}_j , $j = 1, \dots, N$, from the acquired images. Since the observations \mathbf{y}_j , $j = 1, \dots, N$, are noisy we utilize \mathbf{y}'_j , the denoised version of \mathbf{y}_j obtained by applying the BM3D [8] denoising method to the observed images before recovering the different blurred angular views. Then we obtain \mathbf{z}_j by solving the determined linear systems

$$\mathbf{y}'(k) = \mathbf{A}\mathbf{z}(k), \quad k = 1, \dots, P, \quad (2)$$

where the matrix \mathbf{A} is the $N \times N$ system matrix formed from the coefficients a_{ij} , where each row of the matrix contains the coefficients of a coded aperture and $\mathbf{z}(k)$ and $\mathbf{y}'(k)$ are column vectors formed by stacking the pixels at position k of the set of images $\{\mathbf{z}_1, \dots, \mathbf{z}_N\}$ and $\{\mathbf{y}'_1, \dots, \mathbf{y}'_N\}$, respectively.

According to our model each blurred angular view, \mathbf{z}_j , $j = 1, \dots, N$, can be mathematically expressed by

$$\mathbf{z}_j = \mathbf{H}\mathbf{x}_j + \mathbf{n}_j, \quad (3)$$

where the vector \mathbf{n}_j represents the noise, assumed to be Gaussian of variance β^{-1} . Its precision parameter, β , is the same for all the images because, as they are taken under identical conditions, they will have the same noise properties. Notice that \mathbf{n}_j was introduced since \mathbf{z}_j of Eq. (2) will very likely be noisy.

We apply the variational Bayesian approach in a blind deconvolution procedure [7] to recover the blurring kernel \mathbf{h} and the restored angular views \mathbf{x}_j . From Eq. (3), we write the degradation model as

$$p(\mathbf{z}|\mathbf{x}, \mathbf{h}, \beta) = \prod_{j=1}^N p(\mathbf{z}_j|\mathbf{x}_j, \mathbf{h}, \beta) \propto \beta^{P/2} \exp\left(-\frac{\beta}{2} \sum_{j=1}^N \|\mathbf{z}_j - \mathbf{H}\mathbf{x}_j\|^2\right), \quad (4)$$

where \mathbf{z} and \mathbf{x} are column vectors formed by stacking vertically the vectors \mathbf{z}_j and \mathbf{x}_j , $j = 1, \dots, N$, respectively.

We use the general TV function as image prior for each view and, hence, we define

$$p(\mathbf{x}) = \prod_{j=1}^N p(\mathbf{x}_j|\alpha) \propto \exp\left(-\alpha \sum_{j=1}^N \text{TV}(\mathbf{x}_j)\right), \quad (5)$$

where

$$\text{TV}(\mathbf{x}_j) = \sum_{k=1}^P \sqrt{(\Delta^h(\mathbf{x}_j)(k))^2 + (\Delta^v(\mathbf{x}_j)(k))^2}, \quad (6)$$

with the operators $\Delta^h(\mathbf{x}_j)(k)$ and $\Delta^v(\mathbf{x}_j)(k)$ corresponding to the horizontal and vertical first order differences at pixel k , respectively.

To estimate all unknowns $\Theta = \{\mathbf{x}_1, \dots, \mathbf{x}_N, \mathbf{h}\}$ the variational Bayesian approach is used. In this approach, the posterior $p(\Theta|\mathbf{z})$ is approximated by another distribution, $q(\Theta)$, by minimizing the Kulback-Leibler (KL) divergence between both distributions [9]. A convenient factorization of $q(\Theta) = q(\mathbf{x}_1) \dots q(\mathbf{x}_N)q(\mathbf{h})$, named mean field approximation [9], is used in order to get a tractable minimization problem.

3.1. Angular view estimation

Due to use of TV prior, for estimating the distribution of each angular view, $q(\mathbf{x}_j)$, it is necessary to carry out a majorization-minimization procedure, as described in [7]. Thus, $q(\mathbf{x}_j)$ is estimated as a Gaussian distribution with mean $\bar{\mathbf{x}}_j$ and covariance matrix $\Sigma_{\mathbf{x}_j}$ given by

$$\bar{\mathbf{x}}_j = \Sigma_{\mathbf{x}_j} \beta \bar{\mathbf{H}}^T \mathbf{z}_j \quad (7)$$

$$\Sigma_{\mathbf{x}_j} = (\beta \bar{\mathbf{H}}^T \bar{\mathbf{H}} + \alpha((\Delta^h)^T \mathbf{W}_j \Delta^h + (\Delta^v)^T \mathbf{W}_j \Delta^v))^{-1} \quad (8)$$

where $\bar{\mathbf{H}}$ is the convolution matrix obtained from the current estimation of \mathbf{h} , $\bar{\mathbf{h}}$, and $\mathbf{W}_j = \text{diag}((u_j(k))^{-1/2})$, $k = 1, \dots, P$, with $u_j(k)$ a set of additional parameters introduced in the majorization procedure and calculated [7] as

$$u_j(k) = (\Delta^h(\mathbf{x}_j)(k))^2 + (\Delta^v(\mathbf{x}_j)(k))^2. \quad (9)$$

Notice that $\Sigma_{\mathbf{x}_j}$ in Eq. (8) is a $P \times P$ matrix and therefore its computation is extremely expensive. To alleviate this problem, each restored view, $\bar{\mathbf{x}}_j$, is estimated by solving, using conjugate gradient, the linear equation system

$$(\beta \bar{\mathbf{H}}^T \bar{\mathbf{H}} + \alpha((\Delta^h)^T \mathbf{W}_j \Delta^h + (\Delta^v)^T \mathbf{W}_j \Delta^v)) \mathbf{x}_j = \beta \bar{\mathbf{H}}^T \mathbf{z}_j. \quad (10)$$

3.2. Blur estimation

Note that Eq. (3) can also be written as $\mathbf{z}_j = \mathbf{X}_j \mathbf{h} + \mathbf{n}_j$ by forming the matrix \mathbf{X}_j similarly to \mathbf{H} . To estimate the blur, we follow the approximation proposed in [10] where \mathbf{h} is assumed to have a degenerate distribution $q(\mathbf{h})$ and the value where the distribution is degenerate is calculated as the PSF solution of

$$\hat{\mathbf{h}} = \arg \min_{\mathbf{h}} \mathbb{E}[\beta \sum_{i=1}^N \|\mathbf{z}_i - \mathbf{H}\mathbf{x}_i\|^2]. \quad (11)$$

Let us approximate \mathbf{W}_j in Eq. (8), following [7], by $\mathbf{W}_j \approx \text{mean}(\text{diag}(\mathbf{W}_j) \mathbf{I}_{P \times P})$, and then, following [10], utilize

$$\Sigma_{\mathbf{x}_j} \approx s_{\mathbf{x}_j} \mathbf{I}_{P \times P}, \quad (12)$$

with $s_{x_j} = (\beta \sum_{k=1}^K \mathbf{h}(k)^2 + 4\alpha \text{mean}(\text{diag}(\mathbf{W}_j)))^{-1}$. Let

$$\mathbf{C}_h^{-1} = \sum_{j=1}^N (\bar{\mathbf{X}}_j^T \bar{\mathbf{X}}_j + P s_{x_j} \mathbf{I}_{K \times K}), \quad (13)$$

with $\bar{\mathbf{X}}_j$ the convolution matrix obtained from the current estimation of \mathbf{x}_j , $\bar{\mathbf{x}}_j$. Then, $\hat{\mathbf{h}}$ can be approximated as the solution of the restricted quadratic program

$$\begin{aligned} \hat{\mathbf{h}} &= \arg \min_{\mathbf{h}} \mathbf{h}^T \mathbf{b}_h + \frac{1}{2} \mathbf{h}^T \mathbf{C}_h^{-1} \mathbf{h}, \\ \text{subject to } \sum_{k=1}^K \mathbf{h}(k) &= 1, \\ \mathbf{h}(k) &\geq 0, \quad k = 1, \dots, K. \end{aligned} \quad (14)$$

with

$$\mathbf{b}_h = - \sum_{j=1}^N \bar{\mathbf{X}}_j^T \bar{\mathbf{z}}_j. \quad (15)$$

In summary, to recover the light field from the degraded observations we proceeded as follows. First, we denoise the observed images by applying the BM3D [8] denoising method and then recover the different blurred angular views from Eq. (2). Secondly, we estimate the blur from the luminance band of the blurred views by alternatively iterating between Eqs. (10) and (14). The rationale behind this process is that the blur contaminating the R, G, and B bands is the same since these bands were captured under the same conditions and so we can speed up the estimation process by using only the luminance band. Finally, once the blur is obtained, we estimate each one of the RGB bands of the restored angular views by applying the non-blind restoration procedure described by Eq. (10) with the already estimated blur.

4. EXPERIMENTAL RESULTS

We have evaluated the performance of the proposed method with synthetic and real images. In the synthetic experiment, a scene was created with Blender¹ and a set of 9 different angular views were taken by placing a pinhole camera at 9 coplanar positions in the space. Those positions formed a 3×3 grid in a plane perpendicular to the Z scene axis. The angular view at position 5 (center) of the grid is displayed in Fig. 4a.

We then generated the set of 9 coded apertures depicted in Fig. 3 by selecting random 3×3 binary masks that have 5 open blocks. Each single block was in total open the same number of times in the 9 coded apertures set. The observed set of images was obtained by simulating the capture process in Eq. (1), that is, first blurring each view with a Gaussian blur with variance 1 and then multiplexing the blurred views using the set of coded apertures shown in Fig. 3. Finally, Gaussian

¹Available at <http://www.blender.org/>

Table 1. Mean PSNR and SSIM for the R,G,B bands and the mean of the RGB images for the synthetic experiment.

	R	G	B	mean (RGB)
PSNR	37.10	35.15	34.35	35.53
SSIM	0.9888	0.9868	0.9873	0.9876

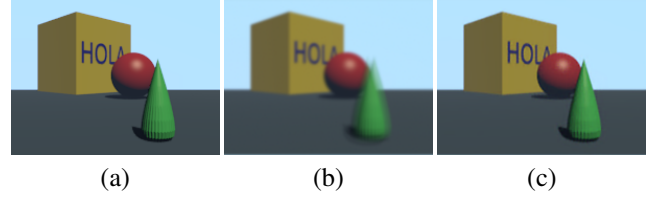


Fig. 4. Synthetic experiment with Gaussian blur ($\sigma = 1$): (a) Original angular view 5, (b) Simulated captured image with mask 5 in Fig. 3, (c) reconstructed angular view 5.

noise with standard deviation 0.001 was added to obtain the observed images, whose observation number 5 is depicted in Fig. 4b. Note that the letters in “Hola” that are at the focal plane are only blurred (since they will be at the same position in all the views) while the cone, that is far from the focal plane, represents the mixture of the different blurred views.

To estimate the original angular views from the observed images we apply the reconstruction algorithm described in the previous section. The initial blur \mathbf{h}^0 is set to a Gaussian with variance 0.16 and support $K_x = K_y = 21$, hence $K = 441$, that is a PSF close to a delta function. The precision parameter β in Eq. (3) is chosen such that the value of $P s_{x_k}$ in Eq. (13) is a fraction (0.1) of the maximum value of $\mathbf{X}^T \mathbf{X}$ in the first iteration of the algorithm. The rationale behind this is that the value of $P s_{x_k} \mathbf{I}_{K \times K}$, that represents the uncertainty of the minimum squares solution, tends to be smaller as we are more certain on the value of the image so, in the first iterations, we are forcing some uncertainty in the blur estimation process that will be reduced as the image is better restored. The image prior parameter α is chosen as a fraction of the value of β . We chose $\alpha = 0.001\beta$ to preserve most of the original data while smoothing out the restoration artifacts and the noise. The estimated angular view 5 is presented in Fig. 4d. Note that the blur has been successfully removed while preserving the structure in the cone. Numerical results, shown in Table 1, show that the reconstructed images have a very high quality both in terms of PSNR and SSIM measures.

We also tested the proposed method on real images. The set of images was taken with the prototype using the set of coded apertures depicted in Fig. 3. To minimize the effects of the spatially variant degradations produced by the LCD, we concentrated on a square of 30×30 pixels in the center of the LCD which was divided in a 3×3 set of square apertures each of size 10×10 pixels. This means that the area of each single block is 16.6 mm^2 . Also, we used only the 512×512 pixel central part of the images to reduce the spatially variant

effects of the lens and prevent vignetting from appearing.

The scene, as seen in Fig. 5, was set at 800 mm from the camera, the distance from the pin to the background is 40 mm, and, when 5 blocks are open, the depth of field is 45.1 mm.

We took pictures focusing at 50 mm from the dice (see Figs. 5a and 5b). For each RGB band, the system matrix \mathbf{A} was obtained by setting its coefficients equal to the mean value of the calibration pictures with the LCD set to opaque or to transparent, depending on whether the corresponding block is opaque or transparent. This allows us to recover the blurred angular views without any additional preprocessing.

We applied BM3D to the observed images using a variance calculated from a flat region of the image. Then we recovered the different blurred angular views from Eq. (2) resulting in the images depicted in Figs. 5c and 5d. Finally, the deconvolution algorithm was applied to the blurred angular views following the procedure described for synthetic images, obtaining the restored views, two of which are shown in Figs. 5e and 5f. As it can be observed, the restored views are sharp, making clearly visible the lines in the background or the details in the thread on the screw, but a bit noisy. This is due to noise amplification in the demultiplexing stage.

5. CONCLUSIONS

We have presented a new programmable aperture camera prototype that allows to capture light fields. We have addressed the problem of recovering blurred light fields that may occur due to the limited depth of field of the cameras. We have developed a method for deconvolving those blurred light fields and tested it on both synthetic and real images.

6. REFERENCES

- [1] M. Levoy, “Experimental platforms for computational photography,” *IEEE Computer Graphics and Applications*, vol. 30, pp. 81–87, 2010.
- [2] Y-R Ng, C. Pitts, and T. Knight, “Light field data acquisition,” U.S. Patent Application 20120327222, 2012.
- [3] C. Perwass and L. Wietzke, “Light field camera technology,” <http://www.raytrix.de/index.php/Technology.html>, Mar. 2013.
- [4] C.-K. Liang, T.-H. Lin, B.-Y. Wong, C. Liu, and H. H. Chen, “Programmable aperture photography: multiplexed light field acquisition,” in *ACM SIGGRAPH*, 2008, pp. 55:1–55:10.
- [5] S. D. Babacan, R. Ansorge, M. Luessi, P. Ruiz, R. Molina, and A. K. Katsaggelos, “Compressive light field sensing,” *IEEE Trans. on Image Processing*, vol. 60, pp. 3964–3977, 2012.
- [6] H. Nagahara, C. Zhou, T. Watanabe, H. Ishiguro, and S. K. Nayar, “Programmable aperture camera using LCoS,” in *Proc. of the ECCV’10*, 2010, pp. 337–350.
- [7] S. D. Babacan, R. Molina, and A. K. Katsaggelos, “Variational Bayesian blind deconvolution using a total variation prior,” *IEEE Trans. on Image Processing*, vol. 18, pp. 12–26, 2009.
- [8] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, “Image denoising by sparse 3D transform-domain collaborative filtering,” *IEEE Trans. on Image Processing*, vol. 16, pp. 2080–2095, 2007.
- [9] C. M. Bishop, *Pattern Recognition and Machine Learning*, Information Science and Statistics. Springer, 2007.
- [10] S. D. Babacan, R. Molina, M. Do, and A. K. Katsaggelos, “Blind deconvolution with general sparse image priors,” in *Proc. of the ECCV’12*, 2012, pp. 341–355.

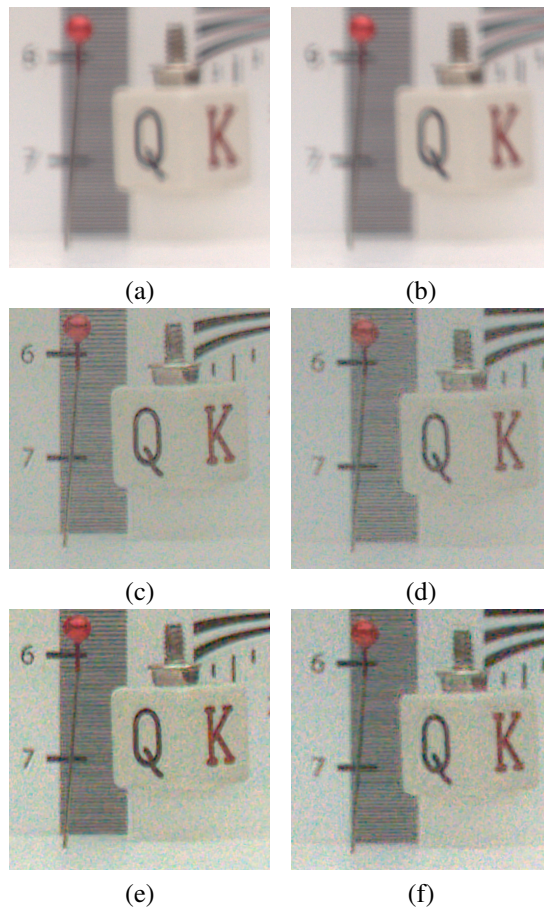


Fig. 5. Real experiment focused at 50 mm from the center of the scene. (a) Observed image 1, (b) Observed image 9, (c) Demultiplexed blurred angular view 1, (d) Demultiplexed blurred angular view 9, (e) deblurred angular view 1, (f) deblurred angular view 9.

6.2 Video Retrieval

6.2.1 Video Retrieval Using Sparse Bayesian Reconstruction

- **P. Ruiz**, S.D. Babacan, L. Gao, Z. Li, R. Molina, and A.K. Katsaggelos, “Video Retrieval Using Sparse Bayesian Reconstruction” in *IEEE International Conference on Multimedia and Expo (ICME2011)*, 1-6, Barcelona (Spain), July 2011.
 - Status: Published
 - Indexed in CORE Conference Ranking as CORE B
 - H index: 23 (Q1:200/1201)

VIDEO RETRIEVAL USING SPARSE BAYESIAN RECONSTRUCTION

Pablo Ruiz¹, S. Derin Babacan², Li Gao³, Zhu Li⁴, Rafael Molina¹, Aggelos K. Katsaggelos³

¹ Depto. de Ciencias de la Computación e I.A. Universidad de Granada.
{mataran, rms}@decsai.ugr.es

² Beckman Institute, University of Illinois at Urbana-Champaign.
dbabacan@illinois.edu

³ Dept. of Electrical Engineering and Comp. Sc. Northwestern University.
gaoli99@yahoo.es, aggk@eecs.northwestern.edu

⁴Dept. of Computing. Hong Kong Polytechnic University Kowloon.
zhu.li@ieee.org

ABSTRACT

Every day, a huge amount of video data is generated for different purposes and applications. Fast and accurate algorithms for efficient video search and retrieval are therefore essential. The interesting properties of sparse representation and the new sampling theory named Compressive Sensing (CS) constitute the core of the new approach to video representation and retrieval we are presenting in this paper. Once the representation (where sparsity is expected) has been chosen and the observations have been taken, the proposed approach utilizes Bayesian modeling and inference to tackle the retrieval problem. In order to speed up the inference process the use of Principal Components Analysis (PCA) to provide an alternative representation of the frames is analyzed. Experimental results validate the proposed approach whose robustness against noise is also examined.

Index Terms— Video retrieval, compressive sensing, Bayesian modeling, Bayesian inference

1. INTRODUCTION

A large amount of video data is generated every day. Searching through huge video databases is an important problem in many applications. For instance, individuals may want to search for video content they are interested in from YouTube videos, media companies may want to locate video content that violates their copyright protection (fingerprint) and, security systems may want to detect suspicious events among

surveillance videos. Fast and accurate algorithms in all these cases are needed for efficient video retrieval.

Due to the different types of query applications (such as query by example, query by video clip, query by semantics, etc), various image/video features are being employed by the different algorithms. For example, the color histogram of video frames is used in [1], both color and motion features are used in [2, 3, 4, 5, 6, 7], visual features and semantic labels are used in [8], and time interval statistics are used in [9]. A survey of this topic can be found in [10]. In [11], the authors compared the use of local and global features.

With the former robust results are obtained with high computational cost, while with the latter computational efficiency is gained at the expense of reduced performance.

Some algorithms also use indexing or hashing to improve search efficiency. For example, in [8] geometric hashing is used to build database indices, while in [4, 5, 7, 9] indexing tree structures are used. In [12], a kd-tree based space partitioning indexing scheme is applied to the video trajectory representations by using scaling and PCA. In [13] several random projections are used to project scaled videos on different search spaces, and then kd-trees on each space are used.

As described in [14] sparsity has emerged in the last decade as one of the important concepts in a wide range of signal processing applications [15, 16]. This interest has been even more elevated by the compressive sensing (CS) theory [17, 18, 19]. Compressive sensing is a new paradigm for signal acquisition where a signal is recovered from a low number of measurement without satisfying the Nyquist rate. CS is based on two main principles. First, the signal of interest can be represented with a sparse set of coefficients in a basis

This work was supported in part by the “Comisión Nacional de Ciencia y Tecnología” under contract TIN2010-15137 and the Spanish research programme Consolider Ingenio 2010: MIPRCV (CSD2007-00018).

(\mathbf{B}_S). The second important property is the incoherence between this representation basis and the measurement basis. It is shown by a large body of work that CS can be applied with great success to many application dramatically reducing the number of measurements needed for signal reconstruction.

Originally, most of the works in the CS and sparse representation literature focused on accurate representation and recovery of a signal in a given dictionary or basis. More recent works, however, exploit the discriminative properties of sparse recovery for classification (see, for instance, [20, 21]). The general principle behind sparse recovery for classification is that the test signals can be represented as linear combinations of the samples in the dictionary. Generally, this linear combination will include only a few coefficients, thus choosing the most relevant samples in the dictionary.

In this paper, we exploit the same discriminative nature of sparse representation for the video retrieval problem. Specifically, our goal is to find the sparsest representation of an input query video clip from the samples of a video database. We first construct the video database that is invariant to the starting video frame, and then formulate the video retrieval problem as sparse reconstruction. We employ a Bayesian compressive sensing algorithm to find the sparse representations of query videos within this database, and apply the classification procedure on the recovered sparse coefficients. Empirical results demonstrate the high retrieval performance of the proposed method compared to some existing algorithms.

The paper is organized as follows. In section 2 we explain how retrieving a video clip can be formulated as finding sparse representation in a convenient domain. In section 3 we formulate the video retrieval problem using the Bayesian framework, describe the inference procedure and explain the classification method for deciding whether a query video is in the database. In section 4 we discuss the feature extraction procedure. In section 5 we analyze the performance of the proposed system and determine its robustness in comparison with other systems.

2. SPARSE REPRESENTATION OF VIDEO CLIPS

In this section, we build a sparse representation for each video clip in the database in order to retrieve a clip of interest from a database using sparse representation principles.

The video database can be represented as a matrix by concatenating the existing video clips as

$$\mathbf{A} = [\mathbf{a}_{1,1}, \mathbf{a}_{1,2}, \dots, \mathbf{a}_{1,N_1}, \dots, \mathbf{a}_{K,1}, \dots, \mathbf{a}_{K,N_K}] \quad (1)$$

where $\mathbf{a}_{i,j}$, $i = 1, \dots, K, j = 1, \dots, N_i$ represents the j -th frame in the i -th video. Each $\mathbf{a}_{i,j}$ is assumed to be a column vector of size M where $M = VH$ with V and H the vertical and horizontal dimensions of each frame. For notational convenience, Eq. (1) is rewritten as

$$\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_N] \quad (2)$$

where $N = \sum_{i=1}^K N_i$. Let \mathbf{y} be a video-clip in the database written in the vector form as

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_S \end{bmatrix} \quad (3)$$

where each \mathbf{y}_i represents the i -th frame, and S is the length of the video clip. Next, we build the following matrix for a query of length S

$$\tilde{\mathbf{A}} = \begin{pmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \mathbf{a}_3 & \dots & \mathbf{a}_{N-(S-1)} \\ \mathbf{a}_2 & \mathbf{a}_3 & \mathbf{a}_4 & \dots & \mathbf{a}_{N-S} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{a}_S & \mathbf{a}_{S+1} & \mathbf{a}_{S+2} & \dots & \mathbf{a}_N \end{pmatrix} \quad (4)$$

by shifting the columns of \mathbf{A} in (1). Using this matrix, it can be observed that \mathbf{y} admits a sparse representation as

$$\mathbf{y} = \tilde{\mathbf{A}}\mathbf{x}_0 \quad (5)$$

where $\mathbf{x}_0 = (0, \dots, 0, 1, 0, \dots, 0)^t$ is a sparse vector with all coefficients equal to zero except for the entry corresponding to the location of the video clip \mathbf{y} in the database, which is equal to 1. Hence, the position of \mathbf{y} in the database is determined by \mathbf{x}_0 .

For a given clip \mathbf{y} , solving for the corresponding \mathbf{x}_0 is an ill-posed problem as the system in (5) is highly underdetermined which leads to non-uniqueness of the solutions. However, as our goal is to find the sparsest solution, i.e., finding the solution with most components equal to zero, this motivates following [21] to seek for the solution of

$$\hat{\mathbf{x}}_0 = \arg \min_{\mathbf{x}} \|\mathbf{x}\|_0 \quad \text{subject to} \quad \tilde{\mathbf{A}}\mathbf{x} = \mathbf{y} \quad (6)$$

where $\|\mathbf{x}\|_0$ is the l_0 -quasinorm (the number of non-zero coefficients). However, as is well known, the solution of this optimization problem is NP-hard. Furthermore, there are other issues like noise, different image sizes or even occlusions that make us resort to the CS formulation of the problem.

The noisy CS acquisition system can be modeled as

$$\mathbf{y} = \tilde{\mathbf{A}}\mathbf{x} + \mathbf{n}, \quad (7)$$

where \mathbf{n} is the $(SM) \times 1$ independent, Gaussian, zero-mean noise vector with variance equal to β^{-1} . The problem (6) can then be relaxed using the l_1 -norm formulation as

$$\hat{\mathbf{x}} = \arg \min_{\mathbf{x}} \{\|\mathbf{y} - \tilde{\mathbf{A}}\mathbf{x}\|_2^2 + \tau\|\mathbf{x}\|_1\}, \quad (8)$$

where $\|\cdot\|_1$ denotes the l_1 -norm. Solving (8) is much easier than (6) and has attracted much interest in the CS community.

3. VIDEO RETRIEVAL BASED ON BAYESIAN COMPRESSIVE SENSING

A number of methods have been proposed to solve the sparse optimization problem in Eq. (8), (see [14] and references therein, see also [22]). In this paper, we formulate the problem using the Bayesian framework following [23] which will also allow us to automatically estimate the regularization parameters, (see [23, 14] for references to parameter estimation). We provide here a brief review of solving (8) using a Bayesian approach.

In Bayesian modeling, all unknowns are treated as stochastic quantities with assigned probability distributions. The joint probability distribution of all quantities is given by

$$p(\mathbf{x}, \boldsymbol{\gamma}, \beta, \mathbf{y}) = p(\mathbf{y}|\mathbf{x}, \beta) p(\mathbf{x}|\boldsymbol{\gamma}) p(\boldsymbol{\gamma}) p(\beta). \quad (9)$$

The observation noise is independent and Gaussian with zero mean and variance equal to β^{-1} , that is, with (7),

$$p(\mathbf{y}|\mathbf{x}, \beta) = \mathcal{N}(\mathbf{y}|\tilde{\mathbf{A}}\mathbf{x}, \beta^{-1}). \quad (10)$$

It is shown in [23] that the l_1 regularization formulation in (8) is equivalent to using a hierarchical Laplace prior on the coefficients of \mathbf{x} , that is,

$$p(\mathbf{x}|\boldsymbol{\gamma}) = \prod_{i=1}^N \mathcal{N}(x_i|0, \gamma_i), \quad (11)$$

$$p(\gamma_i|\lambda) = \frac{\lambda}{2} \exp\left(-\frac{\lambda\gamma_i}{2}\right), \quad \gamma_i \geq 0, \lambda \geq 0, \quad (12)$$

where $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \dots, \gamma_N)$. Using this specification, the signal distribution $p(\mathbf{x}|\mathbf{y}, \lambda, \beta)$ is estimated as a multivariate Gaussian distribution $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \Sigma)$ with parameters

$$\Sigma = \left[\beta \tilde{\mathbf{A}}^t \tilde{\mathbf{A}} + \Lambda \right]^{-1}, \quad (13)$$

$$\boldsymbol{\mu} = \Sigma \beta \tilde{\mathbf{A}}^t \mathbf{y}, \quad (14)$$

with $\Lambda = \text{diag}(1/\gamma_i)$. The hyperparameters $\boldsymbol{\gamma}$ are then estimated by forming the likelihood function

$$\mathcal{L} = -\frac{1}{2} \log |\mathbf{C}| - \frac{1}{2} \mathbf{y}^t \mathbf{C}^{-1} \mathbf{y} + N \log \frac{\lambda}{2} - \frac{\lambda}{2} \sum_i \gamma_i, \quad (15)$$

with $\mathbf{C} = \left(\beta^{-1} \mathbf{I} + \tilde{\mathbf{A}} \Lambda^{-1} \tilde{\mathbf{A}}^t \right)$, and maximizing it with respect to each γ_i and λ in an alternating fashion. This procedure results in the updates

$$\gamma_i = -\frac{1}{2\lambda} + \sqrt{\frac{1}{4\lambda^2} + \frac{\langle x_i^2 \rangle}{\lambda}}, \quad (16)$$

$$\lambda = \frac{N-1}{\sum_i \gamma_i/2}, \quad (17)$$

where $\langle x_i^2 \rangle = x_i^2 + \Sigma_{ii}$. In summary, at each iteration of the algorithm, given an estimate of $\boldsymbol{\gamma}$ and λ , the estimate of the

distribution of \mathbf{x} is calculated using (13) and (14), followed by the estimation of the variances γ_i from (16) and the hyperparameter λ from (17). In addition, [23] proposed a greedy approach that finds the solutions much more efficiently without the need of solving the large linear system in (14). In our work, we use this greedy approach to find the solution of (8).

3.1. Classification Procedure

We finally proceed to decide whether the query video-clip is in the database. If the query video-clip is in database, then its sparse representation will only have one non-zero component, and equal to 1 in the position of the first frame of the query video. Let $\hat{\mathbf{x}}$ be the vector $\boldsymbol{\mu}$ at convergence of the Bayesian algorithm, and $m = \max_i \hat{x}_i$. We then define the vector \mathbf{x}^{comp} with components

$$x_i^{comp} = \begin{cases} 1 & \text{if } \hat{x}_i = m \\ 0 & \text{otherwise} \end{cases} \quad i = 1, 2, \dots, N. \quad (18)$$

We fix a threshold δ and decide that the query video is in the database if, and only if, \mathbf{x}^{comp} only has one non-zero component and

$$\|\hat{\mathbf{x}} - \mathbf{x}^{comp}\|_1 \leq \delta \quad (19)$$

4. FEATURE EXTRACTION

In order to perform an efficient search, and due to the size of frames, feature extraction is needed. We first assume that the frames in the database have been downsampled to a reasonable size (11×8 in our experiments). We then use a linear feature transformation. The projection from the image space to the feature space can be represented by a matrix $\mathbf{D} \in \mathbb{R}^{T \times M}$ with $T \ll M$ which when applied to \mathbf{A} produces

$$\mathbf{D}_{T \times M} \mathbf{A}_{M \times N} = \dot{\mathbf{A}}_{T \times N} \quad (20)$$

Then we can use the proposed retrieval procedure on $\dot{\mathbf{A}}$, which leads to a faster search. In this work we consider \mathbf{D} to be the matrix associated to PCA, see [12]. Notice that we could have also used a matrix of random projections $\Phi_{T \times M}$.

The PCA transformation retains much of the information in only a reduced set of principal components. The number of preserved dimensions, T , determines the energy loss during the PCA transformation. The energy represented by each PCA coefficient obtained from the test database used in the experiments, which consists of 567146 frames, is shown in Figure 1. Notice that for $T=4$ 70 % of the energy is preserved. Furthermore if the CS theoretical conditions are met by $\tilde{\mathbf{A}}$, see [21], then $\hat{\mathbf{x}}$ can be recovered by l_1 -minimization with overwhelming probability if $ST > 2 \log(567146/ST)$. In other words, around $ST \approx 10$ would suffice to recover the only non-zero component. As we will see in the experimental section, when $S = 3$ and $T = 4$ the proposed system retrieves all the relevant clips in the database in the noiseless case.

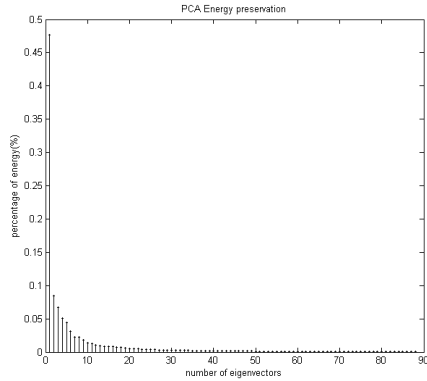


Fig. 1. Energy of the PCA components

No. of frames	Frames in database	Frames not in database
3	1.45 s	2.90 s
7	2.62 s	4.44 s
15	5.72 s	8.07 s

Table 1. CPU time used to find a query.

5. SIMULATION RESULTS

In our experiments we used the 2004 NIST TRECVID shot boundary test set. This data set has approximately 6 hour of video in 12 videos (each of about 30 mins long). We split it in two data sets. The positive video repository (or database) consists of 11 videos and the other video forms the negative data set.

The frames are downsampled with a scaling factor of 32 to produce 11x8 video icons. Then the frames are projected using PCA transformation with $T = 4$. In our test, we select randomly 250 positives and 250 negatives query videos. The query clip lengths are $S = 3, 7$ and 15 frames. All experiments were performed in an Intel Core 2 Duo 2GHz notebook with 2 GB of RAM. The mean times the Bayesian algorithm took to find the sparse representation of a video query is reported in Table 1.

5.1. Noise free test cases

For noise-free test cases our system retrieved all positive cases and rejected all negative one. The results are exactly the same as the ones reported in [12] and [13].

5.2. Noisy test cases

In real world applications video clips can be corrupted by coding and communication losses, as well as, image formation variations. To simulate coding losses in the query clips, we added Gaussian noise to the query clips at PSNR levels of 20, 25, 30, and 35 dB. Figure 2 shows one original image in the



Fig. 2. (a) Original frame in the database, (b) its noisy observation

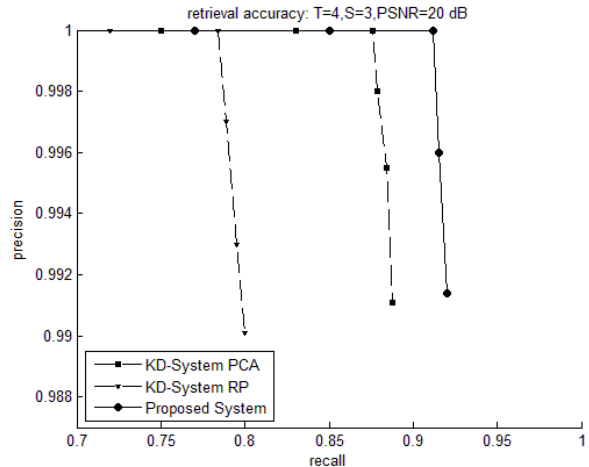


Fig. 3. Precision-Recall curves for three video-retrieval systems: KD-System PCA [12], KD-System RP [13], and Proposed Method.

database and its 20 dB noisy observation. Due to their sizes, the corresponding 11x8 icon frames are not shown.

In order to compare our system with certain state-of-art algorithms, we calculated the precision-recall curves, for the same set of query clips. The precision-recall curve [24] is a typical way of characterizing retrieval performance. For a given threshold, let us assume that a is the number of relevant (present in the database) clips retrieved, b the number of relevant clips not retrieved, and c the number of non relevant clips retrieved, then the precision and recall values are defined by $precision = a/(a + c)$ and $recall = a/(a + b)$, respectively. By changing the threshold value we obtain, for a given method, its precision-recall curve. Notice that as the threshold δ in Eq. (19) decreases the recall value is expected to decrease while the precision value is expected to increase.

In Fig. 3 three systems are compared for the case $T = 4$, $S = 3$, and $PSNR = 20$ dB. We can see that the method proposed in [13] performs worse than the one in [12]. Furthermore, the proposed sparse Bayesian retrieval method performs better than the method in [12].

In Fig. 4 the same systems compared for the case $T = 4$,

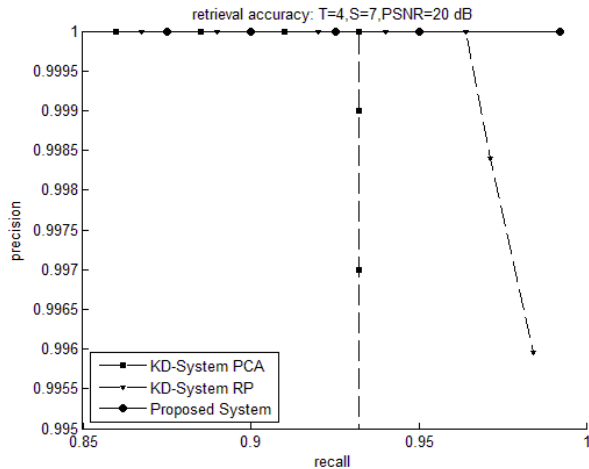


Fig. 4. Precision-Recall curves for three video-retrieval systems: KD-System PCA [12], KD-System RP [13], and Proposed Method.

$S = 7$, and $PSNR = 20$ dB. The method proposed in [12] performs worse than the one in [13]. Furthermore, the proposed sparse Bayesian retrieval method performs better than the method in [13]. As shown in Fig. 4 its precision-recall curve is $precision = 1$ for $recall \leq 0.98$

Finally in Fig. 5 the three systems are compared for the case $T = 4$, $S = 7$ and $PSNR = 25$ dB. Again the method proposed in [12] performs worse than the one in [13]. Furthermore, the proposed sparse Bayesian retrieval method performs better than the method in [13]. As shown in Fig. 5 its precision-recall curve is $precision = 1$ since the threshold values for all relevant clips are smaller than the corresponding to nonrelevant clip. The same behavior is observed for higher PSNR levels. When more frames are included in the query, that is when $S = 15$, the precision-recall curve for the proposed method is $precision = 1$.

6. CONCLUSION

In this paper we have developed a robust and efficient system for video retrieval, based on the use of sparse representation, compressive sensing and Bayesian modeling of the video retrieval problem. Experimental results demonstrate that the proposed method performs better than existing state-of-art systems and also its robustness against noise. Work to tackle the problems of occlusions and missing frames is already in progress.

7. REFERENCES

[1] A.M. Ferman, A.M. Tekalp and R. Mehrotra, "Robust Color Histogram Descriptors for Video Segment Re-

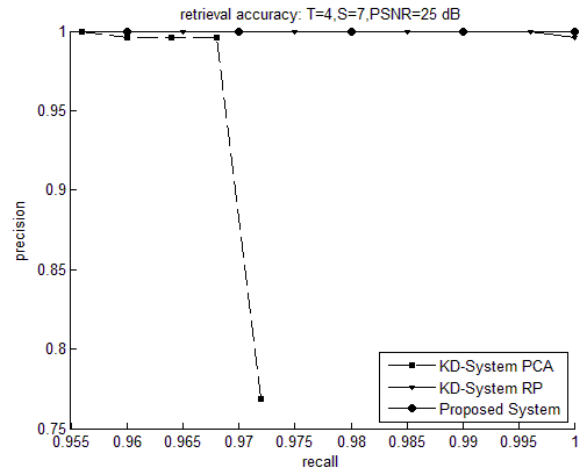


Fig. 5. Precision-Recall curves for three video-retrieval systems: KD-System PCA [12], KD-System RP [13], and Proposed Method.

trieval and Identification," *IEEE Trans. on Image Processing*, vol. 11, no. 5, pp. 497-508, May 2002.

[2] S.-F. Chang Chen, W. Meng, H.J. Sundaram and H. Di Zhong, "A Fully Automated Content-based Video Search Engine Supporting Spatiotemporal Queries," *IEEE Trans. on Circuits and System for Video Technology*, vol. 8, no.5, pp. 602-615, Sept. 1998.

[3] Y. Ho, C.-W. Lin, J.-F. Chen and H.-Y.M. Liao, "Fast Coarse-to-fine Video Retrieval Using Shot-level Spatiotemporal Statistics," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 16, no. 5, pp. 642-648, May 2006.

[4] C-T. Hsu and S.-J. Teng, "Motion Trajectory Based Video Indexing and Retrieval," *Proc. International Conference on Image Processing*, vol. 1, pp. I-605 - I-608, 2002.

[5] V. Mezaris, I. Kompatsiaris, N. V. Boulgouris and M. G. Strintzis, "Real-time Compressed-domain Spatiotemporal Segmentation and Ontologies for Video Indexing and Retrieval," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 14, no. 5, pp. 606-621, May 2004.

[6] C.-W. Ngo, T.-C. Pong and H.-J. Zhang, "On Clustering and Retrieval of Video Shots through Temporal Slices Analysis," *IEEE Trans. on Multimedia*, vol. 4, no. 4, pp. 446-458, Dec. 2002.

[7] J. Yuan, L.-Y. Duan, Qi Tian and C. Xu, "Fast and Robust Short Video Clip Search Using an Index Structure,"

- Proc. of 6th ACM SIGMM international workshop on Multimedia Info Retrieval (MIR)*, pp. 61-68, 2005.
- [8] J. Fan, A.K. Elmagarmid, X. Zhu, W.G. Aref and L. Wu, "ClassView: Hierarchical Video Shot Classification, Indexing, And Accessing," *IEEE Trans. on Multimedia*, vol. 6, no. 1, pp. 70-86, Feb. 2004.
- [9] C.G.M. Snoek and M. Worring, "Multimedia event-based video indexing using time intervals," *IEEE Trans. on Multimedia*, vol. 7, no. 4, pp. 638-647, Aug. 2005.
- [10] A. Joly, O. Buisson and C. Frelicot, "Content-based Copy Retrieval using Distortion-based Probabilistic Similarity Search," *IEEE Trans. on Multimedia*, vol. 9, no. 2, pp. 293-306, Feb. 2007.
- [11] J. Law-to, O. Buisson, L. Chen, M. H. Ipswich, V. Gouet-brunet, A. Joly, N. Boujemaa, I. Laptev, F. Stentiford and M. H. Ipswich, "Video copy detection: a comparative study," *ACM Int. Conf. on Image and Video Retrieval (CIVR)*, pp. 371-378, 2007.
- [12] L. Gao, Z. Li and A.K. Katsaggelos, "Fast Video Shot Retrieval with Luminance Field Trace Indexing and Geometry Matching," *Proc. IEEE Int. Conf. on Image Processing (ICIP)*, pp. 1497-1500, 8-11, Oct. 2006.
- [13] L. Gao, Z. Li and A. K. Katsaggelos, "A Video Retrieval Algorithm using Random Projections," *Proc. IEEE Int. Conf. on Image Processing (ICIP)*, pp. 797-800, 7-10 Nov. 2009.
- [14] J.L. Starck, F. Murtagh, and J. Fadili, *Sparse Image and Signal Processing: Wavelets, Curvelets, Morphological Diversity*, Cambridge University Press, Cambridge, 2010.
- [15] M. Elad, M.A.T. Figueiredo, and M. Yi, "On the Role of Sparse and Redundant Representations in Image Processing," *Proceedings of IEEE*, vol. 6, 972 - 982, June 2010.
- [16] J. Wright, M. Yi, J. Mairal, G. Sapiro, T.S. Huang, and Y. Yan, "Sparse Representation for Computer Vision and Pattern Recognition", vol. 6, 1031 - 1044, June 2010.
- [17] E. J. Candès and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?," *IEEE Trans. Inform. Theory*, vol. 52, 5406-5425, 2006.
- [18] D. Donoho, "Compressed sensing," *IEEE Trans. on Information Theory*, vol. 52, 1289 - 1306, 2006.
- [19] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. on Information Theory*, vol 52, 489-509, 2006.
- [20] J. Mairal, F. Bach, J. Ponce, G. Sapiro and A. Zisserman, "Discriminative Learned Dictionaries for Local Image Analysis," *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 1-8, 2008.
- [21] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210 - 227, Feb. 2009.
- [22] S.D. Babacan, R. Molina and A.K. Katsaggelos, "Parameter Estimation in TV Image Restoration Using Variational Distribution Approximation," *IEEE Transactions on Image Processing*, vol. 17, pp. 326 - 339, Feb. 2008.
- [23] S.D. Babacan, R. Molina and A.K. Katsaggelos, "Bayesian Compressive Sensing Using Laplace Priors," *IEEE Trans. on Image Processing*, vol. 19, no. 1, pp. 53-63, Jan. 2010.
- [24] V. Raghavan, P. Bollmann, and G. S. Jung, "A critical investigation of recall and precision as measures of retrieval system performance," *ACM Trans. on Information Systems*, vol. 7, 1989, pp. 205-229.

6.2.2 Retrieval of Video Clips with Missing Frames using Sparse Bayesian Reconstruction

- **P. Ruiz**, S.D. Babacan, R. Molina, and A.K. Katsaggelos, “Retrieval of Video Clips with Missing Frames using Sparse Bayesian Reconstruction” in *7th International Symposium on Image and Signal Processing and Analysis (ISPA 2011)*, 443-448, Dubrovnik (Croatia), September 2011.
 - Status: Published
 - H index: 12 (Q2: 515/1201)

Retrieval of Video Clips with Missing Frames using Sparse Bayesian Reconstruction

Pablo Ruiz*, S. Derin Babacan[†], Rafael Molina* and Aggelos K. Katsaggelos[‡]

*Depto. de Ciencias de la Computación e I.A., Universidad de Granada, 18071 Granada, Spain

Email: {mataran,rms}@decsai.ugr.es

[†]Beckman Institute for Advanced Sc. and Tech., University of Illinois at Urbana Champaign

405 N Mathews Ave, Urbana, IL 61801, USA. Email: dbabacan@illinois.edu

[‡]Dept. of Electrical Engineering and Comp. Sc., Northwestern University

2145 Sheridan Road, Evanston IL 60208, USA. Email: aggk@eecs.northwestern.edu

Abstract—Fast and accurate algorithms are essential for the efficient search and retrieval of the huge amount of video data that is generated for different purposes and applications every day. The interesting properties of sparse representation and the new sampling theory named Compressive Sensing (CS) constitute the core of the new approach to video representation and retrieval we are presenting in this paper to deal with the search of noisy video clips with also possibly missing frames. Once the representation (where sparsity is expected) has been chosen and the observations have been taken, the proposed approach utilizes Bayesian modeling and inference to tackle the retrieval problem. In order to speed up the inference process the use of Principal Components Analysis (PCA) to provide an alternative representation of the frames is analyzed. Experimental results validate the proposed approach to the retrieval of video clips with missing frames as well as its robustness against noise.

I. INTRODUCTION

With the rapidly increasing growth of digital video content, there is an equally growing need for efficient techniques to analyze, search and retrieve video content. Video retrieval is a key step in many applications including copyright protection, multimedia content search, security and surveillance. Fast and accurate algorithms in all these cases are needed for efficient video retrieval.

A number of methods have been developed for video retrieval. Generally, methods identify features distinguishing video frames and employ classification, indexing and searching based on these features. Among a large number of features, commonly used ones are color histograms [1], color and motion cues [2], [3], [4], visual features and semantic labels [5], and time interval statistics [6]. Surveys and comparisons can be found in [7], [8].

After identifying the distinguishing features, the second step in video retrieval is searching based on these features. Indexing and hashing are commonly used to improve the search efficiency. In [5] geometric hashing is used to build database indices, while [3], [4], [6] used tree-based indexing. A powerful data structure for indexing is the kd-trees. In [9], video trajectories over time are indexed using kd-trees with a dimensionality reduction using PCA. Random projections instead of PCA are utilized in [10], followed by several kd-trees for indexing.

In this paper, we present a new approach to video retrieval using *sparse representation and reconstruction*. The concept of sparsity has emerged in the last decade as a powerful modeling tool with a large number of potential applications [11], [12], [13]. This has been significantly motivated by the emergence of *compressive sensing* [14], [15]. Although compressive sensing and sparse reconstruction have originally aimed at the reconstruction of an original signal, the discriminative properties of sparse representations have been successfully utilized for several applications including face recognition and image classification [16], [17].

In this work, we demonstrate that the discrimination property of sparse representations can be employed very effectively for video retrieval. Specifically, we formulate the problem of searching a query video clip in a video database as a sparse reconstruction problem. We construct the video database directly from the existing video clips, such that no sophisticated feature extraction methods are needed as preprocessing. Moreover, we present a method to handle the problem of missing frames in the query clip, and show that our method is extremely robust to the cases where a large number (almost 80%) of the frames are removed. Finally, we demonstrate with experimental results that the proposed method provides very high retrieval performance in terms of both error rate and retrieval speed.

This paper is organized as follows. In section II, we present the proposed sparse representation framework and the Bayesian reconstruction algorithm for the video retrieval problem. The searching and classification procedure is explained in Section III. The dimensionality reduction step using PCA is presented in Section IV. We analyze the retrieval performance of the proposed system and its robustness to noise and missing frames in Section V and conclude in Section VI.

II. FRAME RETRIEVAL USING SPARSE REPRESENTATION

In this section, we study how to retrieve one frame in the database using sparse representation principles.

A. Sparse representation of frames

The video database can be represented as a matrix by concatenating the existing video clips as

$$\mathbf{A} = [\mathbf{a}_{1,1}, \mathbf{a}_{1,2}, \dots, \mathbf{a}_{1,N_1}, \dots, \mathbf{a}_{K,1}, \dots, \mathbf{a}_{K,N_K}] \quad (1)$$

where $\mathbf{a}_{i,j}$, $i = 1, \dots, K, j = 1, \dots, N_i$ represents the j -th frame in the i -th video. Each $\mathbf{a}_{i,j}$ is assumed to be a column vector of size M where $M = VH$ with V and H the vertical and horizontal dimensions of each frame respectively. For notational convenience, Eq. (1) is rewritten as

$$\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_N] \quad (2)$$

where $N = \sum_{i=1}^K N_i$.

Let \mathbf{Y} be a video-clip in the database with only one frame:

$$\mathbf{Y} = [\mathbf{y}_1] \quad (3)$$

Then, it can be observed that \mathbf{y}_1 admits a sparse representation as

$$\mathbf{y}_1 = \mathbf{A}\mathbf{x}_1 \quad (4)$$

where $\mathbf{x}_1 = (0, \dots, 0, 1, 0, \dots, 0)^t$ is a sparse vector with all coefficients equal to zero except for the entry corresponding to the location of the frame \mathbf{y}_1 in the database, which is equal to 1. Hence, the position of \mathbf{y}_1 in the database is determined by \mathbf{x}_1 .

For a given frame \mathbf{y}_1 , solving for the corresponding \mathbf{x}_1 is an ill-posed problem as the system in (4) is highly under-determined which leads to non-uniqueness of the solutions. However, \mathbf{x}_1 has only one non-zero component, and therefore, our goal is to find the sparsest solution, finding the solution with most components equal to zero. This motivates, following [17], to seek for the solution of

$$\hat{\mathbf{x}}_1 = \arg \min_{\mathbf{x}} \|\mathbf{x}\|_0 \quad \text{subject to} \quad \mathbf{A}\mathbf{x} = \mathbf{y}_1, \quad (5)$$

where $\|\mathbf{x}\|_0$ is the l_0 -quasinorm (the number of non-zero coefficients). However, as is well known, the solution of this optimization problem is NP-hard. Furthermore, there are other issues like noise, different image sizes or even occlusions that make us resort to the CS formulation of the problem.

With $\mathbf{y} = \mathbf{y}_1$, the noisy CS acquisition system can be modeled as

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{n}, \quad (6)$$

where \mathbf{n} is the $M \times 1$ independent, Gaussian, zero-mean noise vector with variance equal to β^{-1} . The problem (5) can then be relaxed using the l_1 -norm formulation as

$$\hat{\mathbf{x}} = \arg \min_{\mathbf{x}} \{\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \tau \|\mathbf{x}\|_1\}, \quad (7)$$

where $\|\cdot\|_1$ denotes the l_1 -norm. Solving (7) is much easier than (5) and has attracted much interest in the CS community.

B. Frame Retrieval Based on Bayesian Compressive Sensing

A number of methods have been proposed to solve the sparse optimization problem in Eq. (7), (see [11] and references therein, see also [18]). In this paper, we formulate the problem using the Bayesian framework following [19] which will also allow us to automatically estimate the regularization parameters, (see [19], [11] for references to parameter estimation). We provide here a brief review of solving (7) using a Bayesian approach.

In Bayesian modeling, all unknowns are treated as stochastic quantities with assigned probability distributions. The joint probability distribution of all quantities is given by

$$p(\mathbf{x}, \boldsymbol{\gamma}, \beta, \mathbf{y}) = p(\mathbf{y}|\mathbf{x}, \beta) p(\mathbf{x}|\boldsymbol{\gamma}) p(\boldsymbol{\gamma}) p(\beta). \quad (8)$$

The observation noise is independent and Gaussian with zero mean and variance equal to β^{-1} , that is, with (6),

$$p(\mathbf{y}|\mathbf{x}, \beta) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x}, \beta^{-1}). \quad (9)$$

It is shown in [19] that the l_1 regularization formulation in (7) is equivalent to using a hierarchical Laplace prior on the coefficients of \mathbf{x} , that is,

$$p(\mathbf{x}|\boldsymbol{\gamma}) = \prod_{i=1}^N \mathcal{N}(x_i|0, \gamma_i), \quad (10)$$

$$p(\gamma_i|\lambda) = \frac{\lambda}{2} \exp\left(-\frac{\lambda\gamma_i}{2}\right), \quad \gamma_i \geq 0, \lambda \geq 0, \quad (11)$$

where $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \dots, \gamma_N)$. Using this specification, the signal distribution $p(\mathbf{x}|\mathbf{y}, \lambda, \beta)$ is estimated as a multivariate Gaussian distribution $\mathcal{N}(\mathbf{x}|\boldsymbol{\eta}, \Theta)$ with parameters

$$\Theta = [\beta\mathbf{A}^t\mathbf{A} + \Lambda]^{-1}, \quad (12)$$

$$\boldsymbol{\eta} = \Theta \beta\mathbf{A}^t\mathbf{y}, \quad (13)$$

with $\Lambda = \text{diag}(1/\gamma_i)$. The hyperparameters $\boldsymbol{\gamma}$ are then estimated by forming the likelihood function

$$\mathcal{L} = -\frac{1}{2} \log |\mathbf{E}| - \frac{1}{2} \mathbf{y}^t \mathbf{E}^{-1} \mathbf{y} + N \log \frac{\lambda}{2} - \frac{\lambda}{2} \sum_i \gamma_i, \quad (14)$$

with $\mathbf{E} = (\beta^{-1}\mathbf{I} + \mathbf{A}\Lambda^{-1}\mathbf{A}^t)$, and maximizing it with respect to each γ_i and λ in an alternating fashion. This procedure results in the updates

$$\gamma_i = -\frac{1}{2\lambda} + \sqrt{\frac{1}{4\lambda^2} + \frac{\langle x_i^2 \rangle}{\lambda}}, \quad (15)$$

$$\lambda = \frac{N-1}{\sum_i \gamma_i/2}, \quad (16)$$

where $\langle x_i^2 \rangle = x_i^2 + \Theta_{ii}$. In summary, at each iteration of the algorithm, given an estimate of $\boldsymbol{\gamma}$ and λ , the estimate of the distribution of \mathbf{x} is calculated using (12) and (13), followed by the estimation of the variances γ_i from (15) and the hyperparameter λ from (16). In addition, [19] proposed a greedy approach that finds the solutions much more efficiently without the need of solving the large linear system in (13). In our work, we use this greedy approach to find the solution of (7).

III. VIDEO RETRIEVAL

A. Searching for video clips without missing frames

In [21] we utilized the CS theory to retrieve a video clip of consecutive frames. More formally, let \mathbf{y} be a video-clip in the database written in the vector form as

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_S \end{bmatrix} \quad (17)$$

where each \mathbf{y}_i represents the i -th frame, and S is the length of the video clip. Next, we built the following matrix for a query of length S

$$\tilde{\mathbf{A}} = \begin{pmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \mathbf{a}_3 & \dots & \mathbf{a}_{N-(S-1)} \\ \mathbf{a}_2 & \mathbf{a}_3 & \mathbf{a}_4 & \dots & \mathbf{a}_{N-S} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{a}_S & \mathbf{a}_{S+1} & \mathbf{a}_{S+2} & \dots & \mathbf{a}_N \end{pmatrix} \quad (18)$$

by shifting the columns of \mathbf{A} in (1). Using this matrix, it can be observed that \mathbf{y} admits a sparse representation as

$$\mathbf{y} = \tilde{\mathbf{A}}\mathbf{x}_0 \quad (19)$$

where $\mathbf{x}_0 = (0, \dots, 0, 1, 0, \dots, 0)^t$ is a sparse vector with all coefficients equal to zero except for the entry corresponding to the location of the video clip \mathbf{y} in the database, which is equal to 1. Hence, the position of \mathbf{y} in the database is determined by \mathbf{x}_0 .

The noisy CS acquisition system for this problem can be modeled as

$$\mathbf{y} = \tilde{\mathbf{A}}\mathbf{x} + \mathbf{n}, \quad (20)$$

where \mathbf{n} is the $(SM) \times 1$ independent, Gaussian, zero-mean noise vector with variance equal to β^{-1} . The retrieval problem is the formulated using the l_1 -norm formulation as

$$\hat{\mathbf{x}} = \arg \min_{\mathbf{x}} \{ \|\mathbf{y} - \tilde{\mathbf{A}}\mathbf{x}\|_2^2 + \tau \|\mathbf{x}\|_1 \}, \quad (21)$$

where $\|\cdot\|_1$ denotes the l_1 -norm.

B. Searching for video clips with missing frames

The method proposed above works well (see [21]) when there are no missing frames in the clip we are looking for. However, when there may be missing frames and their positions in the clip are not known the above $\tilde{\mathbf{A}}$ matrix can not be built and the proposed method is then not applicable. To deal with possibly missing frames we propose the location and classification procedures that are described next.

Let

$$\mathbf{Y} = [\mathbf{y}_1 \ \mathbf{y}_2 \ \dots \ \mathbf{y}_S] \quad (22)$$

be a query video-clip in the database which may contain missing frames. The location procedure consists on finding a candidate video-clip in the database that contains \mathbf{Y} . Following the proposed CS methodology, we can search for each frame independently and then examine if their sparse representations correspond to frames in a video clip in the

database. However, we can reduce the computational time by noticing that the presence of the query video in the database is determined by the location of the first and last frames in the database.

Therefore we start by finding the two sparse vectors $\hat{\mathbf{x}}_1$ and $\hat{\mathbf{x}}_S$ that solve:

$$\hat{\mathbf{x}}_1 = \arg \min_{\mathbf{x}} \{ \|\mathbf{y}_1 - \mathbf{A}\mathbf{x}\|_2^2 + \tau \|\mathbf{x}\|_1 \} \quad (23)$$

and

$$\hat{\mathbf{x}}_S = \arg \min_{\mathbf{x}} \{ \|\mathbf{y}_S - \mathbf{A}\mathbf{x}\|_2^2 + \tau \|\mathbf{x}\|_1 \} \quad (24)$$

respectively. Then, if $\hat{\mathbf{x}}_1$ or $\hat{\mathbf{x}}_S$ do not satisfy the following initial conditions:

- 1) They only have one non-zero component.
- 2) The position marked by $\hat{\mathbf{x}}_1$ is preceding the position marked by $\hat{\mathbf{x}}_S$.

the video clip is classified as not present in the database. If these initial conditions are satisfied, we proceed to examine if \mathbf{Y} is in the database.

We define the candidate video-clip to be retrieved as:

$$\mathbf{C} = [\mathbf{c}_1 \ \mathbf{c}_2 \ \dots \ \mathbf{c}_P] \quad (25)$$

where \mathbf{c}_i $i = 1, \dots, P$ are the frames in the database between the positions marked by $\hat{\mathbf{x}}_1$ and $\hat{\mathbf{x}}_S$.

It is very important to note that P and S do not have to be the same. The query video can have missing intermediate frames. Other systems as [9], [10], [21] strongly utilize the fact that the query video-clip has all intermediate frames, and, as we will see in experimental section, they fail when this does not happen.

Once we have located the tentative position of the first and last frames of the video query in the database we proceed to accept or reject the candidate video clip. We assume that all frames in \mathbf{C} are independent realizations of a Gaussian distribution with mean μ and covariance matrix Σ which are estimated by using

$$\mu = \frac{1}{P} \sum_{i=1}^P \mathbf{c}_i, \quad \Sigma = \frac{1}{P-1} \sum_{i=1}^P (\mathbf{c}_i - \mu)(\mathbf{c}_i - \mu)^t. \quad (26)$$

Therefore if \mathbf{Y} is in the database, then $\bar{\mathbf{y}} = \frac{1}{S} \sum_{i=1}^S \mathbf{y}_i$ will be close to μ when using the Mahalanobis distance. We define the Classification Coefficient (CC), fix a threshold δ and decide that \mathbf{Y} is in the database if and only if:

$$\text{CC}(\mathbf{Y}) \doteq \sqrt{\frac{1}{S} (\bar{\mathbf{y}} - \mu)^t \Sigma^+ (\bar{\mathbf{y}} - \mu)} \leq \delta. \quad (27)$$

where Σ^+ is the Generalized Inverse of Moore-Penrose of Σ .

IV. FEATURE EXTRACTION

In order to perform an efficient search, and due to the size of the frames, feature extraction is needed. We first assume that the frames in the database are downsampled to a reasonable size. This is an important step and a compromise between the size of the downsampled images and the feature extraction process has to be reached. If the downsampled frames are too

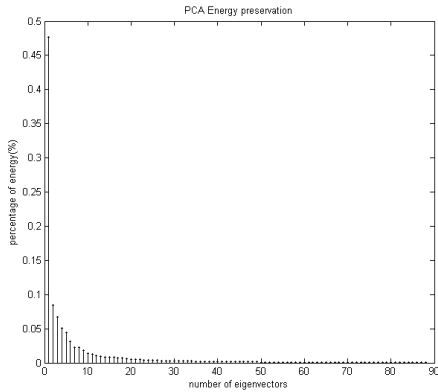


Fig. 1. Energy of the PCA components.

large, feature extraction is very time consuming, on the other hand if the downsampled images are too small, the frames can not be distinguished. See section V for the size of the downsampled images used in the experiments.

We then use a linear feature transformation. The projection from the image space to the feature space can be represented by a matrix $\mathbf{R} \in R^{T \times M}$ with $T \ll M$ which when applied to \mathbf{A} produces

$$\mathbf{R}_{T \times M} \mathbf{A}_{M \times N} = \hat{\mathbf{A}}_{T \times N} \quad (28)$$

Then we can use the proposed retrieval procedure on $\hat{\mathbf{A}}$, which leads to a faster search. In this work we consider \mathbf{R} to be the matrix associated to PCA, (see [9]). Notice that we could have also used a matrix of random projections $\Phi_{T \times M}$.

The PCA transformation retains much of the information in only a reduced set of principal components. The number of preserved dimensions, T , determines the energy loss during the PCA transformation. The energy represented by each PCA coefficient obtained from the test database used in the experiments, which consists of 567146 frames, is shown in Figure 1. Notice that for $T=21$ more than 90 % of the energy is preserved. Furthermore if the CS theoretical conditions are met by \mathbf{A} (see [17]), then $\hat{\mathbf{x}}$ can be recovered by ℓ_1 -minimization with overwhelming probability if $T > 2\log(567146/T)$. In other words, around $T \approx 21$ would suffice to recover the only non-zero component. As we will see in the experimental section, when $T = 21$ the proposed system retrieves all the relevant clips in the database in the noiseless case.

V. SIMULATION RESULTS

In our experiments we used the 2004 NIST TRECVID shot boundary test set. This data set has approximately 6 hours of video in 12 videos (each of about 30 mins long). We split it in two data sets. The positive video repository (or database) consists of 11 videos and the other video forms the negative data set.

The frames are downsampled with a scaling factor of 16 to produce 22×16 downsampled frames. Then the frames are projected using PCA transformation with $T = 21$. In our

TABLE I
CPU TIME USED TO FIND A QUERY.

Methods	15 Frames	30 Frames	60 Frames
KD-PCA [9]	0.04 s	0.06 s	0.12 s
Proposed Method	0.69 s	0.72 s	0.71 s
KD-RP [10]	2.25 s	2.50 s	4.00 s
SR-C [21]	7.57 s	22.17s	out of memory

test, we select randomly 100 positive and 100 negative query videos. The query clip lengths are $S = 15, 30$, and 60 frames. All experiments were performed utilizing an Intel i7 2.8GHz notebook with 8 GB of RAM.

A. Noise free and complete test cases

For noise-free test cases with all its frames our system retrieved all positive cases and rejected all negative one. The results are exactly the same as the ones reported in [9], [10] and [21].

B. Degraded test cases

In real world applications video clips can be corrupted by coding and communication losses, as well as, image formation variations. We evaluate the robustness of our system to both, noise and missing frames, using precision-recall curves. The precision-recall curve [20] is a typical way of characterizing retrieval performance. For a given threshold, let us assume that a is the number of relevant (present in the database) clips retrieved, b the number of relevant clips not retrieved, and c the number of non relevant clips retrieved. Then the precision and recall values are defined by $precision = a/(a + c)$ and $recall = a/(a + b)$. By changing the threshold value we obtain, for a given method, its precision-recall curve. Notice that as the threshold δ in Eq. (27) decreases the recall value is expected to decrease while the precision value is expected to increase.

To simulate the systems in [9], [10], [21] Gaussian noise is added to the query clip to evaluate their robustness to noise. However these systems do not consider test cases with missing frames. This is because they are not designed to retrieve videos with more frames than the query video, and therefore if the query video has missing frames, these systems can not retrieve it.

1) *Noisy test cases.*: We added Gaussian noise to query clips at PSNR of 25dB; this is the noisiest case considered in [9] and [10]. The comparison between our system and KD-PCA [9], KD-RP [10] and SR-C [21], when the query does not contain missing frames, is shown in Fig. 2, for the case $T = 30$. The proposed method obtains a $recall = 0.94$ and a $precision = 1$.

2) *Test cases with missing frames.*: We created queries with missing frames by randomly removing intermediate frames at the following percentages 20%, 50%, and 80%. For noise-free queries all positives cases are retrieved, and all negatives are rejected. For noisy videos with 80% missing frames, out of 30 and 60 frames query clips the $precision = 1$, i.e., all negatives cases are rejected, and $recall = 0.94$ and $recall = 0.96$, respectively. Finally, Fig. 3 shows the precision-recall curves

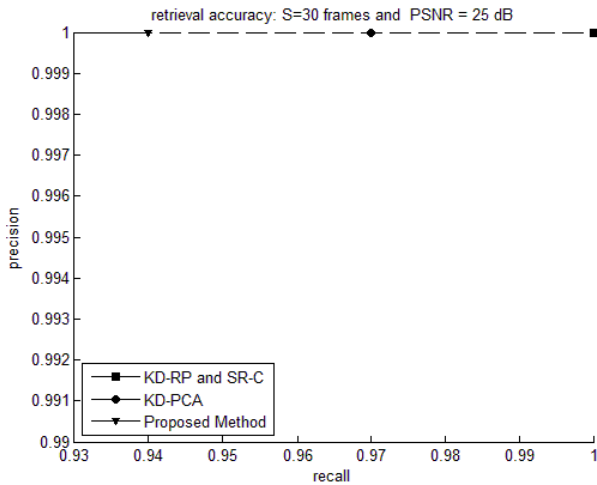


Fig. 2. Comparison of four systems of video retrieval for noisy test.

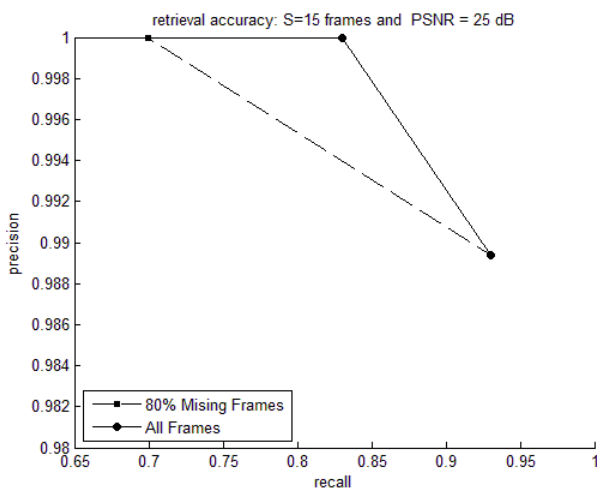


Fig. 3. Precision-Recall curves for noisy query videos with 15 frames and 80% of missing frames, and noisy query videos with all 15 frames.

for a video query of 15 frames and the same query with 80% of missing frames. The system produces more false positives when we remove frames.

VI. CONCLUSION

In this paper we have developed a new system for video retrieval based on the sparse representation framework. We formulate the video retrieval as a sparse reconstruction problem by constructing a database matrix and searching for the sparsest representation of a query clip using the database. We have shown that the proposed system is very effective and robust to noise and missing frames, and does not require sophisticated and data-dependent feature extraction methods. Moreover, the proposed system requires comparable and less computational resources to some of the state-of-the-art methods for video retrieval while providing very high retrieval accuracy.

ACKNOWLEDGMENT

This work has been supported in part by the “Consejería de Innovación, Ciencia y Empresa” of the “Junta de Andalucía” under contract P07-FMQ-02701 and the “Ministerio de Educación y Ciencia” under contract TIN2010-15137.

REFERENCES

- [1] A.M. Ferman, A.M. Tekalp and R. Mehrotra, “Robust Color Histogram Descriptors for Video Segment Retrieval and Identification,” *IEEE Trans. on Image Processing*, vol. 11, no. 5, pp. 497-508, May 2002.
- [2] S.-F. Chang Chen, W. Meng, H.J. Sundaram and H. Di Zhong, “A Fully Automated Content-based Video Search Engine Supporting Spatiotemporal Queries,” *IEEE Trans. on Circuits and System for Video Technology*, vol. 8, no.5, pp. 602-615, Sept. 1998.
- [3] C.-T. Hsu and S.-J. Teng, “Motion Trajectory Based Video Indexing and Retrieval,” *Proc. International Conference on Image Processing*, vol. 1, pp. I-605 - I-608, 2002.
- [4] V. Mezaris, I. Kompatsiaris, N. V. Boulgouris and M. G. Strintzis, “Real-time Compressed-domain Spatiotemporal Segmentation and Ontologies for Video Indexing and Retrieval,” *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 14, no. 5, pp. 606-621, May 2004.
- [5] J. Fan, A.K. Elmagarmid, X. Zhu, W.G. Aref and L. Wu, “ClassView: Hierarchical Video Shot Classification, Indexing, And Accessing,” *IEEE Trans. on Multimedia*, vol. 6, no. 1, pp. 70-86, Feb. 2004.
- [6] C.G.M. Snoek and M. Worring, “Multimedia event-based video indexing using time intervals,” *IEEE Trans. on Multimedia*, vol. 7, no. 4, pp. 638-647, Aug. 2005.
- [7] A. Joly, O. Buisson and C. Frelicot, “Content-based Copy Retrieval using Distortion-based Probabilistic Similarity Search,” *IEEE Trans. on Multimedia*, vol. 9, no. 2, pp. 293-306, Feb. 2007.
- [8] J. Law-to, O. Buisson, L. Chen, M. H. Ipswich, V. Gouet-brunet, A. Joly, N. Boujemaa, I. Laptev, F. Stentiford and M. H. Ipswich, “Video copy detection: a comparative study,” *ACM Int. Conf. on Image and Video Retrieval (CIVR)*, pp. 371-378, 2007.
- [9] L. Gao, Z. Li and A.K. Katsaggelos, “Fast Video Shot Retrieval with Luminance Field Trace Indexing and Geometry Matching,” *Proc. IEEE Int. Conf. on Image Processing (ICIP)*, pp. 1497-1500, 8-11, Oct. 2006.
- [10] L. Gao, Z. Li and A. K. Katsaggelos, “A Video Retrieval Algorithm using Random Projections,” *Proc. IEEE Int. Conf. on Image Processing (ICIP)*, pp. 797-800, 7-10 Nov. 2009.
- [11] J.L. Starck, F. Murtagh, and J. Fadili, *Sparse Image and Signal Processing: Wavelets, Curvelets, Morphological Diversity*, Cambridge University Press, Cambridge, 2010.
- [12] M. Elad, M.A.T. Figueiredo, and M. Yi, “On the Role of Sparse and Redundant Representations in Image Processing,” *Proceedings of IEEE*, vol. 6, 972 - 982, June 2010.
- [13] J. Wright, M. Yi, J. Mairal, G. Sapiro, T.S. Huang, and Y. Yan, “Sparse Representation for Computer Vision and Pattern Recognition,” *Proceedings of IEEE*, vol. 6, 1031 - 1044, June 2010.
- [14] E. J. Candès and T. Tao, “Near-optimal signal recovery from random projections: Universal encoding strategies?,” *IEEE Trans. Inform. Theory*, vol. 52, 5406-5425, 2006.
- [15] D. Donoho, “Compressed sensing,” *IEEE Trans. on Information Theory*, vol. 52, 1289 - 1306, 2006.
- [16] J. Mairal, F. Bach, J. Ponce, G. Sapiro and A. Zisserman, “Discriminative Learned Dictionaries for Local Image Analysis,” *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 1-8, 2008.

- [17] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210 - 227, Feb. 2009.
- [18] S.D. Babacan, R. Molina and A.K. Katsaggelos, "Parameter Estimation in TV Image Restoration Using Variational Distribution Approximation," *IEEE Transactions on Image Processing*, vol. 17, pp. 326 - 339, Feb. 2008.
- [19] S.D. Babacan, R. Molina and A.K. Katsaggelos, "Bayesian Compressive Sensing Using Laplace Priors," *IEEE Trans. on Image Processing*, vol. 19, no. 1, pp. 53-63, Jan. 2010.
- [20] V. Raghavan, P. Bollmann, and G. S. Jung, "A critical investigation of recall and precision as measures of retrieval system performance," *ACM Trans. on Information Systems*, vol. 7, 1989, pp. 205-229.
- [21] P. Ruiz, S.D. Babacan, L.Gao, Z. Li, R. Molina and A.K. Katsaggelos, "Video Retrieval Using Sparse Bayesian Reconstruction," *IEEE Int. Conf. on Multimedia and Expo (ICME), Barcelona (Spain), July 2011.*(Accepted for publication.)

Chapter 7

Conclusions and Future Works

7.1 Conclusions

In this dissertation we have applied Bayesian modeling and inference to image recovery and classification problems. We have shown that image restoration, blind image deconvolution, multispectral image classification, pansharpening, active learning, light field acquisition and video retrieval problems can be modeled using a Bayesian framework, and Bayesian inference has allowed us to find the solutions of these problems. In some cases, point estimators have been utilized to reduce the inference problems to optimization problems. In other cases, the variational inference has allowed us to approximate the posterior distribution, and estimate the model parameters. In the performed experiments, the proposed methods have shown to be very accurate and efficient and, in almost all cases, they outperformed the state-of-the-art methods.

The dissertation has been presented in the modality of “compendium” and has been structured in three blocks: image restoration and blind deconvolution, multispectral image classification and other related problems. Below we detail the specific conclusions and contributions of each area.

7.1.1 Image Restoration and Blind Deconvolution

- First, we have presented a novel image restoration method that uses the Bayesian paradigm to combine two prior models: the total variation (TV) model that preserves edge structure while imposes smoothness on the solution and controls the noise, and the Poisson singular integral (PSI) model which is capable to preserve the textures but cannot differentiate between highly detailed textures and noise. The final product is a restoration algorithm that combines the advantages of the two models. A study of TV and PSI models and the parameters that control their shape has been carried out. The work concludes that neither the TV nor the PSI image models alone can suc-

cessfully recover the textures and control the noise. A set of experiments has been carried out, where the proposed method has been compared against both classical and state of art methods. The experimental results supported that for images with a combination of detailed and smooth regions, the proposed restoration method, which combines TV and PSI prior models, provides the best restorations.

- For the BID problem we have written a review of the recent literature on Bayesian blind image deconvolution (BID) methods. We have stated that two events have marked the recent history of BID: the predominance of variational Bayes (VB) inference as a tool to solve BID problems and the increasing interest of the computer vision community in solving BID problems. We have shown that VB inference in combination with recent image models like the ones based on Super Gaussian (SG) and scale mixture of Gaussians (SMG) representations have led to the use of very general and powerful tools to provide clear images from blurry observations. In the provided review emphasis has been paid on VB inference and the use of SG and SMG models with coverage of recent advances in sampling methods. We have also provided examples of current state of the art BID methods and have discussed problems that very likely will mark the near future of BID.

7.1.2 Multispectral Image Classification Problems

- We have shown that pansharpening techniques can be used to increase the performance of classification methods when they are applied to multispectral images. We have addressed the problem of adaptively modifying the parameter of a pansharpening method in order to improve the precision and recall figures of merit of a classifier on a given class without deteriorating its performance over the other classes. The validity of the proposed technique has been demonstrated using a real Quickbird image.
- We have also presented a new method to jointly filter and classify a signal or an image. Using Bayesian modeling and variational inference we have developed an iterative procedure to jointly estimate the classifier parameters, the filterbank and the model parameters. We have experimentally shown that the estimated filters improves the classifier performance. The proposed method has been compared with other classification/filtering approaches, and experimental results have shown that the new method is both more accurate and more efficient.
- We have presented a non-parametric Bayesian learning approach based on kernels for remote sensing image classification. The Bayesian methodology efficiently tackles purely supervised and active learning approaches, and shows

competitive performance when compared to support vector machines (SVMs) and recent active learning (AL) approaches. An incremental learning approach based on three different approaches was presented: maximum differential of entropies, minimum distance to decision boundary, and minimum normalized distance. Automatic parameter estimation is solved by using the evidence Bayesian approach, the kernel trick, and the marginal distribution of the observations instead of the posterior distribution of the adaptive parameters. The proposed approach was tested on several scenes dealing with urban monitoring problems using multispectral and SAR data. We observed that, while similar results are obtained by SVMs in supervised mode, an improvement in accuracy and convergence is observed for the active learning scenario. Interestingly our methods do not only provide point-wise class predictions but confidence intervals.

- We have also developed a multiclass classification system using a prior on the adaptive coefficients based on the ℓ_p pseudo-norm. The contribution of the adaptive coefficients corresponding with no-relevant data will be zero, which allow us to identify the irrelevant coefficients. Variational inference has been used to estimate all the model parameters and connections with independent Gaussian priors established. The predictive distribution of the classes has been calculated. This distribution has been used to define two active learning methods, named Minimum Probability Criteria and Maximum Entropy Criteria. Experimental results have shown that the use of ℓ_p -priors allows the classifier to select discriminative features and discard non-relevance components. The proposed approach has shown higher accuracy than SVM methods in both classification and AL problems.

7.1.3 Other Related Problems (Light Field Acquisition and Video Retrieval)

- We have developed a new programmable aperture camera prototype to capture the light field. The prototype was constructed in collaboration with the Instituto de Astrofísica de Andalucía (IAA). In [15] we developed a system which uses the compressive sensing theory to capture the light field by taking much less observations than views of light field. In [16], we addressed the problem of recovering blurred light fields. We developed a method to deconvolve blurred light fields and experimentally shown that it is possible to obtain sharp images from blurred observations using both synthetic and real images.
- We have developed a robust and efficient system for video retrieval, based on the use of sparse representation, compressive sensing and Bayesian modeling of the video retrieval problem. Experimental results demonstrate that the proposed method performs better than existing state-of-art systems and also

its robustness against noise. We have also shown that the new system is very effective and robust to noise and missing frames, and does not require sophisticated and data-dependent feature extraction methods.

7.2 Future Works

7.2.1 Image Restoration and Blind Deconvolution

- In Chapter 3 we proposed a restoration method based on model combination to simultaneously preserve edges and textures while controlling noise. The new restoration method depends on a set of parameters that need to be estimated for each image. However we have not so far addressed the parameter estimation problem. For future work we want to introduce variational inference in order to estimate the model parameters.

7.2.2 Multispectral Image Classification Problems

- In [71] we proposed a model where we jointly estimates an optimal filterbank and trained a GP classifier. To link both procedures, in [71] we used a parameter which was increasing in each iteration. We are currently working on solving this constrained optimization problem using the Alternating Method of Multipliers (ADMM) [100]. ADMM is often utilized to transform a constrained optimization problem into an unconstrained one through the use of the augmented Lagrangian. The use of this approach and in our case, it will allow to automatically estimate all the model parameters.
- In [73] we used kernel methods to find non-linear decision boundaries for classification problems. In [74] we presented a method to eliminate non-relevant classification features. In future work we want to develop systems capable of finding non-linear decision boundaries and discarding information non-relevant for classification by combining kernel method and sparse priors.

7.2.3 Other Related Problems

- As we have studied in [16] the camera prototype utilize a mask in front of the lens, which produces vigneting in the observed images. To avoid this problem, we plan to develop a new camera prototype where the mask will be located in the aperture plane of the lens.
- In order to obtain a robust video copy detector, the system proposed in [75] has to work with transformed videos. In [76] we addressed the problem of missing frames. For future works the system will be improved to deal with other transformations like, video re-coding or cropped videos.

-
- We have shown that Bayesian modeling and inference are powerful tools to address image recovery and classification problems. In the future we want to apply the learned methodology to solve more inverse problems. In particular we want to explore the use of Bayesian modelling and inference in crowdsourcing [101, 102], sensor fusion [103, 104], as well as threat detection in millimeter images [105, 106].

Bibliography

- [1] C. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, 1st edition, Feb. 2007.
- [2] D. Barber, *Bayesian Reasoning and Machine Learning*, Cambridge University Press, 2012.
- [3] K.P. Murphy, *Machine Learning: A Probabilistic Perspective*, The MIT Press, Cambridge, MA, 1st edition, Aug. 2012.
- [4] S.J.D. Prince, *Computer Vision: Models, Learning, and Inference*, Cambridge University Press, New York, 1st edition, June 2012.
- [5] S.D. Babacan, *Bayesian Techniques for Image Recovery*, Ph.D. thesis, Northwestern University, 2009.
- [6] R. Molina, J. Nuñez, F.J. Cortijo, and J. Mateos, “Image Restoration in Astronomy: A Bayesian Perspective,” *IEEE Signal Processing Magazine*, vol. 18, no. 2, pp. 11–29, Mar. 2001.
- [7] A. Carasso, “Singular Integrals, Image Smoothness, and the Recovery of Texture in Image Deblurring,” *SIAM Journal on Applied Mathematics*, vol. 64, no. 5, pp. 1749–1774, 2004.
- [8] T.E. Bishop, S.D. Babacan, B. Amizic, T. Chan, R. Molina, and A.K. Katsaggelos, “Blind Image Deconvolution: Problem Formulation and Existing Approaches,” in *Blind Image Deconvolution: Theory and Applications*. CRC Press, Campisi, P. and Egiazarian, K. edition, 2007.
- [9] S.D. Babacan, R. Molina, M.N. Do, and A.K. Katsaggelos, “Bayesian Blind Deconvolution with General Sparse Image Priors,” in *Computer Vision – ECCV 2012*, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds., number 7577 in Lecture Notes in Computer Science, pp. 341–355. Springer Berlin Heidelberg, 2012.
- [10] P. Ruiz, X. Zhou, J. Mateos, R. Molina, and A.K. Katsaggelos, “Variational Bayesian Blind Image Deconvolution: A review,” *Digital Signal Processing*, 2015.

-
- [11] A.K. Katsaggelos, R. Molina, and J. Mateos, *Super Resolution of Images and Video*, Morgan and Claypool, 2007.
- [12] M.G. Kang and S. Chaudhuri, “Super-resolution Image Reconstruction,” *IEEE Signal Processing Magazine*, vol. 20, no. 3, pp. 19–20, May 2003.
- [13] S. Chaudhuri and J. Manjunath, *Motion-free Super-resolution*, Springer, 2005.
- [14] C.-K. Liang, T.-H. Lin, B.-Y. Wong, C. Liu, and H.H. Chen, “Programmable Aperture Photography: Multiplexed Light Field Acquisition,” in *ACM SIGGRAPH 2008 Papers*, New York, NY, USA, 2008, SIGGRAPH ’08, pp. 55:1–55:10, ACM.
- [15] S.D. Babacan, R. Ansorge, M. Luessi, P. Ruiz-Matarán, R. Molina, and A.K. Katsaggelos, “Compressive Light Field Sensing,” *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4746–4757, 2012.
- [16] P. Ruiz, J. Mateos, M.C. Cardenas, S. Nakajima, R. Molina, and A.K. Katsaggelos, “Light Field Acquisition from Blurred Observations Using a Programmable Coded Aperture Camera,” in *Proceedings of the 21st European Signal Processing Conference (EUSIPCO 2013)*, Sept. 2013, pp. 1–5.
- [17] I. Amro, J. Mateos, M. Vega, R. Molina, and A. K. Katsaggelos, “A Survey of Classical Methods and New Trends in Pansharpening of Multispectral Images,” *EURASIP Journal on Advances in Signal Processing*, vol. 2011, no. 1, pp. 79, September 2011.
- [18] J.M. Bioucas-Dias and M.A.T. Figueiredo, “An Iterative Algorithm for Linear Inverse Problems with Compound Regularizers,” in *15th IEEE International Conference on Image Processing (ICIP 2008)*, Oct. 2008, pp. 685–688.
- [19] P. Mitra, B. Uma-Shankar, and S.K. Pal, “Segmentation of Multispectral Remote Sensing Images Using Active Support Vector Machines,” *Pattern Recognition Letters*, vol. 25, no. 9, pp. 1067–1074, July 2004.
- [20] L. Gomez-Chova, D. Fernández-Prieto, J. Calpe, E. Soria, J. Vila, and G. Camps-Valls, “Urban Monitoring Using Multi-temporal SAR and Multispectral Data,” *Pattern Recognition Letters*, vol. 27, no. 4, pp. 234–243, Mar. 2006.
- [21] J. Li, J.M. Bioucas-Dias, and A. Plaza, “Hyperspectral Image Segmentation Using a New Bayesian Approach With Active Learning,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 10, pp. 3947–3960, Oct. 2011.

-
- [22] B. Settles, “Active Learning Literature Survey,” Technical Report, Department of Computer Science, University of Wisconsin-Madison, Madison, WI, USA, 2009.
- [23] J. Paisley, X. Liao, and L. Carin, “Active Learning and Basis Selection for Kernel-Based Linear Models: A Bayesian Perspective,” *IEEE Transactions on Signal Processing*, vol. 58, no. 5, pp. 2686–2700, May 2010.
- [24] D. Tuia, F. Ratle, F. Pacifici, M.F. Kanevski, and W.J. Emery, “Active Learning Methods for Remote Sensing Image Classification,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 7, pp. 2218–2232, July 2009.
- [25] L. Gao, Z. Li, and A.K. Katsaggelos, “Fast Video Shot Retrieval with Luminance Field Trace Indexing and Geometry Matching,” in *IEEE International Conference on Image Processing (ICIP 2006)*, Oct. 2006, pp. 1497–1500.
- [26] L. Gao, Z. Li, and A.K. Katsaggelos, “A Video Retrieval Algorithm Using Random Projections,” in *16th IEEE International Conference on Image Processing (ICIP 2009)*, Nov. 2009, pp. 797–800.
- [27] J. Law-To, L. Chen, A. Joly, I. Laptev, O. Buisson, V. Gouet-Brunet, N. Boujemaa, and F. Stentiford, “Video Copy Detection: A Comparative Study,” in *Proceedings of the 6th ACM International Conference on Image and Video Retrieval*, New York, NY, USA, 2007, CIVR ’07, pp. 371–378, ACM.
- [28] J. Wang, Y. Shi, W. Ding, and B. Yin, “A Low-rank Matrix Completion Based Intra Prediction for H.264/AVC,” in *IEEE 13th International Workshop on Multimedia Signal Processing*, 2011, pp. 1–6.
- [29] S.D. Babacan, M. Luessi, R. Molina, and A.K. Katsaggelos, “Sparse Bayesian Methods for Low-Rank Matrix Estimation,” *IEEE Transactions on Signal Processing*, vol. 60, no. 8, pp. 3964–3977, 2012.
- [30] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, H. Mobahi, and Y. Ma, “Toward a Practical Face Recognition System: Robust Alignment and Illumination by Sparse Representation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 2, pp. 372–386, 2012.
- [31] T. Goodall, S. Gibson, and M.C. Smith, “Parallelizing Principal Component Analysis for Robust Facial Recognition Using CUDA,” in *Symposium on Application Accelerators in High Performance Computing*, 2012, pp. 121–124.
- [32] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, and Yi Ma, “Robust Face Recognition via Sparse Representation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

-
- [33] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma, “RASL: Robust Alignment by Sparse and Low-rank Decomposition for Linearly Correlated Images,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 763–770.
- [34] J.P. Haldar and Z. Liang, “Spatiotemporal Imaging with Partially Separable Functions: A Matrix Recovery Approach,” in *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, 2010, pp. 716–719.
- [35] Z. Chen, S.D. Babacan, R. Molina, and A.K. Katsaggelos, “Variational Bayesian Methods For Multimedia Problems,” *IEEE Transactions on Multimedia*, vol. 16, no. 4, pp. 1000–1017, June 2014.
- [36] M.J. Beal, *Variational Algorithms for Approximate Bayesian Inference*, Ph.D. thesis, The Gatsby Computational Neuroscience Unit, University College London, 2003.
- [37] J. Miskin, *Ensemble Learning for Independent Component Analysis*, Ph.D. thesis, Astrophysics Group, University of Cambridge, 2000.
- [38] S.T. Jaakkola and Jordan I.M., “Bayesian Parameter Estimation via Variational Methods,” *Statistics and Computing*, vol. 10, no. 1, pp. 25–37, 2000.
- [39] D.G. Tzikas, C. L. Likas, and N. P. Galatsanos, “The Variational Approximation for Bayesian Inference,” *IEEE Signal Processing Magazine*, vol. 25, no. 6, pp. 131–146, 2008.
- [40] C.W. Fox and S.J. Roberts, “A Tutorial on Variational Bayesian Inference,” *Artificial Intelligence Review*, vol. 38, no. 2, pp. 85–95, 2012.
- [41] S. Horaczek, “How Many Photos Are Uploaded to The Internet Every Minute?,” May 2013.
- [42] M. Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*, Springer, New York, 2010 edition, 2010.
- [43] J. Hadamard, *Lectures on Cauchy’s Problem in Linear Partial Differential Equations*, Yale University Press, New Haven, CT, 1923.
- [44] S. Liang, *Quantitative Remote Sensing of Land Surfaces*, Wiley-Interscience, Hoboken, N.J, 1st edition, Dec. 2003.
- [45] T.M. Lillesand, R.W. Kiefer, and J. Chipman, *Remote Sensing and Image Interpretation*, John Wiley & Sons, New York, Dec. 2008.

-
- [46] G. Camps-Valls, D. Tuia, L. Gómez-Chova, S. Jiménez, and J. Malo, “Remote Sensing Image Processing,” *Synthesis Lectures on Image, Video, and Multimedia Processing*, vol. 5, no. 1, pp. 1–192, 2011.
- [47] F. Palsson, J.R. Sveinsson, J.A. Benediktsson, and H. Aanaes, “Classification of Pansharpened Urban Satellite Images,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, no. 1, pp. 281–297, Feb 2012.
- [48] R. Flamary, D. Tuia, B. Labbe, G. Camps-Valls, and A. Rakotomamonjy, “Large Margin Filtering,” *IEEE Transactions on Signal Processing*, vol. 60, no. 2, pp. 648–659, Feb. 2012.
- [49] Y. LeCun and Y. Bengio, “Convolutional Networks for Images, Speech, and Time Series,” in *The Handbook of Brain Theory and Neural Networks*, M. A. Arbib, Ed., pp. 255–258. MIT Press, 1998.
- [50] B. De Vries and J.C. Principe, “The Gamma Model—A New Neural Model for Temporal Processing,” *Neural Networks*, vol. 5, pp. 565–576, 1992.
- [51] S. Lawrence, A.C. Tsoi, and A.D. Back, “The Gamma MLP for Speech Phoneme Recognition,” in *Advances in Neural Information Processing Systems*. 1996, pp. 785–791, MIT Press.
- [52] B. Demir, C. Persello, and L. Bruzzone, “Batch-Mode Active-Learning Methods for the Interactive Classification of Remote Sensing Images,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 3, pp. 1014–1031, Mar. 2011.
- [53] D. Tuia, M. Volpi, L. Copa, M. Kanevski, and J. Muñoz-Mari, “A Survey of Active Learning Algorithms for Supervised Remote Sensing Image Classification,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 3, pp. 606–617, June 2011.
- [54] M. Levoy, “Experimental Platforms for Computational Photography,” *IEEE Computer Graphics and Applications*, vol. 30, pp. 81–87, 2010.
- [55] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, Cambridge, UK, New York, 2nd edition, Apr. 2004.
- [56] D. Farin, Y. Morvan, and P.H.N. de With, “View Interpolation Along a Chain of Weakly Calibrated Cameras,” in *IEEE Workshop on Content Generation and Coding for 3D-Television*, 2006.

-
- [57] Y-R Ng, C. Pitts, and T. Knight, “Lytro ILLUM,” <https://www.lytro.com>, 2013.
- [58] C. Perwass and L. Wietzke, “Light Field Camera Technology,” <http://www.raytrix.de/index.php/Technology.html>, Mar. 2013.
- [59] H. Nagahara, C. Zhou, T. Watanabe, H. Ishiguro, and S.K. Nayar, “Programmable Aperture Camera Using LCoS,” in *Computer Vision – ECCV 2010*, K. Daniilidis, P. Maragos, and N. Paragios, Eds., vol. 6316 of *Lecture Notes in Computer Science*, pp. 337–350. Springer Berlin Heidelberg, 2010.
- [60] J. Yuan, L.-Y. Duan, Q. Tian, and C. Xu, “Fast and Robust Short Video Clip Search Using an Index Structure,” in *Proceedings of the 6th ACM SIGMM International Workshop on Multimedia Information Retrieval*, New York, NY, USA, 2004, MIR ’04, pp. 61–68, ACM.
- [61] J. Fan, A.K. Elmagarmid, X. Zhu, W.G. Aref, and L. Wu, “ClassView: Hierarchical Video Shot Classification, Indexing, and Accessing,” *IEEE Transactions on Multimedia*, vol. 6, no. 1, pp. 70–86, Feb. 2004.
- [62] V. Mezaris, I. Kompatsiaris, N.V. Boulgouris, and M.G. Strintzis, “Real-time Compressed-domain Spatiotemporal Segmentation and Ontologies for Video Indexing and Retrieval,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 5, pp. 606–621, May 2004.
- [63] Y.-H. Ho, C.-W. Lin, J.-F. Chen, and H.-Y.M. Liao, “Fast Coarse-to-fine Video Retrieval Using Shot-level Spatio-temporal Statistics,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 5, pp. 642–648, May 2006.
- [64] C.-T Hsu and S.-J. Teng, “Motion Trajectory Based Video Indexing and Retrieval,” in *2002 Proceedings of International Conference on Image Processing*, 2002, vol. 1, pp. I-605–I-608 vol.1.
- [65] C.-W Ngo, T.-C Pong, and H.-J Zhang, “On Clustering and Retrieval of Video Shots Through Temporal Slices Analysis,” *IEEE Transactions on Multimedia*, vol. 4, no. 4, pp. 446–458, 2002.
- [66] C.G.M. Snoek and M. Worring, “Multimedia Event-based Video Indexing Using Time Intervals,” *IEEE Transactions on Multimedia*, vol. 7, no. 4, pp. 638–647, 2005.
- [67] A. Joly, O. Buisson, and C. Frelicot, “Content-Based Copy Retrieval Using Distortion-Based Probabilistic Similarity Search,” *IEEE Transactions on Multimedia*, vol. 9, no. 2, pp. 293–306, Feb. 2007.

- [68] H. Madero-Orozco, P. Ruiz, J. Mateos, R. Molina, and A.K. Katsaggelos, "Image Deblurring Combining Poisson Singular Integral and Total Variation Prior Models," in *21th European Signal Processing Conference (EUSIPCO 2013)*. Sept. 2013, p. 1569744251, Marrakech (Morocco).
- [69] P. Ruiz, H. Madero-Orozco, J. Mateos, O.O. Vergara-Villegas, R. Molina, and A.K. Katsaggelos, "Combining Poisson Singular Integral and Total Variation Prior Models in Image Restoration," *Signal Processing*, vol. 103, pp. 296–308, Oct. 2014.
- [70] P. Ruiz, J.V. Talens, J. Mateos, R. Molina, and A.K. Katsaggelos, "Interactive Classification Oriented Superresolution of Multispectral Images," in *7th International Workshop Data Analysis in Astronomy (DAA2011)*, L. Scarsi and V.D. Gesù, Eds. Apr. 2011, pp. 77–85, Erice (Italy).
- [71] P. Ruiz, J. Mateos, R. Molina, and A.K. Katsaggelos, "Learning Filters in Gaussian Process Classification Problems," in *IEEE International Conference on Image Processing (ICIP 2014)*, Oct. 2014, pp. 2913–2917.
- [72] P. Ruiz, J. Mateos, R. Molina, and A.K. Katsaggelos, "A Bayesian Active Learning Framework for a Two-Class Classification Problem," in *MUSCLE International Workshop on Computational Intelligence for Multimedia Understanding*, E. Salerno, A.E. Çetin, and O. Salvetti, Eds. 2012, vol. LNCS-7252, pp. 42–53, Pisa (Italy).
- [73] P. Ruiz, J. Mateos, G. Camps-Valls, R. Molina, and A.K. Katsaggelos, "Bayesian Active Remote Sensing Image Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 4, pp. 2186–2196, Apr. 2014.
- [74] P. Ruiz, N. Perez de la Blanca, R. Molina, and A.K. Katsaggelos, "Bayesian Classification and Active Learning Using Lp-priors. Application to Image Segmentation," in *Proceedings of the 22nd European Signal Processing Conference (EUSIPCO2014)*, Sept. 2014, pp. 1183–1187.
- [75] P. Ruiz, S.D. Babacan, L. Gao, Z. Li, R. Molina, and A.K. Katsaggelos, "Video Retrieval Using Sparse Bayesian Reconstruction," in *IEEE International Conference on Multimedia and Expo (ICME2011)*. Barcelona (Spain), July 2011, pp. 1–6.
- [76] P. Ruiz, S.D. Babacan, R. Molina, and A.K. Katsaggelos, "Retrieval of Video Clips with Missing Frames Using Sparse Bayesian Reconstruction," in *7th International Symposium on Image and Signal Processing and Analysis (ISPA 2011)*. Dubrovnik (Croatia), September 2011, pp. 443–448.

-
- [77] R. Molina, A.K. Katsaggelos, and J. Mateos, “Bayesian and Regularization Methods for Hyperparameter Estimation in Image Restoration,” *IEEE Transactions on Image Processing*, vol. 8, no. 2, pp. 231–246, 1999.
- [78] L.I. Rudin, S. Osher, and E. Fatemi, “Nonlinear Total Variation Based Noise Removal Algorithms,” *Physica D: Nonlinear Phenomena*, vol. 60, no. 1–4, pp. 259–268, Nov. 1992.
- [79] S.D. Babacan, R. Molina, and A.K. Katsaggelos, “Variational Bayesian Blind Deconvolution Using a Total Variation Prior,” *IEEE Transactions on Image Processing*, vol. 18, no. 1, pp. 12–26, 2009.
- [80] B. Amizic, R. Molina, and A.K. Katsaggelos, “Sparse Bayesian Blind Image Deconvolution with Parameter Estimation,” *Eurasip Journal on Image and Video Processing*, vol. 2012, no. 1, nov 2012.
- [81] A. Levin, R. Fergus, F. Durand, and W.T. Freeman, “Image and Depth from a Conventional Camera with a Coded Aperture,” *ACM Transactions on Graphics (TOG) - Proceedings of ACM SIGGRAPH 2007*, vol. 26, no. 3, pp. 70, 2007.
- [82] Q. Shan, J. Jia, and A. Agarwala, “High-quality Motion Deblurring from a Single Image,” *ACM Transactions on Graphics (TOG) - Proceedings of ACM SIGGRAPH 2008*, vol. 27, no. 3, pp. 73, 2008.
- [83] M. Vega, R. Molina, and A.K. Katsaggelos, “Parameter Estimation in Bayesian Blind Deconvolution with Super Gaussian Image Priors,” in *European Signal Processing Conference (EUSIPCO)*, 2014, pp. 1632–1636.
- [84] X. Zhou, R. Molina, F. Zhou, and A. Katsaggelos, “Fast Iteratively Reweighted Least Squares for Lp Regularized Image Deconvolution and Reconstruction,” in *IEEE International Conference on Image Processing (ICIP 2014)*, oct 2014, pp. 1783–1787.
- [85] G.C. Cawley, N.L.C. Talbot, and M. Girolami, “Sparse Multinomial Logistic Regression via Bayesian L1 Regularisation,” in *Neural Information Processing Systems*, 2006, pp. 209–216.
- [86] S. Ryali, K. Supekar, D.A. Abrams, and V. Menon, “Sparse Logistic Regression for Whole-brain Classification of fMRI Data,” *NeuroImage*, 2010.
- [87] S.D. Babacan, R. Molina, and A.K. Katsaggelos, “Bayesian Compressive Sensing Using Laplace Priors,” *IEEE Transactions on Image Processing*, vol. 19, no. 1, pp. 53–63, 2010.

-
- [88] J.A. Palmer, K. Kreutz-Delgado, and S. Makeig, “Strong Sub- and Super-Gaussianity,” in *Latent Variable Analysis and Signal Separation*, V. Vigneron, V. Zarzoso, E. Moreau, R. Gribonval, and E. Vincent, Eds., vol. 6365 of *Lecture Notes in Computer Science*, pp. 303–310. Springer Berlin Heidelberg, 2010.
- [89] H. Zhang and D. Wifp, “Non-Uniform Camera Shake Removal Using a Spatially-Adaptive Sparse Penalty,” in *Advances in Neural Information Processing Systems*, 2013, pp. 1556–1564.
- [90] A. Mohammad-Djafari, “Bayesian Blind Deconvolution of Images Comparing JMAP, EM and BVA with a Student-t a Priori Model,” in *International Workshops on Electrical Computer Engineering Subfields*, 2014, pp. 98–103.
- [91] J.O. Berger, *Statistical Decision Theory and Bayesian Analysis*, Springer Series in Statistics. Springer New York, 1985.
- [92] H. Raïffa and R. Schlaifer, *Applied Statistical Decision Theory*, Division of Research, Graduate School of Business Administration, Harvard University, 1961.
- [93] A. Levin, Y. Weiss, F. Durand, and W.T. Freeman, “Efficient Marginal Likelihood Optimization in Blind Deconvolution,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 2657–2664.
- [94] A. Levin, Y. Weiss, F. Durand, and W.T. Freeman, “Understanding Blind Deconvolution Algorithms,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 12, pp. 2354–2367, Dec. 2011.
- [95] J.J.K. O Ruanaidh and W.J. Fitzgerald, *Numerical Bayesian Methods Applied to Signal Processing*, Springer Verlag, 1996.
- [96] W.J. Fitzgerald, “Markov Chain Monte Carlo Methods with Applications to Signal Processing,” *Signal Processing*, vol. 81, no. 1, pp. 3–18, 2001.
- [97] S. Gulam-Razul, W.J. Fitzgerald, and C. Andrieu, “Bayesian Deconvolution in Nuclear Spectroscopy Using RJMCMC,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2002, vol. 2, pp. 1309–1312.
- [98] G. Parisi, *Statistical Field Theory*, Perseus Books, Reading, Mass, new edition, Nov. 1998.
- [99] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

-
- [100] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, “Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers,” *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, Jan. 2011.
- [101] V.C. Raykar, S. Yu, L.H. Zhao, G.H. Valadez, C. Florin, L. Bogoni, and L. Moy, “Learning from Crowds,” *Journal of Machine Learning Research*, vol. 11, pp. 1297–1322, apr 2010.
- [102] Y. Yan, R. Rosales, G. Fung, R. Subramanian, and J. Dy, “Learning from Multiple Annotators with Varying Expertise,” *Machine Learning*, vol. 95, no. 3, pp. 291–327, 2014.
- [103] A. Kapoor, H. Ahn, and R.W. Picard, “Mixture of Gaussian Processes for Combining Multiple Modalities,” in *Multiple Classifier Systems*, N.C. Oza, R. Polikar, J. Kittler, and F. Roli, Eds., number 3541 in Lecture Notes in Computer Science, pp. 86–96. Springer Berlin Heidelberg, Jan. 2005.
- [104] A.R. Groves, C.F. Beckmann, S.M. Smith, and M.W. Woolrich, “Linked Independent Component Analysis For Multimodal Data Fusion,” *Neuroimage*, vol. 54, no. 3, pp. 2198–2217, 2011.
- [105] M.C. Kemp, “Millimetre Wave and Terahertz Technology for Detection of Concealed Threats - A Review,” in *Joint 32nd International Conference on Infrared and Millimeter Waves, and the 15th International Conference on Terahertz Electronics. IRMMW-THz.*, Sept 2007, pp. 647–648.
- [106] S.W. Harmer, N. Bowring, D. Andrews, N.D. Rezgui, M. Southgate, and S. Smith, “A Review of Nonimaging Stand-Off Concealed Threat Detection with Millimeter-Wave Radar [Application Notes],” *Microwave Magazine, IEEE*, vol. 13, no. 1, pp. 160–167, Jan 2012.