TESIS DOCTORAL

New Reconstruction Techniques for Image/Video Communication

Universidad de Granada

Departamento de Teoría de la Señal, Telemática y Comunicaciones

Programa de Doctorado en Sistemas Multimedia

Autor:

Ján Koloda

Directores:

Antonio Miguel Peinado Herreros

Victoria Eugenia Sánchez Calle

Granada, Julio de 2014

Editor: Editorial de la Universidad de Granada
Autor: Ján Koloda
D.L.: GR 2147-2014
ISBN: 978-84-9083-265-3

El doctorando **D. Ján Koloda** y los directores de la tesis **D. Antonio Miguel Peinado Herreros** y **Dª. Victoria Eugenia Sánchez Calle** garantizamos, al firmar esta tesis doctoral, que el trabajo ha sido realizado por el doctorando bajo la dirección de los 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, julio de 2014

Directores de la Tesis | Doctorando

Fdo.: Antonio Peinado | Fdo.: Victoria Sánchez | Fdo.: Ján Koloda

Agradecimientos

Esta tesis está dedicada a todas las personas que han hecho posible su realización, que me han apoyado y motivado durante este interesante camino.

En primer lugar querría expresar mi más profunda gratitud a mis directores de tesis. A Antonio por su infinita dedicación y pasión por nuestra investigación y trabajo. Por todo el conocimiento y experiencia que me ha transferido. Y también por su sentido de humor y la pasión que compartimos por la bici y el senderismo. A Victoria por todos sus valiosos consejos y la inmensa ayuda en la elaboración de los artículos. Victoria y Antonio, me siento muy afortunado por poder conoceros, por poder compartir estos años con vosotros y por todo lo que me habéis enseñado. ¡Gracias!

Obrovská vďaka patrí mojim rodičom, Štefanovi a Viere, ako i sestre Ľudmile, za ich všestrannú podporu.

Me gustaría agradecer a los miembros de nuestro fructífero congresillo semanal ICOM: mis amigos Iván, Domingo, Ángel y José Luis. Y a mis amigos Martin, Pedro y José por nuestras estimulantes conversaciones, salidas y viajes.

I would also like to thank Jan and Søren, for all those fruitful discussions we had in Aalborg, and Jürgen, for the interest he showed for our research during my stay in Erlangen.

A osobitne na záver som si nechal osobu, ktorá bezprostredne zažila úsilie, čo si vyžiadala táto práca, a ktorá už pravdepodobne vie o multimediálnych signáloch viac, ako by sama chcela. Vďaka Ti, Kristínka!

Y finalmente, una última y gran mención a todos aquellos que no aparecen explícitamente en esta memoria, pero que siempre estarán en la mía. En este trabajo podéis encontrar vuestra huella también.

GRACIAS A TODOS

Contents

1 Introduction to the PhD Thesis **3**
1.1 Introduction . . . 3
1.1.1 Interpolation based EC . . . 7
1.1.2 Statistically driven EC . . . 8
1.1.3 EC in transformed domains . . . 8
1.1.4 Filling order . . . 9
1.2 Starting hypotheses . . . 9
1.3 Objectives . . . 10
1.4 Thesis proposals . . . 11
1.4.1 Interpolation based EC: Spatial Error Concealment Based on Edge Visual Clearness for Image/Video Communication . . . 11
1.4.2 Statistically driven EC . . . 12
1.4.2.1 Multimedia Signal Reconstruction by Sparse Linear Prediction . . . 12
1.4.2.2 Multimedia Signal Reconstruction by Kernel-based MMSE . . . 13
1.4.3 EC in a transformed domain: Frequency Selective Extrapolation with Residual Filtering . . . 13
1.4.4 Improved filling order . . . 14

2 Publications: Published and Accepted Papers **15**
2.1 Interpolation based EC . . . 15
2.1.1 Spatial Error Concealment Based on Edge Visual Clearness for Image/Video Communication . . . 15
2.2 Statistically driven EC . . . 27
2.2.1 Multimedia Signal Reconstruction by Sparse Linear Prediction . . . 27
2.2.1.1 Sequential Error Concealment for Video/Images by Sparse Linear Prediction . . . 27
2.2.1.2 Speech Reconstruction by Sparse Linear Prediction . . . 43
2.2.2 Multimedia Signal Reconstruction by Kernel-based MMSE . . . 55
2.2.2.1 On the Application of Multivariate Kernel Density Estimation to Image Error Concealment . . . 55
2.2.2.2 Kernel-based MMSE Multimedia Signal Reconstruction and its Application to Spatial Error Concealment . . . 63
2.3 EC in transformed domain . . . 75
2.3.1 Frequency Selective Extrapolation with Residual Filtering for Image Error Concealment . . . 75
2.4 Improved filling order . . . 83

2.4.1 Sequential Error Concealment for Video/Images by Weighted Template Matching 83
2.4.2 An Error-based Recursive Filling Ordering for Image Error Concealment . . 95

3 Conclusions and Future Work **103**
3.1 Conclusions 103
3.2 Future work 105

Bibliography **107**

Chapter 1

Introduction to the PhD Thesis

1.1 Introduction

Recent advances in computing and communication technologies boosted the bandwidth expansion and the processing capabilities of personal computers and battery-powered terminals. These advances naturally yield a rapid growth of multimedia applications. In particular, video streaming (e.g. mobile TV, video-calling, etc.) and image transmission comprise nowadays by far the largest fraction of all the consumer traffic [1]. Thus, achieving high quality of service (QoS) for these signals is of utmost importance and it is a challenging task since multimedia streams are usually transmitted over error-prone channels such as the internet.

The majority of multimedia applications relies on state-of-the-art video codecs, such as H.264/AVC (Advanced Video Coding) or H.265/HEVC (High Efficiency Video Coding). Among the popular applications, we can mention video conferencing (e.g. Skype), video streaming services (Youtube, Vimeo, iTunes Store) or Bluray discs. In addition, high-definition television broadcasts over cable (DVB-C), satellite (DVB-S), handheld (DVB-H) and terrestrial (DVB-T) are also based on this type of codec [2]. These codecs are block-based and, in general terms, split the image/video signals into so called macroblocks which are coded using inter or intra prediction. Macroblocks within a frame can be split into several slices [3, 4]. A slice forms the payload of a network abstraction layer unit (NALU), which is a data sequence that can be decoded independently [5]. The loss of a NALU will therefore not affect other macroblocks within the current frame. However, due to temporal interframe prediction, error propagation may occur.

In many common multimedia applications the retransmission of the lost data is not possible due to real-time constraints (video-calling, sport events retransmissions, etc.) or lack of bandwidth. Although the aforementioned standards include several error resilience tools, such as arbitrary slice ordering (ASO) or flexible macroblock ordering (FMO) [6], error concealment (EC) techniques are required and even mandatory when packet losses occur. EC algorithms try to recover the lost signal at the decoder and without intervention of the encoder. In order to do this, these algorithms reconstruct the signal from correctly received data and other available information. EC algorithms can be classified into two categories:

1. Spatial EC (SEC) that relies on the information provided only by the current frame. It mainly involves surrounding pixels that have been received and correctly decoded.

2. Temporal EC (TEC) which utilizes temporal information such as motion vectors and previous or already available future frames.

TEC usually provides higher performance since temporal correlations in video signals tend to be higher than spatial ones. However, utilizing temporal information for the recovery of intracoded frames is not always possible, since these are inserted mainly to reset the prediction error when a change of scenes occurs. Thus, when the temporal information is unreliable or not available at all, SEC techniques are employed. Note that the intracoded frames (or I-frames) serve as prediction templates for intercoded frames. Thus, high quality reconstruction is desirable since any reconstruction error will be propagated until the next I-frame arrives and resets the prediction error.

Moreover, it is worth mentioning that SEC can be as well adapted to various image enhancement applications, such as inpainting, object removal, superresolution or texture diffusion.

EC techniques are of utmost importance for mobile video-streaming (especially video-calling) since video streams are in general highly compressed which favours error propagation. Applying accurate EC algorithms to real-time video applications will greatly improve the visual quality, especially during the hours of highly occupied bandwidth.

In this thesis we have studied the possibility of improving the reconstruction quality provided by the state-of-the-art techniques. We have designed and implemented several reconstruction algorithms applying different points of view. We have focused our attention on four main issues:

1. EC techniques based on **spatial interpolation**. These methods, in general, involve some type of edge detection and multiple interpolations can be combined in order to obtain the final reconstruction.

2. Algorithms that pursue the reconstruction by gathering **data statistics** from known surrounding samples.

3. EC techniques carried out in a **transformed domain**. These techniques generally involve an initial coarse estimation which is then iteratively refined.

4. High reconstruction quality is obtained if the missing region is divided into smaller areas and reconstructed sequentially. In such a case, the **filling order** is of utmost importance since error propagation is involved.

An exhaustive comparison with state-of-the-art algorithms is provided in order to asses the quality of our techniques. In the simulations, we have considered different loss scenarios and tested the performance over a large variety of images and video sequences. In fact, comparisons so extensive are rare in other work on EC. We present quality evaluations with different error patterns applied and values for each image as well as the overall average are also provided. In order to measure the reconstruction quality, the classical peak signal-to-noise ratio (PSNR) is utilized. Moreover, in order to better take into account the perceptual quality, the multiscale structural similarity (MS-SSIM) index is employed [7]. The MS-SSIM index consists in measuring the SSIM index for different image resolutions (obtained by low-pass filtering and subsampling). The SSIM index aims at approximating the human visual system (HVS) response by looking for similarities in luminance, contrast and structure [8]. It is worth mentioning that MS-SSIM is one of the metrics most highly correlated with subjective scores [9].

This memory is divided in three chapters and is organized as follows. In Chapter 1, after this introductory part, an overview of the four main research topics treated in this work is provided

in the next subsections (1.1.1-1.1.4). In Section 1.2 we address the open problems and describe the starting hypotheses that justify the elaboration of this work. The objectives of this thesis are detailed in Section 1.3. In Section 1.4, the proposed techniques are briefly summarized. Chapter 2 consists of the publications that deal with the proposed objectives. The last Chapter is devoted to conclusions and it outlines some important aspects to be taken into account in the future work.

Introducción

Los avances recientes en las tecnologías de la comunicación y la computación han propiciado la expansión del ancho de banda y las capacidades de procesamiento de ordenadores, tabletas y smartphones. Estos avances han supuesto un rápido crecimiento de las aplicaciones multimedia. En concreto, el streaming de vídeo (televisión móvil, videoconferencias, etc.) y la transmisión de imágenes constituyen hoy en día la mayor parte del tráfico de datos [1]. Como consecuencia, conseguir una alta calidad de servicio es de vital importancia puesto que las señales multimedia se suelen transmitir por redes no fiables como Internet.

La mayoría de las aplicaciones multimedia se basa en los codecs más avanzados, como el H.264/ AVC (Advanced Video Coding) o el H.265/HEVC (High Efficiency Video Coding). Entre las aplicaciones más populares, mencionar la videoconferencia (por ejemplo Skype), los servicios de streaming de vídeo (por ejemplo Youtube, Vimeo, iTunes Store) o los discos Bluray. Además, los sistemas de transmisión de televisión de alta definición por cable (DVB-C), satélite (DVB-S), móvil (DVB-H) y terrestre (DVB-T) también hacen uso de este tipo de codecs [2]. Estos codecs emplean la codificación por bloques. En términos generales, las imágenes/vídeo se descomponen en los llamados macrobloques que se codifican utilizando inter- ó intrapredicción. Los macrobloques de una trama se pueden agrupar en varios *slices* [3, 4]. Un *slice* forma el datagrama de la llamada NALU (Network Abstraction Layer Unit), que es una secuencia de datos que se puede decodificar de forma independiente [5]. Así, la pérdida de una NALU no afectará a otros macrobloques de la misma trama. No obstante, debido a la predicción temporal (entre tramas), la propagación del error es posible.

En muchas aplicaciones multimedia comunes la retransmisión de los datos perdidos no es posible debido a las restricciones de tiempo real (videollamadas, retransmisiones deportivas, etc.) o la falta de ancho de banda. Aunque los estándares, mencionados anteriormente, incluyen varias herramientas de protección frente a errores como ASO (Arbitrary Slice Ordering) ó FMO (Flexible Macroblock Ordering) [6], cuando se pierde un paquete es necesario emplear técnicas de mitigación de errores o EC (en inglés: *error concealment*). Los algoritmos EC taratan de recuperar la señal perdida utilizando los datos disponibles y se aplican en el decodificador, sin la intervención del codificador. Los algoritmos EC se pueden clasificar en dos categorías:

1. Técnicas EC espaciales (SEC, de *Spatial EC*) que se basan en la información contenida exclusivamente en el frame actual, es decir, en los píxeles adyacentes que se han recibido y decodificado correctamente.

2. Técnicas EC temporales (TEC, de *Temporal EC*) que hacen uso de la información temporal como los vectores de movimiento o las tramas pasadas y/o futuras.

TEC normalmente ofrece un mejor rendimiento puesto que las correlaciones temporales tienden a ser más fuertes que las espaciales. No obstante, utilizar la información temporal para la

reconstrucción de tramas intracodificadas no siempre es posible, dado que este tipo de tramas se inserta sobre todo para reiniciar el error de predicción cuando ocurren cambios de escenas. Como consecuencia, cuando la información temporal no es fiable o ni siquiera está disponible, se aplican las técnicas SEC. También notar que las tramas intracodificadas sirven como plantillas de predicción para las tramas intercodificadas. Por lo tanto, es deseable obtener reconstrucciones de alta calidad ya que los errores se propagarán hasta que una nueva trama intracodificada los reinicie.

Además, mencionemos que SEC puede adaptarse también a diferentes aplicaciones de realce de imágenes tales como inpainting, eliminación de objetos, superresolución o difusión de texturas.

Las técnicas EC son de importancia vital para videollamadas móviles ya que los flujos de datos están fuertemente comprimidos lo cual favorece la propagación de errores. Emplear técnicas EC de alta calidad para las aplicaciones de vídeo en tiempo real supone una gran mejora en cuanto a la calidad visual, especialmente durante las horas de alto uso del ancho de banda.

En esta tesis hemos estudiado la posibilidad de mejorar la calidad de reconstrucción de los algoritmos EC del estado de arte. Hemos diseñado e implementado varios algoritmos de reconstrucción aplicando diferentes puntos de vista. Nos hemos centrado en los siguientes cuatro aspectos:

1. Técnicas EC basadas en **interpolación** espacial. Estos métodos, en general, incluyen algún tipo de detección de fronteras y es posible combinar múltiples interpolaciones para obtener la reconstrucción final.

2. Algoritmos que consiguen la reconstrucción extrayendo la **estadística** a partir de las muestras adyacentes que son conocidas.

3. Técnicas EC llevadas a cabo en el **dominio transformado**. Estas técnicas, en general, parten de una estimación inicial de poca resolución que se refina iterativamente.

4. Reconstrucciones de alta calidad se pueden obtener si la región perdida se divide en áreas más pequeñas y se reconstruye secuencialmente. En este caso, el **orden de relleno** es muy importante puesto que tiene mucha influencia sobre la propagación de errores.

Se ha realizado una comparación exhaustiva con los algoritmos del estado de arte para poder evaluar la calidad de nuestras técnicas. En las simulaciones hemos asumido diferentes escenarios de pérdidas y hemos medido el rendimiento utilizando una gran variedad de imágenes y secuencias de vídeo. De hecho, unas comparaciones tan extensivas son escasas en otros trabajos. Se presentan evaluaciones de calidad utilizando diferentes patrones de error y se proporcionan valores para cada imagen así como el valor promedio. Para medir la calidad de reconstrucción, empleamos la medida clásica de PSNR (*peak signal-to-noise ratio*). Además, para tener en cuenta también la calidad perceptual, se emplea el índice MS-SSIM (*multiscale structural similarity index*) [7]. El índice MS-SSIM consiste en medir el índice SSIM para una imagen utilizando distintas resoluciones espaciales (obtenidas mediante un filtrado paso-baja y submuestreo). El índice SSIM trata de aproximar la respuesta del sistema visual humano mediante la búsqueda de similitudes en luminancia, constraste y estructura [8]. Cabe mencionar que MS-SSIM es una de las medidas más correladas con la evaluación sujetiva [9].

Esta memoria se divide en tres capítulos y se estructura de la siguiente forma. En el Capítulo 1, después de esta parte introductoria, se ofrece una descripción general de los cuatro aspectos mencionados anteriormente (subsecciones 1.1.1-1.1.4). En la Sección 1.2 identificamos los problemas

abiertos y describimos las hipótesis de partida que justifican la elaboración de este trabajo. Los objetivos de esta tesis se detallan en la Sección 1.3. Un breve resumen de las técnicas propuestas se halla en la Sección 1.4. El capítulo 2 consiste en las publicaciones que tratan de los objetivos propuestos. El último Capítulo está dedicado a conclusiones donde además se esbozan algunos aspectos importantes a tener en cuenta en el futuro.

1.1.1 Interpolation based EC

Various EC techniques have been already proposed for block-coded video/images. Many of them are based on some type of pixel interpolation, trying to exploit the correlations between adjacent pixels. In [10], a simple spatial interpolation is used. This technique aims at reconstructing each lost pixel by spatial interpolation from the four nearest undamaged pixels. This approach yields only moderate reconstruction quality but due to its computational inexpensiveness it is used as part of the non-normative error concealment algorithm set for H.264/AVC [11]. However, modern terminals are powerful enough to deal with much more complex EC algorithms without any significant time delay or surcharge in hardware usage that could excessively drain their batteries.

Since high frequencies, such as edges, are visually more relevant than uniform textures [12], more advanced interpolation techniques have been proposed exploiting directional features in the neighbourhood of the missing area. The authors in [13] combined the edge recovery and selective directional interpolation in order to achieve more visually pleasing reconstructions. However, the edge detection is quite inaccurate yielding errors when more complex edge structures (e.g. with a noisy background) are involved. A more robust edge detection based on a voting mechanism was proposed in [14]. A directional interpolation approach is applied here if there are only a few edges crossing the missing macroblock and a best-match approach is applied if the macroblock is decided to contain fine texture. For this algorithm, and in general for all switching EC techniques, a correct classification is crucial since an erroneous decision on the macroblock behaviour could have a very negative effect on the final reconstruction [15, 16]. In addition, trying to match the entire macroblock may generate artificial edges (the so called blocking). In [17], spatial direction vectors are introduced to obtain more accurate edge directions.

In order to determine which edges and under which angle enter the missing area, it is convenient to examine not only the pixels directly adjacent to the corrupt region but also a wider neighbourhood. Such an approach is treated in [18]. Sobel's operator along with an adaptive thresholding is applied in order to retain only significant edges. A weighted directional interpolation is applied afterwards. In [19], the Hough transform, a powerful tool for edge description, was used to set the angle for the directional interpolation. However, the performance drops when multiple edges need to be connected. A more robust approach is proposed in [20] permitting to connect several edges. This technique is highly adapted to consecutive block losses so it may suffer from a lack of generality.

A pixel-wise sequential recovery based on Wiener filtering was proposed in [21]. The error propagation is alleviated by a linear interpolation strategy. Another pixel-wise technique, that, after determining the direction of the edge which traverses the pixel to be recovered, extrapolates the missing pixel along the corresponding direction was introduced in [22]. However, as we will show later, pixel by pixel recovery suffers from smoothing high frequency textures.

1.1.2 Statistically driven EC

Interpolation based techniques are, in general, highly efficient. Problems occur when more complex edges and fine textures are involved. Although edges and borders are visually very important, there are other visual features that require more sophisticated description. In such cases, statistical driven approaches, that consider the correlation among pixels, are more suitable. The modelling of natural images as Markov random fields for EC was treated in [23]. This scheme produces relatively small squared reconstruction errors at the expense of an oversmoothed (i.e. blurred) reconstruction. The frames can be successfully modelled as AR processes, as in [24]. This work, however, assumes that (small) groups of macroblocks can be modelled using the same AR process which for low resolution videos or complex scenes may be inaccurate.

Bilateral filtering, exploiting a pair of Gaussian kernels, is treated in [25]. One kernel is employed to measure the similarity between the lost macroblock and the available surrounding area (range distance filter) and the other one is applied to take into account their spatial separation (domain distance filter). It has been shown that penalizing the spatial distance may be deteriorative [26] since the key issue is the selection of the appropriate variance for the kernels.

A patch-wise reconstruction technique based on sparse representation is treated in [27]. This algorithm, based on boundary matching, applies a fixed ℓ_0-norm sparsity level and the entries of the resulting sparse dictionary are combined using fixed weights. An improved sparsity based scheme is proposed in [28]. Here, a computationally expensive double optimization approach is involved.

Estimating a probability density function (pdf) from a given data set can also provide interesting results. High performance is achieved by reconstructing the unknown samples by minimum mean square error (MMSE) estimation [29]. The MMSE criterion can also be applied to reduce the error propagation while decoding a corrupt video sequence [30]. In [31], a Gaussian mixture model (GMM) is obtained from spatial and temporal surrounding information. This model, however, requires an extensive offline training. A computationally lighter version is described in [32].

1.1.3 EC in transformed domains

These techniques aim at recovering the missing samples by taking advantage of the fact that the transform basis are well suited for modelling visual features. Usually, the missing region and the surrounding available samples comprise the reconstruction area. Since the most popular transforms in image/video communication (DCT, Fourier, etc.) are block-based, recovering the reconstruction area will naturally yield the reconstruction of unknown samples.

An iterative technique for restoring the damaged areas, based on the method of projections onto convex sets, is developed in [33]. It tries to preserve borders, applying constraints on edge continuity and smoothness. It is a switching algorithm that depends on edge detection. An EC algorithm based on DCT coefficients recovery is presented in [34]. Smoothness constraint on image intensity is assumed. Another DCT based method is treated in [35]. The basic assumption is that, if the set of pixels consisting of the missing block and its border pixels is transformed by DCT, the high frequency coefficients obtained can be set to zero. Thus, a system of linear equations can be used to solve this problem although the final reconstruction may lack in important high frequency details.

A multiscale estimation approach with a DCT pyramid is treated in [36]. The missing blocks are recovered from low to high frequencies, starting from a low resolution image (low scale) and refining the details in later approximations. During the reconstruction, Gaussian weights with fixed bandwidth are used, regardless of the input data.

An alternative approach to image EC is the frequency selective extrapolation (FSE) proposed in [37]. In particular, the complex-valued FSE implementation [38] can provide high quality reconstructions with a low computational burden. This technique develops a signal model from the set of Fourier basis functions which can be used to replace the unknown samples. An improved version that deals with the orthogonality deficiency among windowed basis functions is described in [39]. In [40], a simplified and computationally much lighter version is presented.

1.1.4 Filling order

Previous work on image recovery has shown that applying sequential recovery yields better reconstruction quality [41, 42, 43]. The lost region is divided into smaller blocks that are estimated one by one following a certain order. The estimation, in turn, also relies on pixels that have been already reconstructed. In general, the filling order is crucial [42] since errors can be propagated throughout the lost area. Several filling orders have been proposed in the literature. Raster scan or concentric layer filling use a fixed filling order. An adaptive filling order based on external boundary matching criterion is treated in [44]. The method in [41] is based on the amount of the correctly received pixels around the missing block. The more pixels there are, the higher is the block priority. Although this recursive concealment performs considerably better than a single step recovery, it does not distinguish between correctly received pixels and already extrapolated pixels. This issue is dealt with in [42] by introducing a confidence parameter. This technique, however, is based on isophotes and therefore prioritizes linear structures which may lead to considerable error propagation. In [43], the confidence term is combined with a parameter based on fractional derivative. This approach prioritizes strong edges which serves well for inpainting purposes but may not be convenient for EC tasks. In fact, it is worth noticing that the majority of the work on filling order is related to inpainting and the reconstruction error is neglected.

1.2 Starting hypotheses

In this section, we outline the main drawbacks of the different approaches treated in Section 1.1 and identify some related issues which deserve a more in-depth research.

- **Interpolation based EC**: The most sophisticated interpolators extract visual features from the available surrounding area in order to find the right angle for the directional interpolation. In addition, interpolations can be combined in order to avoid artefacts. However, interpolation techniques that can be found in the literature perform rather modestly when dealing with more complex edge structures and textures. In order to improve the reconstruction quality, the following issues should be addressed:
    - Edges that enter the missing area need to be precisely described. In order to do so, a robust and accurate edge detection is crucial. Voting mechanisms based on Sobel's operator are inaccurate and highly affected by noise while Canny's edge detector may mask relevant edges due to the application of hysteresis.

  - In addition, not all the edges that enter the corrupt macroblock are equally relevant. The (visually) clearer the edge, the more it should affect the final reconstruction.

- **Statistically driven EC**: These techniques aim at extracting signal statistics from the available samples. In order to do so, several approaches are possible:

  - Multimedia signals can be well modelled using a sparse representation from a dictionary of prototypes. The problem formulation and the sparsity level are the key issues here. Moreover, although there are efficient algorithms to solve problems with sparsity constraints, the processing time is still too high for real-time applications given the multidimensionality of image/video signals.
  - For many EC techniques the estimation of a pdf from a given data set is the key issue. Kernel density estimation (KDE) techniques, in general, can capture local statistics more suitably than the GMM-based MMSE estimator. Nevertheless, a critical issue related to KDE is the estimation of a suitable kernel bandwidth. In spite of the extensive bibliography on bandwidth estimation, none of the techniques is oriented to multimedia signal reconstruction.

- **EC in a transformed domain**: These techniques develop a signal model from the set of basis functions (DCT, Fourier, etc.) which can be used to replace the unknown pixels.

  - The state-of-the-art algorithms do not take into account the low-pass behaviour of natural images. This leads to overfitting which can negatively affect the reconstruction quality.

- In addition, sequential reconstruction, if applicable, provides better results than a single step recovery. In this situation, the order in which the lost region is recursively filled will clearly condition the resulting reconstruction. The majority of the work on **filling order** is related to inpainting and the reconstruction error, which is the ultimate goal in EC, is neglected.

1.3 Objectives

The main goal of this thesis is to design and implement EC techniques that will outperform state-of-the-art algorithms in terms of reconstruction quality. In the following, we will specify the sub-objectives that permit us to accomplish the proposed challenge:

- Interpolation based EC

  - **To perform a robust edge analysis for directional interpolation**. A scanning procedure based on the Hough transform is developed in order to find the relevant edges and the visually clearest ones are employed in an interpolation based reconstruction.
  - **To design an interpolation scheme able to reconstruct more complex edge structures**. Specifically, we will combine several interpolations according to a set of weights. These weights are derived from the visual clearness associated to an edge and are unique for every pixel within the missing macroblock.

- Statistically driven EC

  - **To provide powerful tools for capturing local signal statistics**. These statistics will be then utilized to obtain high quality reconstructions. We will develop a sparse linear predictor able to recover even the finest textures. Moreover, we will adapt KDE to MMSE multimedia signal reconstruction.

- EC in transformed domain

  - **To improve the performance by considering the low-pass behaviour of natural images**. The goal is to find a suitable way to incorporate this information into existing EC scenarios.

- Filling order

  - **To design a novel filling order aimed to improve the performance of EC algorithms**. First, regions surrounded by a larger amount of available and reliable data should be prioritized. Moreover, in order to determine the filling order, the reconstruction quality of the already concealed blocks should be considered.

1.4 Thesis proposals

In this section we provide a summary of the different techniques proposed in this thesis along with a brief discussion about the obtained results.

1.4.1 Interpolation based EC: Spatial Error Concealment Based on Edge Visual Clearness for Image/Video Communication

In previous sections, we have addressed the issues related to interpolation-based concealment techniques. In order to obtain an accurate reconstruction, a reliable description of the edges affecting the lost area is required.

In this work, we propose an EC technique based on the concept of visual clearness of an edge. We explore the directional behaviour in the neighbourhood of the missing area by applying a novel scanning procedure.

- In order to estimate the angle of the directional interpolation, an edge detector must be applied first to provide a binary edge map. In most cases, Sobel's operator is utilized given its simple implementation [13, 14, 18, 20, 45]. The resulting edge map can be further thinned [20] but it still provides relatively low precision and is highly affected by noise. In our work, we use Canny's edge detector which is less sensible to noise as it first smooths the image by filtering it with a Gaussian kernel. Moreover, due to its non-maximum suppression feature the detected edges are clear and no additional thinning is required. However, extremely strong edges may mask other relevant ones when applying thresholding during the edge detection process. We solve this issue by applying a novel **scanning procedure** that isolates those edges, making the edge detection more robust. A Hough transform [19, 20] is then employed to obtain a suitable edge descriptor.

- The most advanced interpolation-based EC methods employ multiple directional interpolations which are combined by clustering [13] or linear combination [14]. These approaches, however, yield artefacts and oversmoothing. The proposed algorithm fixes this problem by

assuming that pixels lying in the proximity of the prolongation of the edge tend to be more influenced by the interpolation defined by the corresponding direction. We introduce the concept of **visual clearness** of an edge that allows a robust selection of relevant edges and is used to compute the pixel-wise weights that are used to combine the interpolations.

The results show that the proposed technique significantly outperforms other state-of-the-art interpolation based EC algorithms, both on objective and subjective levels. Better quality is achieved also with respect to other modern EC techniques, such as content adaptive EC [14] or bilateral filtering [25], among others.

The journal article associated to this part is:

- J. Koloda, V. Sánchez and A.M. Peinado, "Spatial Error Concealment Based on Edge Visual Clearness for Image/Video Communication", ***Circuits, Systems, and Signal Processing (Springer)***, vol. 32, pp. 815-824, April 2013.

1.4.2 Statistically driven EC

1.4.2.1 Multimedia Signal Reconstruction by Sparse Linear Prediction

In order to exploit the inner correlations among samples, a signal/image model is required. Several models can be found in the literature: Markov random fields [23, 46], fields of experts [47, 48], autorregressive processes [24, 49, 50], Gaussian mixture models [31, 51] or hidden Markov models [29, 52] The drawbacks of such approaches have been discussed in previous sections. In our work, the correlations are modelled and exploited by means of vector linear prediction (LP).

- Since natural images can be locally non-stationary, we design a **sparse LP** that dynamically adapts itself to the amount of useful and available data. In order to solve this sparsity constrained problem more effectively, we apply convex relaxation [53] that allows to employ efficient convex optimization tools [54].
    - This approach can be extended to speech signals. We propose a new variant of the least square autorregressive (LSAR) [55] method for speech reconstruction, which can estimate via least squares the segment of missing samples. The LP model of speech is assumed and a sparsity constraint on the AR coefficients is applied.
- A study on the distribution of the predictor coefficients is carried out which suggests a **fast exponential approximation** of the sparsely distributed LP coefficients.

Experimental comparisons reveal that our proposal yields higher quality reconstructions of complex structures and fine textures. The proposed method outperforms other state-of-the-art techniques both on objective and subjective levels.

The journal and conference papers associated to this part are:

- J. Koloda, J. Østergaard, S.H. Jensen, A.M. Peinado and V. Sánchez, "Sequential Error Concealment for Video/Images by Sparse Linear Prediction", ***IEEE Transaction on Multimedia***, vol. 4, pp. 957-969, June 2013.
- J. Koloda, A.M. Peinado and V. Sánchez, "Speech Reconstruction by Sparse Linear Prediction", *IberSPEECH*, selected for publication in *Communications in Computer and Information Science (Springer)*, pp. 247-256, Madrid, Spain, November 2012.

1.4.2.2 Multimedia Signal Reconstruction by Kernel-based MMSE

It can be noticed that the exponential approximation derived in the previous subsection consists in a Nadaraya-Watson regressor with a fixed bandwidth. We can generalize this regressor by adopting a **multivariate kernel-based MMSE estimation** framework. In fact, signal reconstruction can be viewed as a regression problem, where the regressor can be expressed as an expectation over a KDE-estimated probability density function [56]. However, the objectives of signal reconstruction are quite different from those of KDE or regression. The goal of reconstruction is the estimation of a specific group of samples rather than a global or even local fitting like in KDE or regression.

- As for any KDE problem, the main issue of kernel-based reconstruction is the estimation of a suitable bandwidth, which, under our multidimensional formalism, becomes a matrix. In spite of the extensive bibliography on bandwidth estimation (BE), there is no approach specifically oriented to multimedia signal reconstruction. We propose a novel **multivariate bandwidth estimation** method which is especially conceived for EC tasks.

Simulations reveal that the proposed technique achieves an average improvement of up to 1dB (in terms of PSNR) with respect to the classical plug-in BE [57]. The improvement is even larger with respect to a classical GMM-based MMSE reconstruction [31].

The articles associated to this part are:

- J. Koloda, A.M. Peinado and V. Sánchez, "On the Application of Multivariate Kernel Density Estimation to Image Error Concealment", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013.
- J. Koloda, A.M. Peinado and V. Sánchez, "Kernel-based MMSE Multimedia Signal Reconstruction and its Application to Spatial Error Concealment", ***IEEE Transactions on Multimedia***, accepted.

1.4.3 EC in a transformed domain: Frequency Selective Extrapolation with Residual Filtering

As mentioned in Section 1.1.3, various EC algorithms in a transformed domain have been proposed. One of the most efficient is frequency selective extrapolation (FSE) [37], with its further improvements regarding the fast complex-valued implementation [38] and orthogonality deficiency compensation [39, 40]. FSE is carried out from a parametric model based on two-dimensional basis functions. In our work, we will consider Fourier basis functions. FSE is an iterative procedure that, iteration by iteration, updates the parametric model by maximizing the decrease of the residual energy over the known neighbouring samples.

Natural images tend to be low-pass signals [58]. This is a priori knowledge not considered in the original FSE algorithm which could be incorporated into it in order to improve both reconstruction quality and robustness against overfitting. We propose a **frequency weighting (filtering)** to exploit this a priori knowledge. A special low-pass filter is designed for such a purpose.

Experimental results show that the proposed method improves the reconstruction quality by almost 1dB (in terms of PSNR) with negligible additional computational cost. This proposal also suppresses the performance decrease after a critical number of iterations has been achieved, making this technique much more robust against overfitting.

The article associated to this part is:

- J. Koloda, J. Seiler, A. Kaup, V. Sánchez and A.M. Peinado, "Frequency Selective Extrapolation with Residual Filtering for Image Error Concealment", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, May 2014.

1.4.4 Improved filling order

Dividing the missing area into smaller blocks yields high quality reconstructions [41, 42]. In this scenario, the filling order is of utmost importance since error propagation may occur. In order to deal with this issue, we introduce a **reliability parameter** associated to each subblock. The missing area is then concealed sequentially, prioritizing subblocks with higher reliabilities. When a subblock is concealed, the reliabilities are recalculated and the reconstruction evolves from the outer layer towards the centre of the missing area.

On the other hand, the majority of work on filling order is related to inpainting and object removal and the reconstruction error is neglected. We propose a novel filling order approach that, **exploiting the reconstruction error**, improves the quality reconstruction. Regions that yield better reconstructions will be prioritized in order to reduce error propagation and achieve better overall reconstruction quality.

The proposed method is applicable to a large variety of EC algorithms, providing improvements of up to 1dB with respect to other state-of-the-art filling techniques.

The articles associated to this part are:

- J. Koloda, J. Østergaard, S.H. Jensen, A.M. Peinado and V. Sánchez, "Sequential Error Concealment for Video/Images by Weighted Template Matching", *IEEE Data Compression Conference (DCC)*, Snowbird, Utah (USA), April 2012.
- J. Koloda, J. Seiler, A. Kaup, V. Sánchez and A.M. Peinado, "An Error-based Recursive Filling Ordering for Image Error Concealment", *IEEE Internationl Conference on Image Processing (ICIP)*, Paris, France, October 2014.

Chapter 2

Publications: Published and Accepted Papers

2.1 Interpolation based EC

The journal paper associated to this part is:

2.1.1 Spatial Error Concealment Based on Edge Visual Clearness for Image/Video Communication

- J. Koloda, V. Sánchez and A.M. Peinado, "Spatial Error Concealment Based on Edge Visual Clearness for Image/Video Communication", *Circuits, Systems, and Signal Processing (Springer)*, vol. 32, pp. 815-824, April 2013.
    - Status: Published.
    - Impact Factor (JCR 2012): 0.982
    - Subject Category: Engineering, Electrical & Electronic. Ranking 136/243 (Q3).

Circuits Syst Signal Process (2013) 32:815–824
DOI 10.1007/s00034-012-9504-0

Spatial Error Concealment Based on Edge Visual Clearness for Image/Video Communication

Ján Koloda · Victoria Sánchez · Antonio M. Peinado

Received: 24 February 2012 / Revised: 20 September 2012 / Published online: 12 October 2012

Abstract In this paper, we propose a technique for concealing missing image/video blocks based on the concept of visual clearness of an edge. A scanning procedure based on the Hough transform allows us to find the relevant edges, and the visually clearest ones are employed in an interpolation based reconstruction. Specifically, several interpolations are combined according to a set of weights which allows the reconstruction of more complex textures. These weights are derived from the visual clearness associated to an edge and are unique for every pixel within the missing macroblock. The resulting algorithm is quite efficient, simple, and competitive in comparison with other state-of-the-art techniques.

Keywords Error concealment · Directional interpolation · Block-coded image/video · Hough transform

1 Introduction

Multimedia transmission applications are prone to suffer from deterioration of QoS. Due to strict real-time requirements, the retransmission of lost or severely damaged packages can be impossible. The block-based video coding standard H.264/AVC has introduced several error resilience mechanisms that draw on specific data organization tools such as network abstraction layer units (NALU), flexible macroblock order-

J. Koloda (✉) · V. Sánchez · A.M. Peinado
Department of Signal Theory, Networking and Communications and CITIC-UGR, Universidad de Granada, 18071 Granada, Spain
e-mail: janko@ugr.es

V. Sánchez
e-mail: victoria@ugr.es

A.M. Peinado
e-mail: amp@ugr.es

ing (FMO), or arbitrary slice ordering (ASO) [4, 17]. These tools allow the decoder to apply error concealment (EC) algorithms [8], in order to achieve an acceptable visual quality of the received stream.

The EC algorithms benefit from the fact that video signals are highly correlated, spatially and/or temporally. This criterion is used to classify the EC algorithms into two groups: spatial EC (SEC), which utilizes only the information provided by the current frame, and temporal EC (TEC), which makes use of temporal information such as motion vectors (MV). This classification is nonexcluding, and combining temporal and spatial information leads to significant improvements [6]. However, the most extended block-based coding standards, such as H.264 and MPEG-4, use both intracoding (I-frames) and prediction (P/B-frames). Since the intracoded frames serve as a "firewall", that separates visually different scenes or resets the prediction error, utilizing temporal information for their concealment could be risky. Therefore, SEC techniques are the most suitable choice to conceal the I-frames [17]. Moreover, these frames are employed as templates to predict several consecutive P/B-frames [4], so a poor reconstruction of one I-frame would distort not only the frame itself but the sequence of several frames in a row.

A simple technique for spatial concealment is bilinear interpolation [11]. Since high-frequency features, such as edges, are visually more relevant than uniform textures, more advanced interpolation techniques have been proposed exploiting directional features in the neighborhood of the missing macroblock [5]. A reconstruction of broken edges in the transformed domain is treated in [13]. In [9], the Hough transform, a powerful tool for edge description, was used to set the angle for the interpolation. However, the performance drops when multiple edges need to be connected. A more robust approach is suggested in [2] permitting one to connect several edges. Nevertheless, the technique is highly adapted to consecutive block loss, and so it may suffer from a lack of generality. Moreover, the applied interpolation process is rather simple and is unable to restore more complicated edges and textures. Restoration of broken edges based on extrapolation is introduced in [19]. A block-matching technique with a decision algorithm is treated in [10]. The algorithm restores successfully fine textures although correct classification of the missing macroblock is crucial. In addition, trying to match the entire macroblock may generate artificial edges (the so-called blocking). This is partially solved by utilizing block-based bilateral filtering [18]. Inpainting-based techniques are also used for quality texture reconstruction [3]. Modeling an image as a Markov random field allowed the authors in [12] to implement an efficient concealment algorithm. However, the reconstruction of high-frequency features tends to oversmoothing.

In this paper, the concept of visual clearness associated to an edge is introduced, and the most visually relevant edges are utilized for concealing missing image/video blocks based on a weighted combination of directional interpolations. Furthermore, we consider that pixels in the corrupted macroblock are not equally affected by the interpolations, and so pixel-dependent weights need to be assigned when combining these interpolations. The resulting technique is able to restore complicated edges and more complex textures, thus providing high-quality reconstructions.

The paper is organized as follows. In Sect. 2, the concept of visual clearness associated to an edge is introduced. Reconstruction based on the visually clearest direc-

tions is described in Sect. 3, and simulation results and comparisons with other SEC techniques are presented in Sect. 4. The last section is devoted to conclusions.

2 Visual Clearness Associated to an Edge

In this section, we establish a criterion to assess the importance of an edge. This criterion is based on the visual clearness of a given edge. At the end of the section, we will be able to introduce the parameter representing this visual clearness.

In order to obtain the relevant directions, it is necessary to explore first the directional behavior in the neighborhood of the missing macroblock. Thus, an edge detection must be applied first to provide a binary edge map. For this purpose, we use one of the most simple and efficient procedures, the Canny's edge detector [1]. In comparison with other detectors, such as Sobel's one, it is less sensible to noise as it first smooths the image by filtering it with a Gaussian kernel. Moreover, due to its nonmaximum suppression feature, the detected edges are clear, and no thinning algorithm needs to be applied.

Next, the predominant directions need to be computed. For this purpose, a gradient-based voting mechanism [10] has been widely used. However, the angular resolution of such a technique is rather poor, and edges with the same directions but different spatial location are treated as a single line. Instead, a Hough transform-based procedure is applied.

The Hough transform provides simple yet powerful descriptors and is based on the fact that many shapes can be expressed in a parametric form. In this paper, for the sake of simplicity, we use a linear kernel that assumes that any line can be expressed as

$$\rho = x\cos\theta + y\sin\theta, \tag{1}$$

where ρ is the perpendicular distance between the line and the origin, and θ is the slope of the normal. The Hough transform is applied to the binary image provided by the edge detector. Thus, Eq. (1) involves that every pixel (x, y) that belongs to a linear edge produces the same (ρ, θ). Therefore, the set of pixels that comprise a linear segment are transformed into a single point with position (within the transformed matrix) indicated by the parameters ρ (row) and θ (column). Thus, the bidimensional spatial domain (x, y) is transformed into a new domain (ρ, θ) where the transform value at each point (Hough coefficient, $H(\rho, \theta)$) is directly related to the number of pixels contained in the segment defined by (ρ, θ).

Directional interpolation is based on the directional behavior of the missing macroblock. Given the high spatial correlation, this behavior can be deduced from that of its neighborhood. To explore it, a scanning is carried out as shown in Fig. 1. The proposed scanning consists in moving a mask with the same dimensions as the lost macroblock, pixel by pixel (scanning step of 1 pixel) along its four sides: top (Fig. 1(b)), left, bottom, and right. At each step (or mask position), the Canny's edge detector is applied, and the Hough transform is calculated over the corresponding binary image in order to find relevant edges. An edge is said to be relevant if its prolongation crosses the missing macroblock. The direction of every relevant edge, θ, is stored in

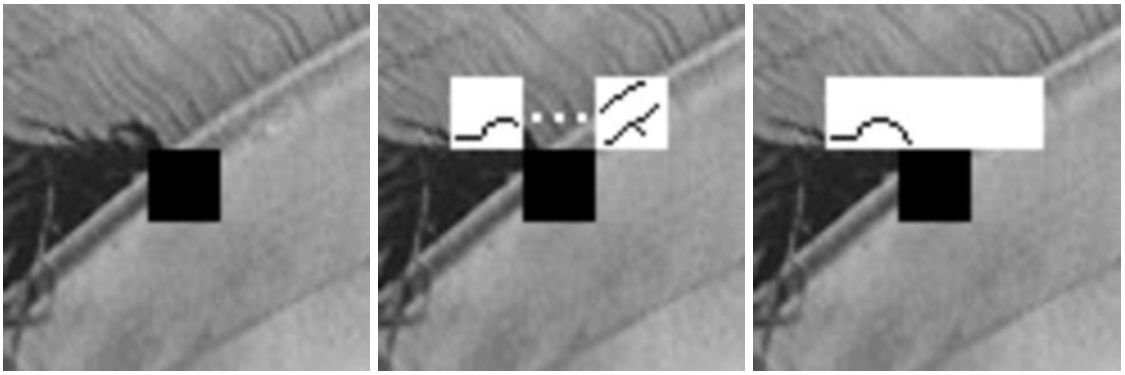


Fig. 1 Edge detection. (**a**) Received image with a missing macroblock, (**b**) top side scanning, (**c**) top side edge detection without scanning

a set $\mathcal{D}$. Note that if the available neighborhood were treated as a whole (as in [2]), extremely strong edges would mask other relevant ones when applying thresholding in the edge detection process (Fig. 1(c)). The scanning procedure isolates those edges, making the edge detection more robust.

The candidates for relevant edges are found as follows. First, the strongest direction in every mask position is selected from the transformation matrix (the strength of a direction is expressed by the Hough coefficient value). Then, using the information provided by the Hough coefficient coordinates, θ (slope) and ρ (offset), we can determine whether the edge crosses the corrupt macroblock. If it does, the edge is considered relevant and stored in $\mathcal{D}$, otherwise the direction is discarded, and the next strongest one is examined. If none of the directions satisfies the aforementioned condition, no direction is stored.

Due to the image resolution, perfectly straight lines might not lead to a single pair (ρ, θ). Moreover, curved edges are treated as a set of linear segments with similar (ρ, θ). Using a coarser resolution of the Hough transform, these (ρ, θ) pairs lead to a single value of H allowing the edge to be treated as a single line. However, too coarse resolutions introduce significant imprecision and should be avoided. In our simulations, pixel-by-pixel resolution is applied for ρ and steps of 2^o for θ.

Finally, we introduce the concept of visual clearness, σ_i, associated to an edge. In order to define the visual clearness σ_i of an edge i, let $\boldsymbol{E}_i$ be the set of all pixels that comprise the edge. We will then compute σ_i as the product of the strength of the edge direction and the visual separation between the regions at both of its sides. The strength of an edge is proportional to its associated Hough coefficient value (which is proportional to the edge length), and the visual separation is given by the average 2D spatial gradient per edge pixel,

$$\sigma_i = H_i \frac{1}{|\boldsymbol{E}_i|} \sum_{j \in \boldsymbol{E}_i}^{|\boldsymbol{E}_i|} \sqrt{dx_j^2 + dy_j^2}, \quad i = 1, \ldots, |\mathcal{D}|, \tag{2}$$

where H_i is the associated Hough coefficient, dx_j and dy_j are the horizontal and vertical gradients centred over the jth pixel of the edge, and $|\mathcal{D}|$ is the number of relevant directions found by the scanning procedure.

3 Reconstruction Based on the Visually Clearest Directions

Among the directions stored in the set $\mathcal{D}$, the N visually clearest ones are selected. That is, those edges with the N largest σ_i values are selected. The missing macroblock is reconstructed by combining directional interpolations based on these N selected directions. Previously, the N corresponding interpolations are computed as

$$I_i(x,y) = \frac{d_2}{d_1+d_2} p_1^{(i)} + \frac{d_1}{d_1+d_2} p_2^{(i)}, \quad i = 1,\ldots,N, \tag{3}$$

where the missing pixel $p(x,y)$ is replaced by a weighted mean of $p_1^{(i)}$ and $p_2^{(i)}$, the two closest pixels in the ith direction that have been correctly received and decoded. The variable d_1 (d_2) is the Euclidean distance between p and $p_1^{(i)}$ ($p_2^{(i)}$).

Clear lines tend to be visually more important, so the N interpolations are combined according to the visual clearness of their corresponding edges. Therefore, every interpolation I_i has an associated weight defined as

$$w_i = \frac{\sigma_i}{\sum_{j=1}^{N} \sigma_j}. \tag{4}$$

Scalar weights are widely used in state-of-the-art techniques that involve combination of interpolations [10]. In more complex environments where several directions are present, this approach leads to oversmoothing. The proposed algorithm fixes this problem by assuming that the interpolations are not equally relevant for every pixel of the missing macroblock. In fact, pixels lying in the proximity of the prolongation of an edge tend to be more influenced by the interpolation defined by the corresponding direction. Therefore, every pixel has an associated weight computed as

$$\pi_i(x,y) = 1 - \delta_i^2(x,y), \tag{5}$$

where $\delta_i(x,y)$ is the normalized distance between the pixel $p(x,y)$ and the line defined by the ith edge. The normalization factor corresponds to the maximum distance between a pixel and a line within a macroblock, that is, to the length of its diagonal. The use of square power has been set heuristically since it provides better results.

Finally, the corrupt macroblock is reconstructed by means of a weighted superposition of N directional interpolations. That is, for every pixel $p(x,y)$ of the missing macroblock, we apply

$$p(x,y) = \sum_{i=1}^{N} \frac{w_i \pi_i(x,y)}{\sum_{j=1}^{N} w_j \pi_j(x,y)} I_i(x,y). \tag{6}$$

Some of the state-of-the-art algorithms, such as [2, 5], divide the available neighborhood into support regions. Thus, a pixel can be only interpolated relying on pixels within the same support region. This hard division, however, may create artificial borders and false textures. Weights $\pi_i(x,y)$ smooth the transitions from one region to another, preserving the continuity of the image signal.

Finally, it should be noted that N is the maximum number of directions to be used and is, by no means, mandatory. If there are fewer directions available or they are not visually clear enough, the number of interpolations to be considered will be less than N. The algorithm thus automatically adapts itself to the type of texture surrounding the missing macroblock.

In summary, for each missing macroblock, the proposed algorithm can be resumed as follows:

Step 1: perform the scanning at each side of the corrupt macroblock in order to obtain the set of relevant directions $\mathcal{D}$.

Step 2: compute the visual clearness for every relevant direction.

Step 3: select the N visually clearest directions and compute I_i $(i = 1, \ldots, N)$ as shown in Eq. (3).

Step 4: calculate the corresponding weighting factors ω_i and $\pi_i(x, y)$ for $i = 1, \ldots, N$ and for every missing pixel $p(x, y)$.

Step 5: combine the N directional interpolations corresponding to the N visually clearest directions by applying Eq. (6).

4 Simulation Results

The performance of the proposed algorithm is tested over the images of *Lena* (512×512), *Pirate* (1024×1024), the first frame of *Foreman* sequence (288×352), *Office* (592×896), *Airplane* (512×512), *Zelda* (512×512), *Boat* (512×512), and *House* (256×256). The test is carried out for macroblock dimensions of 16×16, and the rate of block loss is approximately 25 %, corresponding to a single packet loss of a frame with dispersed slicing structure [4]. The reconstruction quality of the proposed algorithm is compared with others, such as the bilinear interpolation (BIL) [11], projections onto convex sets (POC) [13], SEC based on the Hough transform (SHT) [2], content adaptive SEC (CAD) [10], directional extrapolation (EXT) [19], nonnormative SEC for H.264/AVC codec (AVC) [14], an adaptive Markov random fields method (MRF) [12], inpainting (INP) [3], and bilateral filtering [18].[1] Our proposal was tested for $N = 2$, $N = 3$, $N = 4$, and $N = 5$, where N is the number of interpolations combined according to Eq. (6). In addition, larger scanning steps of 2 pixels (N_2), 4 pixels (N_4), and 8 pixels (N_8) are also tested for $N = 5$.

In order to better take into account the perceptual quality, the multiscale structural similarity (MS-SSIM) index [16] is used for comparison along with the objective PSNR measure. Regarding MS-SSIM, the image is sequentially low-pass filtered and subsampled, and so a set of images is obtained, including the original resolution. Then, the SSIM index is applied for every subimage within the set. The SSIM index aims at approximating the human visual system (HVS) response looking for similarities in luminance, contrast, and structure [15]. This index can be seen as a convolution

[1] Implementations of most of these techniques, as well as the implementation of our algorithm, is available online at [20].

Table 1 PSNR values (in dB) and MS-SSIM values (scaled by 100) of test images reconstructed by several algorithms. The best performances for each image are shown in *boldface*

PSNR MS-SSIM	*Lena*	*Pirate*	*Foreman*	*Office*	*Airplane*	*Zelda*	*Boat*	*House*
BIL	30.00	27.82	27.12	27.54	25.58	33.44	26.95	26.83
	96.82	94.90	95.36	94.12	93.49	97.12	93.12	94.39
POC	28.04	26.42	28.49	27.56	26.18	29.91	26.05	27.38
	93.17	91.85	93.82	93.62	93.87	92.51	92.63	92.22
SHT	30.55	28.12	28.09	27.58	25.62	33.45	27.20	28.41
	97.13	95.18	95.94	94.09	93.41	97.22	93.21	96.20
CAD	31.96	28.44	34.85	29.29	27.41	34.01	27.73	31.11
	97.38	95.38	98.30	96.12	95.70	97.74	94.29	97.37
EXT	29.10	27.57	29.59	27.72	26.23	32.14	26.63	27.38
	95.83	94.89	97.18	94.84	95.36	96.73	94.01	93.98
AVC	30.42	28.74	29.11	29.99	27.79	34.43	28.25	28.37
	96.72	95.62	97.06	96.07	95.90	97.63	94.95	94.92
MRF	32.17	29.52	32.99	29.77	27.98	35.03	27.91	30.08
	97.75	96.33	98.21	96.45	96.12	98.03	94.92	96.89
INP	30.85	28.44	34.44	29.65	26.29	33.62	27.79	29.90
	97.17	95.27	98.35	96.66	94.91	97.44	95.22	96.97
BLF	32.15	29.36	34.75	30.06	28.34	33.83	28.37	30.52
	97.52	95.99	98.27	96.30	96.59	97.25	95.55	97.05
$N = 2$	32.46	29.60	34.93	30.98	28.51	35.07	28.44	31.30
	98.00	96.55	98.37	97.15	96.72	98.26	95.82	97.46
$N = 3$	32.70	29.90	35.09	31.34	28.71	35.44	28.66	31.45
	98.04	96.71	98.44	97.35	96.77	98.36	95.99	97.45
$N = 4$	32.74	29.98	35.15	31.44	28.76	35.67	28.77	31.47
	98.06	96.77	98.49	97.37	96.82	98.44	96.11	97.47
$N = 5$	**32.80**	**30.03**	**35.21**	**31.52**	**28.80**	**35.75**	**28.84**	**31.48**
	98.09	**96.80**	**98.53**	**97.43**	**96.82**	**98.46**	**96.11**	**97.50**
$N_2 = 5$	32.58	29.98	35.00	31.50	28.77	35.72	28.81	31.46
	97.91	96.53	98.40	97.39	96.66	98.36	96.03	97.36
$N_4 = 5$	32.53	29.96	34.97	31.49	28.74	35.61	28.73	31.46
	97.85	96.51	98.37	97.31	96.60	98.30	95.94	97.37
$N_8 = 5$	32.28	29.76	34.64	31.47	28.62	35.36	28.50	31.45
	97.66	96.48	98.22	97.33	96.64	98.24	95.69	97.36

of a fixed-sized mask with the residual error between the reference image and the concealed image [7]. A unique mask size is used for each of the images within the testing set. Thus, both fine and coarse textures and objects are taken into account.

The results in Table 1 show that the proposed algorithm outperforms the others for all the tested images both in terms of PSNR and MS-SSIM. Note that the algorithm performance saturates as N increases. In many cases, two directions are sufficient for

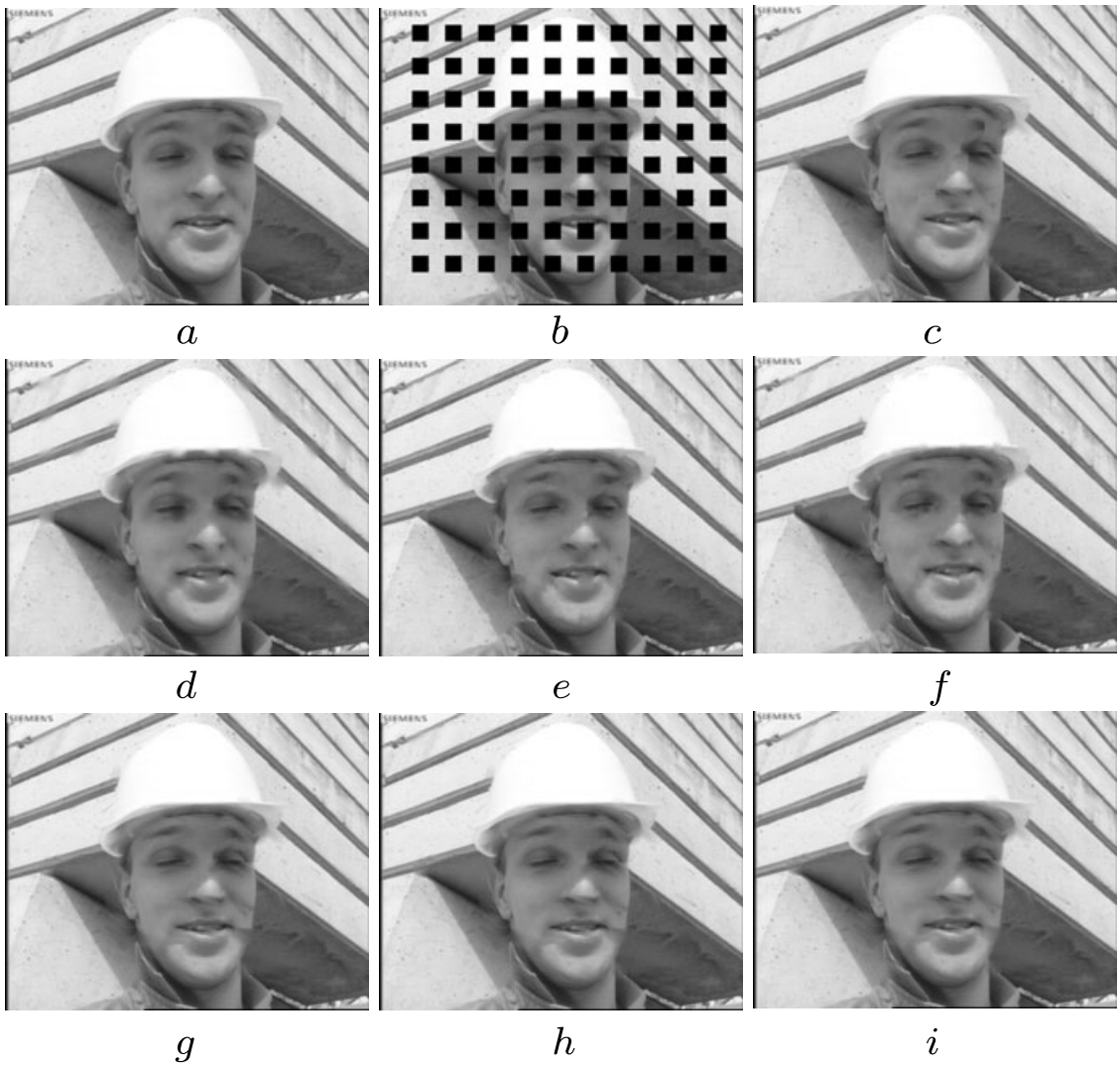


Fig. 2 Subjective comparison of different algorithms for 16×16 pixels macroblocks. (**a**) Original image, (**b**) corrupted image, (**c**) reconstructed image by CAD, (**d**) MRF, (**e**) INP, (**f**) BLF, (**g**) proposed algorithm with $N = 2$, (**h**) $N = 3$, and (**i**) $N = 4$

a good reconstruction, and utilizing more directions might not achieve any considerable improvement. In fact, Table 1 shows that even the simple reconstruction with two directions provides better reconstruction quality than all the other state-of-the-art techniques listed in the table. However, as a general rule, the more complex the texture, the more directions should be considered.

In order to better illustrate the subjective quality, Fig. 2 shows a comparison of the methods that provide the highest perceptual quality reconstructions (highest MS-SSIM) in Table 1. We see that the superiority of our algorithm, in terms of PSNR, is also corroborated at the subjective level.

Finally, the simulations reveal that the scanning procedure (including edge detection, Hough transform, and computing the visual clearness) comprises up to 90 % of the overall computational load. Using pixel-by-pixel scanning is relatively computationally expensive; however, in many cases, coarser resolutions would achieve almost identical results. Increasing the scanning step to 4 pixels, the processing time becomes similar or even lower than the processing time of more complex tested algo-

rithms such as CAD, INP, or BLF. By applying 4-pixel scanning step the reconstruction quality is reduced in roughly 0.1 dB on average (in comparison to pixel-by-pixel scanning) and still does outperform the other techniques for all the tested images. Moreover, note that if we continue increasing the scanning step, the computational load can be reduced even further with relatively moderate effect on the reconstruction quality (see N_8 in Table 1).

5 Conclusions

Block-based coding standards, such as H.264 and MPEG-4, are widely spread in video transmission over packet based networks. However, the packetized stream suffers from packet losses, and so some of the macroblocks are not received or decoded properly. Moreover, as the aforementioned standards use interframe prediction, a single packet loss could cause an error propagation, distorting the whole video sequence. In this paper, we have developed an error concealment scheme that utilizes only the spatial information of the current frame, making it specially suitable for reconstruction of I-blocks. The concept of visual clearness of an edge is introduced in order to find a suitable set of weights that are used in a pixel-level weighted combination of interpolations that allows one to reconstruct even nonlinear features more accurately. The proposed algorithm shows a significant improvement while keeping a relatively moderate computational complexity.

Regarding future work, the reconstruction of more complex visual features by combining edge and texture reconstruction is a goal worth exploring.

Acknowledgements This work has been supported by the Spanish MEC/FEDER project TEC2010-18009.

References

1. J. Canny, A computational approach to edge detection. IEEE Trans. Pattern Anal. Mach. Intell. **8**, 679–698 (1986)
2. H. Gharavi, S. Gao, Spatial interpolation algorithm for error concealment, in *Proceedings of ICASSP*, (2008), pp. 1153–1156
3. P.F. Harrison, Texture synthesis, texture transfer and plausible restoration. Ph.D. Thesis, Monash University (2005)
4. ITU-T, ITU-T Recommendation H.264, International Telecommunication Union (2005)
5. W.Y. Kung, C.S. Kim, C.C.J. Kuo, Spatial and temporal error concealment techniques for video transmission over noisy channels. IEEE Trans. Circuits Syst. Video Technol. **16**, 789–802 (2006)
6. M. Ma, O.C. Au, S.H. Gary Chan, M.T. Sun, Edge-Directed error concealment. IEEE Trans. Circuits Syst. Video Technol. **20**, 382–394 (2010)
7. J. Østergaard, M.S. Derpich, S.S. Channappayya, The high-resolution rate-distortion function under the structural similarity index. EURASIP J. Adv. Signal Process. (2011)
8. A.M. Peinado, A.M. Gómez, V. Sánchez, Error concealment based on MMSE estimation for multimedia wireless and IP applications, in *Proceedings of PIMRC*, (2008), pp. 1–5, (invited paper)
9. D.L. Robie, R.M. Mersereau, The use of hough transforms in spatial error concealment, in *Proceedings of ICASSP*, vol. 4, (2000), pp. 2131–2134
10. Z. Rongfu, Z. Yuanhua, H. Xiaodong, Content-adaptive spatial error concealment for video communication. IEEE Trans. Consum. Electron. **50**, 335–341 (2004)

11. P. Salama, N.B. Shroff, E.J. Coyle, E.J. Delp, Error concealment techniques for encoded video streams, in *Proceedings of ICIP*, (1995), pp. 9–12
12. S. Shirani, F. Kossentini, R. Ward, An adaptive Markov random field based error concealment method for video communication in error prone environment, in *Proceedings of ICIP*, vol. 6, (1999), pp. 3117–3120
13. H. Sun, W. Kwok, Concealment of damaged block transform coded images using projections onto convex sets. IEEE Trans. Image Process. **4** (1995)
14. V. Varsa, M.M. Hannuksela, Non-normative error concealment algorithms, in *ITU-T SG16, VCEG-N62*, vol. 50 (2001)
15. Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assesment: from error visibility to structural visibility. IEEE Trans. Image Process. **13**, 600–612 (2004)
16. Z. Wang, E.P. Simoncelli, A.C. Bovik, Multi-scale structural similarity for image quality assessment. IEEE Signal. Syst. Comput. **2**, 1398–1402 (2003)
17. T. Wiegand, G.J. Sullivan, G. Bjontegaard, A. Luthra, Overview of the H.264/AVC video coding standard. IEEE Trans. Circuits 560–576 (2003)
18. G. Zhai, J. Cai, W. Lin, X. Yang, W. Zhang, Image error-concealment via block-based bilateral filtering, in *IEEE International Conference on Multimedia and Expo*, (2008), pp. 621–624
19. Y. Zhao, H. Chen, X. Chi, J.S. Jin, Spatial error concealment using directional extrapolation, in *Proceedings of DICTA*, (2005), pp. 278–283
20. http://dtstc.ugr.es/~jkoloda/download.html. Available online

2.2 Statistically driven EC

The papers associated to this part are:

2.2.1 Multimedia Signal Reconstruction by Sparse Linear Prediction

2.2.1.1 Sequential Error Concealment for Video/Images by Sparse Linear Prediction

- J. Koloda, J. Østergaard, S.H. Jensen, A.M. Peinado and V. Sánchez, "Sequential Error Concealment for Video/Images by Sparse Linear Prediction", *IEEE Transactions on Multimedia*, vol. 4, pp. 957-969, June 2013.
    - Status: Published.
    - Impact Factor (JCR 2012): 1.754
    - Subject Category: Computer Science, Information Systems. Ranking 24/132 (Q1).
    - Subject Category: Computer Science, Software Engineering. Ranking 15/105 (Q1).
    - Subject Category: Telecommunications. Ranking 14/78 (Q1).

Sequential Error Concealment for Video/Images by Sparse Linear Prediction

Ján Koloda, Jan Østergaard, *Senior Member, IEEE*, Søren H. Jensen, *Senior Member, IEEE*, Victoria Sánchez, *Member, IEEE*, and Antonio M. Peinado, *Senior Member, IEEE*

***Abstract*—In this paper, we propose a novel sequential error concealment algorithm for video and images based on sparse linear prediction. Block-based coding schemes in packet loss environments are considered. Images are modelled by means of linear prediction, and missing macroblocks are sequentially reconstructed using the available groups of pixels. The optimal predictor coefficients are computed by applying a missing data regression imputation procedure with a sparsity constraint. Moreover, an efficient procedure for the computation of these coefficients based on an exponential approximation is also proposed. Both techniques provide high-quality reconstructions and outperform the state-of-the-art algorithms both in terms of PSNR and MS-SSIM.**

***Index Terms*—Block-coded images/video, convex optimization, error concealment, missing data imputation, sparse representation.**

I. Introduction

BLOCK-BASED video coding standards, such as MPEG-4 or H.264/AVC, are widely used in multimedia applications. Video signals are split into macroblocks that are coded using inter-or intraframe prediction. Quantization is carried out in the DCT domain and lossless arithmetic compression is applied [1]. This leads to low distortions at moderate bit-rates. However, achieving high quality reception is a challenging task since data streams are usually transmitted over error-prone channels.

For real-time transmission applications, the H.264/AVC standard has introduced several error resilience tools, such as arbitrary slice order (ASO) and flexible macroblock ordering (FMO) [2]. Macroblocks within a frame can be split into several slices. A slice forms the payload of a network abstraction layer unit (NALU), which is a data sequence that can be decoded independently [1]. The loss of a NALU will therefore not affect other macroblocks within the current frame. However, due to temporal interframe prediction, error propagation does occur.

H.264/AVC allows both bit-and packet-oriented delivery. For bit-oriented transmissions, an error burst that surpasses the channel-coding protection may result in loss of synchronization as well as fatal data damage since H.264/AVC utilizes variable length coding (VLC) or Exponential-Golomb coding for lossless compression [3]. Errors would thus propagate throughout the packet, making the current slice unusable. In packet oriented delivery, damaged packets, containing NALUs, are usually detected and discarded by network or transmission layers. Also, there may be packets which are not received at all due to congestion, routing problems, etc. In both cases, we are facing the problem of the loss of, at least, one slice.

Error concealment (EC) techniques form a very challenging field, since QoS is of utmost importance for the users. In many cases, retransmission of lost data is not possible due to real-time constraints or lack of bandwidth. This last case also applies to additional transmission of media-specific forward error correction (FEC) codes which, in addition, may not be standard compliant [4]. In contrast to channel coding techniques, which are carried out at the encoder and are designed to minimize the negative impact of packet losses, EC is applied at the decoder and can significantly improve the quality of the received stream [5]. EC algorithms can be classified into two categories: spatial EC (SEC), which relies on the information provided within the current frame and temporal EC (TEC), which utilizes temporal information such as motion vectors (MV) and previous or already available future frames. Some TEC techniques use both temporal and spatial information for image restoration and they are often referred to as combined or hybrid SEC/TEC algorithms. Both categories, SEC and TEC, exploit the redundancy due to the high spatial and temporal correlation within a video sequence. Temporal correlations tend to be higher than the spatial ones, so TEC techniques usually provide better results. This would be the straightforward choice when concealing a P/B-frame (intercoded). However, utilizing temporal information for the recovery of I-frames (intracoded) is not always possible, since they may be inserted to reset the prediction error when a change of scene occurs. Thus, when all the available temporal information belongs to a different scene or there is no temporal information available, SEC algorithms are necessary. Every I/P-frame in the video sequence usually serves as a prediction template for, at least, one intercoded frame. Thus, high quality concealment is required since any reconstruction error will be propagated until the next I-frame arrives and resets the prediction error.

Manuscript received April 17, 2012; revised July 23, 2012; accepted September 24, 2012. Date of publication January 09, 2013; date of current version May 13, 2013. This work was supported by the Spanish MEC/FEDER Project TEC 2010-18009. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Eckehard Steinbach.

J. Koloda, V. Sánchez, and A. M. Peinado are with the Department of Signal Theory, Networking and Communications, University of Granada, Granada 18011, Spain (e-mail: janko@ugr.es; victoria@ugr.es; amp@ugr.es).

J. Østergaard and S. H. Jensen are with the Department of Electronic Systems, Aalborg University, Aalborg 9220, Denmark (e-mail: jo@es.aau.dk; shj@es.aau.dk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2013.2238524

Several SEC techniques have been proposed for block-coded video/images. Many of them are based on some type of interpolation, trying to exploit the correlations between adjacent pixels. In [6], a simple spatial interpolation is used. In [7] a directional extrapolation algorithm was proposed, which exploits the fact that high frequencies, and especially edges, are visually the most relevant features. An algorithm for preservation of edges and borders in the transformed domain based on projections onto convex sets has been also proposed [8]. A technique including edge detectors combined with a Hough transform, a powerful tool for edge description, was utilized in [9]. A more advanced Hough transform based method was proposed in [10]. However, the performance of these methods drops when multiple edges or fine textures are involved. Modelling natural images as Markov random fields for EC was treated in [11]. This scheme produces relatively small squared reconstruction errors at the expense of an oversmoothed (and, therefore, blurred) image. The authors in [12] combined edge recovery and selective directional interpolation in order to achieve a more visually pleasing texture reconstruction. A content adaptive algorithm was introduced in [13]. A simple interpolation is applied if there are only a few edges crossing the missing macroblock and a best-match approach is applied if the macroblock is decided to contain texture. For this algorithm, and in general for all switching SEC techniques, a correct classification is critical since an erroneous decision on the macroblock behaviour could have a very negative effect on the final reconstruction. Inpainting-based methods can also be adopted for SEC purposes [14], [15]. Sequential pixel-wise recovery based on orientation adaptive interpolation is treated in [16]. As we will show later, pixel by pixel recovery usually suffers from smoothing high frequency textures. In [17], Bayesian restoration is combined with DCT pyramid decomposition. Bilateral filtering exploiting a pair of Gaussian kernels is treated in [18]. The algorithm seems quite competitive although some high frequency textures may be found overfiltered. Recently, SEC techniques in transform domains [19] have shown promising results although ringing can be observed in some cases.

TEC techniques take advantage of temporal and/or spatial redundancy as well. A joint video team (JVT) reference software TEC algorithm includes frame copying and motion vector copying [20]. A more advanced recovery of lost motion vectors is based on the boundary matching algorithm (BMA) [21] that minimizes the squared error between the outer boundary of the lost macroblock and the inner boundary of macroblocks found in the reference frame. A slight modification of BMA, overlapping BMA (OBMA), matches the outer boundaries of both the missing macroblock and the reference, leading to more accurate reconstructions [21]. These techniques, however, consider a linear movement and assume that the entire macroblock has been moved the same way. This issue is palliated by a multi-hypothesis approach (based on BMA) [22] which, however, lacks in generality. In [23], MV's are estimated by a Lagrangian interpolation of previously extrapolated MV's. This technique is entirely based on MV's so maintaining spatial continuity may be an issue. An edge-directed hybrid EC algorithm was proposed in [24]. Strong edges are estimated first and regions along these edges are recovered afterwards. Another combined EC technique is presented in [25]. It is a modification of the classic BMA under spatio-temporal constraints with an eventual posterior refinement based on partial differential equations. However, the improvement over the BMA is rather moderate. A MAP estimator, using an adaptive Markov random field process, is used to conceal the lost macroblocks in [26]. A statistically driven technique, based on a Gaussian mixture model is obtained in [27] from spatial and temporal surrounding information. This model, however, requires an extensive offline training. A computationally lighter version is described in [28]. Interesting results are obtained in [29] where a sparse representation based on local dictionaries is used for image reconstruction. This method, however, lacks in flexibility when complex textures are present and the concealment in scanning order may not always be appropriate. Recently, refinement technique [30] based on spatial and temporal AR models has been proposed. However, it is highly dependent on the previous MV estimate (using BMA, for example) and it assumes that (small) groups of macroblocks can be modelled using the same AR process which for low resolution videos or complex scenes may be inaccurate.

In this paper we propose an error concealment technique that automatically adapts itself to SEC [31], TEC or a combined SEC/TEC scheme according to the available information. Our proposal tries to fix or palliate some of the weak points of the previously referenced work such as blurring, blocking or filling order. The lost regions are recovered sequentially using a linear predictor whose coefficients are estimated by an adaptive procedure based on sparsity and a missing data imputation approach. First, we formulate the problem of estimating the predictor coefficients (only for SEC) as a convex optimization problem and then we derive an efficient alternative based on an exponential approximation. Although different exponential estimators have been used in EC algorithms [17], [18], a thorough treatment, combined with a linear prediction model, sparse recovery and sequential filling is proposed in this paper. This leads to a more generic and flexible EC technique. We also show that our EC scheme can be straightforwardly extended to also account for temporal correlations in video sequences (TEC and SEC/TEC). The experimental results show that our proposals provide better performance than other existing state-of-the-art algorithms on a wide selection of images and video sequences. In particular, the exponential approximation provides the best perceptual results.

The paper is organized as follows. In Section II we formulate the problem and introduce the linear prediction image model employed in the optimization process as well as the estimator (linear predictor) used for EC. The convex optimization based error concealment algorithm and its exponential approximation are presented in Section III. The model for video sequences is treated in Section IV. Simulations results and comparisons with other SEC and TEC techniques are presented in Section V. The last section is devoted to conclusions.

II. Linear Prediction Modelling and its Application to Error Concealment

Our aim is to conceal a lost region by optimally exploiting the correlations with the correctly received and decoded pixels in

its neighbouring area. These correlations will be modelled and exploited by means of vector linear prediction as it is described in Section III.

The Sections II-A–II-C describe how this model can be suitably estimated and applied to our concealment task.

A. Vector LP-Based Spatial Modelling

Let us assume that our image can be modelled as a stationary random field. Then, we can expect that every pixel z can be linearly predicted from a small set of surrounding pixels. The corresponding linear prediction (LP) model is defined by,

$$z = \sum_{(k,l)\in\mathcal{R}_z} w(k,l)z(k,l) + \nu \tag{1}$$

where $w(k,l)$ are the LP coefficients, $\mathcal{R}_z$ is the region of surrounding pixels employed for prediction, and ν is the residual error. We will assume integer pixel values belonging to $\Psi = [0, 255]$ for each colour space component.

In our case, we are interested in LP-based reconstruction of groups of lost pixels. Thus, it is convenient to re-formulate the above LP spatial modelling into a vector form by replacing the pixels $z(k,l)$ in (1) by pixel vectors. Let $\boldsymbol{z}$ be an arbitrarily shaped group of pixels that we want to express in terms of our LP model. Writing $\boldsymbol{z}$ as a column vector, we have that $\boldsymbol{z} \in \Psi^n$, where n is the number of pixels contained in $\boldsymbol{z}$. Also, let $\mathcal{Z} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{|\mathcal{Z}|}\}$ be the set of all possible spatially shifted versions of $\boldsymbol{z}$ which are employed to predict it. Then, the whole region employed to predict $\boldsymbol{z}$ is,

$$\mathcal{N}_{\boldsymbol{z}} = \bigcup_{j=1}^{|\mathcal{Z}|} \boldsymbol{z}_j. \tag{2}$$

Again, we can expect that prediction can be carried out with a small number $|\mathcal{Z}|$ of neighbouring vectors. Now, (1) can be extended to a vector form as follows,[1]

$$\boldsymbol{z} = \sum_{j=1}^{|\mathcal{Z}|} w_j \boldsymbol{z}_j + \boldsymbol{\nu}, \tag{3}$$

where $\boldsymbol{\nu}$ is the corresponding vector of residuals and $w_j \geq 0$ for all $j = 1, \ldots, |\mathcal{Z}|$.

The previous LP model can be applied to estimate $\boldsymbol{z}$ from the known neighbour vectors in region $\mathcal{N}_{\boldsymbol{z}}$ as,

$$\hat{\boldsymbol{z}} \;=\; \sum_{j=1}^{|\mathcal{Z}|} w_j \boldsymbol{z}_j \,. \tag{4}$$

In order to obtain optimal LP coefficients, the residual energy

$$\epsilon(\boldsymbol{w}) \triangleq \|\boldsymbol{\nu}\|^2 = \left\| \boldsymbol{z} - \sum_{j=1}^{|\mathcal{Z}|} w_j \boldsymbol{z}_j \right\|_2^2 \tag{5}$$

is usually minimized by solving a system of normal equations.

[1]Note that the intraprediction scheme used the in H.264 codec is a particular case of (3).

B. Application to Error Concealment: Sparse LP

We will denote $\mathcal{S}$ as the set of known pixels and $\mathcal{L}$ will denote the set of lost pixels (see Fig. 1(a)). When applying the above LP estimator of (4) to compute a lost group of pixels $\boldsymbol{z}$, we are facing two problems:

1) Since $\boldsymbol{z}$ is not known, it is not possible to find the residual energy function $\epsilon(\boldsymbol{w})$ exactly. In order to solve this problem, a solution based on missing-data imputation is proposed later in this section.
2) The region $\mathcal{N}_{\boldsymbol{z}}$ required for prediction is not known either. Instead, we have to employ a support area $\mathcal{S}$ of available (correctly received and decoded) pixels which provides us with a set $\mathcal{Z}'$ containing $M = |\mathcal{Z}'|$ available neighbour vectors $\boldsymbol{z}_j$ $(j = 1, \ldots, M)$, that is,

$$\mathcal{S} = \bigcup_{j=1}^{M} \boldsymbol{z}_j. \tag{6}$$

Then, some pixels required for prediction in (4) may be missing. Also, since the image is, in general, non-stationary, the support area $\mathcal{S}$ may include a high number of alien pixels not useful for predicting $\boldsymbol{z}$ ($M \gg |\mathcal{Z}|$, typically). As a result, the usual least-squares solution based on solving a system of normal equations is not suitable in our case. Typically, this solution involves the inversion of a huge $M \times M$ correlation matrix of small rank which would lead us to a poor solution. This small rank indicates that the number of vectors $\boldsymbol{z}_j \in \mathcal{S}$ useful for prediction is quite small. In other words, we can say that the solution $\boldsymbol{w} = (w_1, \ldots, w_M)^t$ we are seeking will be a sparse vector.

In order to overcome this last problem, the classical least-squares estimation of the LP coefficients can be replaced by a joint optimization of the squared error of (5) and the level of sparsity of the solution (typically represented by the ℓ_0-norm), which leads to a sparse linear prediction (SLP) scheme [32]. This scheme yields an unconstrained minimization problem, that we will represent as the following constrained optimization [33]:

$$\begin{aligned} \text{minimize} \quad & \epsilon(\boldsymbol{w}) = \left\| \boldsymbol{z} - \sum_{j=1}^{M} w_j \boldsymbol{z}_j \right\|_2^2 \\ \text{subject to} \quad & \|\boldsymbol{w}\|_0 \leq \delta_0 \text{ and } \boldsymbol{w} \succeq 0 \end{aligned} \tag{7}$$

where δ_0 is a parameter that controls the sparsity level and $\boldsymbol{w} \succeq 0$ is imposed to prevent negative pixels from the estimator (9) which is introduced later in this section. Moreover, preliminary experiments have shown that not using this last constraint would yield a worse performance.

This optimization involves two problems that will be addressed in Section III. First, we have that the ℓ_0-norm is non-convex and unfortunately also computationally infeasible for problems of higher dimensions. This problem is usually solved through convex relaxation. Second, we have the problem of selecting a suitable maximum value for sparsity parameter δ_0. We will shortly see that convex relaxation of (7) also

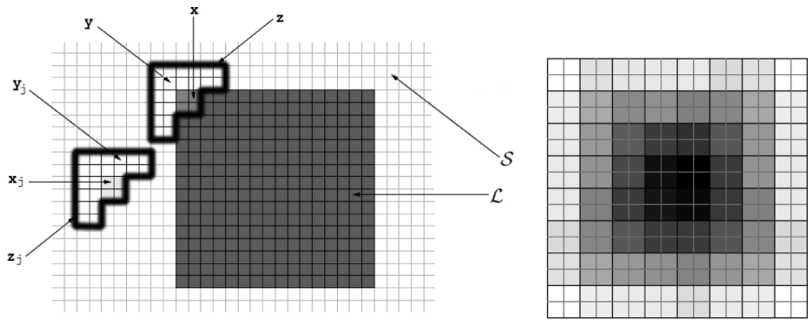


Fig. 1. (a) Example of configuration for the vectors $\boldsymbol{x}$, $\boldsymbol{y}$ and $\boldsymbol{z}$. $\mathcal{S}$ denotes the set of known pixels and $\mathcal{L}$ denotes the set of lost pixels. (b) Filling order for sequential reconstruction with 2×2 patches ($p = 2$). The regions illustrated by brighter level are recovered first.

provides a natural and smart solution to this issue which is proposed in Section III.

The LP formulation in (7) provides us with an adaptive procedure which dynamically obtains both the LP coefficients and the region of support $\mathcal{N}_{\boldsymbol{z}}$ (defined by those vectors $\boldsymbol{z}_j$ with $\boldsymbol{w}_j \neq 0$) for every image block $\boldsymbol{z}$. We still have the problem of $\boldsymbol{z}$ being unknown. As a consequence, the squared error $\epsilon(\boldsymbol{w})$ cannot be directly computed. In order to solve this, we will adopt a missing data approach where lost pixels can be imputed from known ones [34]. Instead of having a vector $\boldsymbol{z}$ completely unknown, we will consider that it contains both known and unknown pixels. Without loss of generality, let $\boldsymbol{z}$ be a group of pixels as shown in Fig. 1(a). Let the vector $\boldsymbol{z} = \boldsymbol{x} \cup \boldsymbol{y}$ consist of the two subvectors $\boldsymbol{x}$ and $\boldsymbol{y}$, where $\boldsymbol{x}$ denotes the missing pixels and $\boldsymbol{y}$ denotes correctly received and decoded pixels and can be seen as the spatial context of $\boldsymbol{x}$. Every $\boldsymbol{z}_j \in \mathcal{Z}'$ is split in a similar way, as shown in Fig. 1(a). Since $\boldsymbol{z}$ is (locally) stationary and $\boldsymbol{y} \subset \boldsymbol{z}$, then we can approximate the weights obtained from (7) by means of the following procedure:

$$\begin{aligned} \text{minimize} \quad & \epsilon_{\boldsymbol{y}}(\boldsymbol{w}) = \left\| \boldsymbol{y} - \sum_{j=1}^{M} w_j \boldsymbol{y}_j \right\|_2^2 \\ \text{subject to} \quad & \|\boldsymbol{w}\|_0 \leq \delta_0 \text{ and } \boldsymbol{w} \succeq 0. \end{aligned} \tag{8}$$

Section III will be devoted to the search for solutions to this optimization problem.

Finally, according to (4) the concealed group of pixels, $\hat{\boldsymbol{x}}$, can be approximated by a linear combination of blocks within its neighbourhood

$$\hat{\boldsymbol{x}} = \sum_{j=1}^{M} w_j^* \boldsymbol{x}_j, \tag{9}$$

where $\boldsymbol{w}^* = (w_1^*, \ldots, w_M^*)^t$ is the vector of optimal weights (LP coefficients) obtained by (8).

C. Application to Error Concealment: Sequential Filling

The H.264/AVC encoder packetizes the stream by slices so a loss of one packet implies a loss of, at least, one 16×16 macroblock. Applying (9) to $\boldsymbol{x} \in \Psi^{16\times 16}$ would lead to significant imprecisions due to blocking as well as blurring since it is often not possible to find a combination of $\boldsymbol{x}_j$'s suitably matching $\boldsymbol{x}$ due to the high number of dimensions in $\Psi^{16\times 16}$. This means that the residual error from (3) may still carry significant energy. This is the reason why the H.264/AVC standard also includes submacroblock prediction [3]. In order to manage with this problem, we introduce sequential recovery. Thus, the macroblock is recovered using a set of square patches $\hat{\boldsymbol{x}} \in \Psi^{p\times p}$ with $1 \leq p \leq 16$. Pixel-wise reconstructions ($p = 1$), as in [16], may introduce considerable blurring when high frequencies are involved (Fig. 11(b)). By using groups of pixels the correlation within a group is better preserved and so is the texture (Fig. 11(c)). Let us consider, without loss of generality, $p = 2$ and let $\boldsymbol{y}$ include all the received and already recovered pixels within the 6×6 block with the lost pixels $\boldsymbol{x}$ placed in its centre, as shown in Fig. 1(a). The macroblock is recovered sequentially by filling it with $\hat{\boldsymbol{x}}$ obtained by applying (8) and (9). The filling order is critical and it should preserve the continuity of image structures [15]. In [15], the filling priorities of every patch are set in order to maintain the continuity of isophotes and according to the amount of information within the patch. Our proposal, due to the shape of the context $\boldsymbol{y}$, can achieve an appropriate filling order in a much simpler way by using contexts reliabilities. We define the reliability ρ of context $\boldsymbol{y}$ as the sum of reliabilities of all its pixels. Initially, the reliability of a pixel is set to 1 if it has been correctly received and decoded. Missing pixels have reliability zero. When a pixel $x \in \boldsymbol{x}$ is concealed, its reliability is set to $\alpha\rho/m$, where $0 < \alpha < 1$ and m is the number of pixels contained in $\boldsymbol{y}$. We use $\alpha = 0.9$ in our simulations. The lost region $\boldsymbol{x}$, whose context $\boldsymbol{y}$ produces the highest reliability, is recovered first. The reliability is non-increasing and the reconstruction evolves from the outer layer towards the centre of the missing macroblock. Fig. 1(b) shows the filling order of a 16×16 macroblock using 2×2 patches. Note that the first patches to be concealed are the corners as their contexts are the largest ones, and thereby providing more reliable information (which leads to a more accurate estimate of the LP coefficients).

III. LP Parameter Estimation

The scheme proposed in the previous section requires the computation of a set of LP coefficients by solving the optimization problem of (8). In this section, we propose first a solution based on convex relaxation. Then, we derive a computationally less expensive algorithm by applying several approximations.

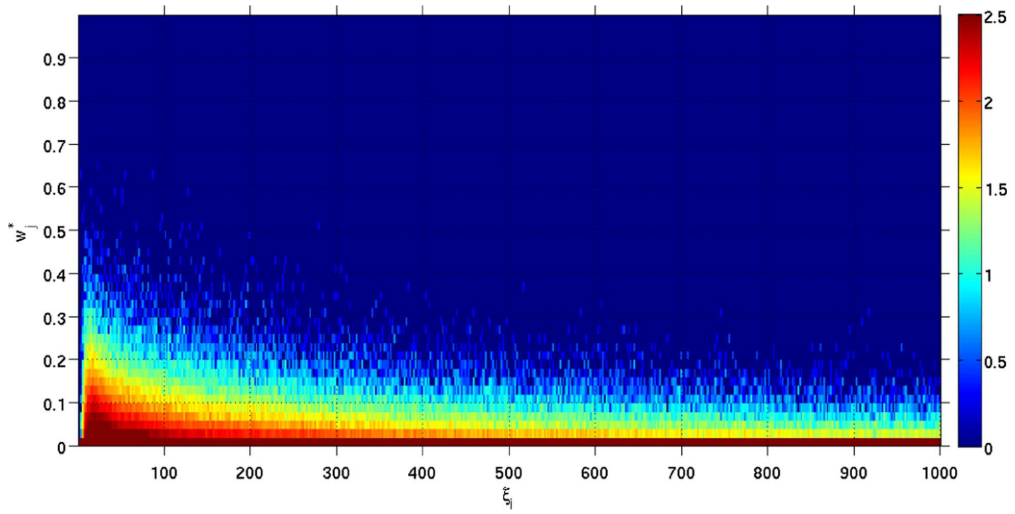


Fig. 2. Histogram of pairs squared-error/weight (ξ_j, w_j^*) for *Lena*. Logarithmic scale is employed for more clarity. For reconstruction purposes 2×2 patches are used and loss pattern from Fig. 6(b) is applied.

A. SLP Via Convex Relaxation (SLP-C)

The main problem that arises when solving (8) is that the ℓ_0-norm is non convex, so that this optimization usually requires exhaustive search and is therefore computationally prohibitive. Applying convex relaxation [35], the solution to the optimization defined by (8) can be modified in terms of the ℓ_1-norm as follows:

$$\begin{aligned} \text{minimize} \quad & \epsilon_{\boldsymbol{y}}(\boldsymbol{w}) = \left\| \boldsymbol{y} - \sum_{j=1}^{M} w_j \boldsymbol{y}_j \right\|_2^2 \\ \text{subject to} \quad & \|\boldsymbol{w}\|_1 \leq \delta_1 \text{ and } \boldsymbol{w} \succeq 0. \end{aligned} \tag{10}$$

In our simulations, this optimization is solved by the primal-dual interior point (IP) method [36].

The remaining problem is the selection of a suitable sparsity level δ_1 (redefined under the ℓ_1-norm). In order to do this, we will assume smoothness in the visual features of an image. This implies that the reconstructed block should not contain any singular features. In the particular case of luma, it means that a reconstructed pixel could not be brighter (darker) than the brightest (darkest) pixel in $\mathcal{S}$. This requires that (9) must be a convex combination and it implies that $\delta_1 = 1$. The resulting technique will be referred to as SLP-C in the following.

B. SLP With Exponentially Distributed Weights (SLP-E)

Although there are efficient algorithms for solving convex optimization problems, such as the IP method employed above, the processing time still remains very high and far from real-time. In this section we develop a fast approximation for solving the minimization problem in (10). Specifically, we show that the optimal weights $\boldsymbol{w}^*$ obtained from (10) can be well modelled by an exponential function.

According to (10), every context $\boldsymbol{y}_j$ has a weight w_j^* associated. Due to the high spatial correlation of an image, it is likely that contexts that produce smaller squared error, ξ_j, would generate larger weights, where we define the squared error ξ_j associated to a context $\boldsymbol{y}_j$ as,

$$\xi_j = \frac{\|\boldsymbol{y} - \boldsymbol{y}_j\|_2^2}{m}. \tag{11}$$

Fig. 2 represents the joint 2D histogram of pairs (ξ_j, w_j^*) for the image of *Lena*. The loss pattern applied is the one shown in Fig. 6(b). The histogram suggests that there is an exponential relationship between the squared errors ξ_j and the weights w_j^*. With this in mind, we propose the following approximation for the LP weights:

$$\hat{w}_j = C \exp\left(-\frac{1}{2}\frac{\xi_j}{\sigma^2}\right), \tag{12}$$

where σ^2 is a decay factor that controls the slope of the exponential and C is a normalization factor that ensures the sparsity constraint $\|\boldsymbol{w}\|_1 = 1$, that is,

$$C = \frac{1}{\sum_{i=1}^{M} \exp\left(-\frac{1}{2}\frac{\xi_i}{\sigma^2}\right)}. \tag{13}$$

Note that this normalization always forces the solution $\hat{\boldsymbol{w}}$ to have the maximum value of sparsity considered in (10), i.e., $\delta_1 = 1$. The corresponding LP estimator is obtained by replacing the optimal weights w_j^* by their exponential approximation $\hat{w}_j$ in (9). The resulting EC technique will be referred to as SLP-E in the following.

Let us analyze the approximation proposed in (12) and (13). We can see that the exponential trend observed in Fig. 2 cannot be written down as a single exponential function for the whole image. In fact, the figure shows lots of exponential contours. There are two reasons for this:

1) We must take into account the effect of the mild sparsity constraint applied in (10). Thus, given several similar contexts $\boldsymbol{y}_j$ (representing a certain context type) with small quadratic errors ξ_j (that is, relevant for reconstruction), the

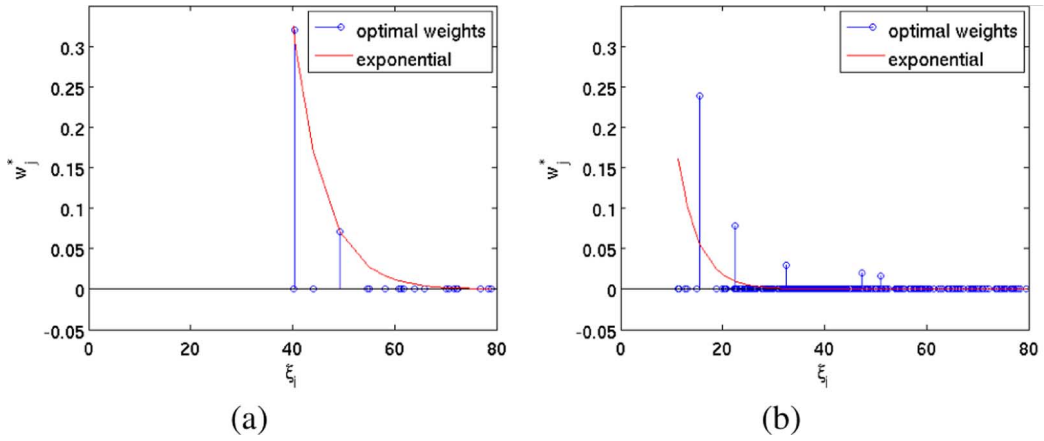


Fig. 3. Example of the exponential estimated by means of the optimal weights $\boldsymbol{w}^*$ for two different patches.

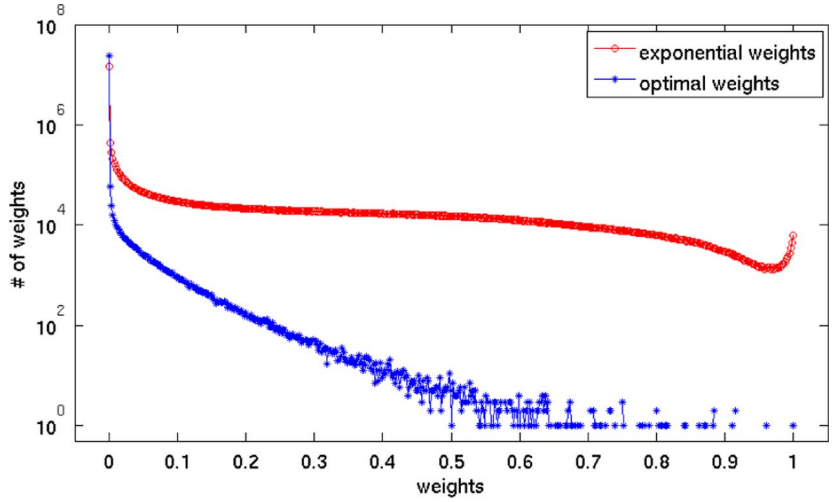


Fig. 4. Comparison of the weights histograms obtained by SLP-E (red) and SLP-C (blue) for the image of *Lena*. The vertical axis uses a logarithmic scale for a clearer visualization and σ^2 has been fixed to 10 for the whole image.

optimization algorithm picks one context and suppresses the others instead of using all of them. On the contrary, the exponential approximation relaxes the sparsity constraint and keeps all the relevant contexts.

2) We must also consider that Fig. 2 shows all the pairs (ξ_j, w_j^*) for all the patch linear predictors in the image. However, clearly all these linear predictors are different and must have a different factor σ^2, since this is the only free parameter in (12).

Let us consider first the issue of obtaining a suitable value of σ^2 for each patch predictor. This factor is related to the squared error ϵ_y and, therefore, to the local predictability of the image signal. In order to estimate a suitable value of σ^2 for every predictor, a logical solution is that of minimizing the prediction error $\epsilon_{\boldsymbol{y}} = \epsilon_{\boldsymbol{y}}(\sigma^2)$ defined in (10) but constrained to the LP weights defined by (12) and (13). Fig. 3 illustrates two examples of the optimal weights and their corresponding exponential approximations with factors σ^2 estimated as described above. In the first example, the exponential function mainly follows the most relevant optimal weights. However, in the second one, the exponential approximation leads to weights which are smaller than the optimal ones. In order to understand this, we must take into account that there is a considerable number of zero-valued optimal weights in the small squared error area, which is due, as previously explained, to the mild sparsity constraint. On the contrary, the exponential approximation introduces a sparsity relaxation and the weight assigned to a certain type of context is distributed among the contexts of that type through the normalization in (13) and the selection of a suitable σ^2. We must point out that the sparsity relaxation just described is quite limited. In order to see this, the histograms for both optimal and exponential weights are depicted in Fig. 4. We can see that although the exponential approximation reduces sparsity, most of the weights are still close to zero.

TABLE I
ESTIMATED VARIANCE (MEAN VALUE AND STANDARD DEVIATION) FOR TESTED IMAGES

σ^2	*Lena*	*Clown*	*Office*	*Barbara*	Average
mean	6.70	12.01	7.35	12.99	9.76
std	14.69	35.53	15.50	20.09	21.45

Table I shows the mean value and the standard deviation of σ^2 for several tested images. $\epsilon_{\boldsymbol{y}}(\sigma^2)$ minima have been obtained

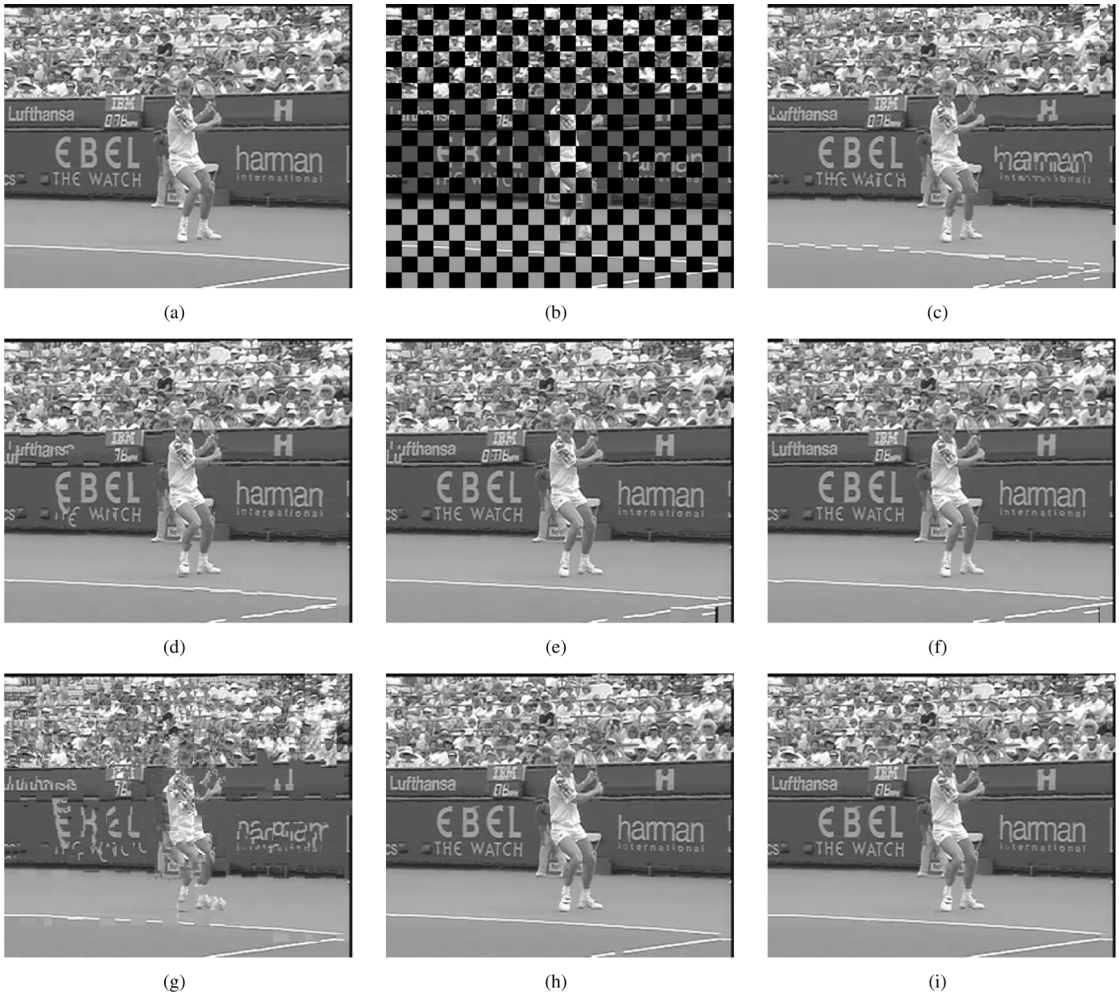


Fig. 5. Comparison of reconstructions obtained by different procedures. (a) Original frame. (b) Received frame. (c) Reconstruction by average MV replacement ($\mathrm{PSNR} = 21.36, \mathrm{MS}-\mathrm{SSIM} = 86.52$). (d) Reconstruction by BMA ($\mathrm{PSNR} = 22.44, \mathrm{MS}-\mathrm{SSIM} = 90.24$). (e) Reconstruction by OBMA ($\mathrm{PSNR} = 24.14, \mathrm{MS}-\mathrm{SSIM} = 96.86$). (f) Reconstruction by MHEC ($\mathrm{PSNR} = 24.37, \mathrm{MS}-\mathrm{SSIM} = 96.93$). (g) Reconstruction by SLP-E using spatial information only ($\mathrm{PSNR} = 18.56, \mathrm{MS}-\mathrm{SSIM} = 77.80$). (h) Reconstruction by SLP-E using temporal information only ($\mathrm{PSNR} = 25.79, \mathrm{MS}-\mathrm{SSIM} = 97.73$). (i) Reconstruction by SLP-E using both spatial and temporal information ($\mathrm{PSNR} = 25.93, \mathrm{MS}-\mathrm{SSIM} = 97.74$).

by exhaustive search. In the following and for the sake of computational simplicity, a fixed value of σ^2 will be used. Simulations reveal that this simplification, along with the exponential approximation, leads to a factor of 100 of computational saving with respect to SLP-C. For natural images, σ^2 values around 10 lead to visually good results (Fig. 7(b)). Larger values of σ^2 may lead to oversmoothing (Fig. 7(c)) while smaller values may lead to numerical instability and should be avoided (Fig. 7(a)) (unless the image is extremely stationary).

Finally, we must also point out that the approach developed here can be alternatively interpreted as a non-parametric kernel-based regression, in particular, as a multivariate Nadaraya-Watson estimator.

IV. Temporal Model of a Video Sequence

The importance of temporal correlations is reflected by the fact that they are a crucial issue in video coding. However, in the temporal domain, video signals tend to be non-stationary due to motion. That is, the pixel $z(i,j)$ in the current frame usually cannot be predicted using the pixels with the same location in previous frames [37]. This can be palliated by applying motion compensation. In fact, the H.264/AVC standard encodes the submacroblock $sMB^{(n)}_{(i,j)}$ belonging to the current P-frame n as

$$sMB^{(n)}_{(i,j)} = sMB^{(n-\tau)}_{(i+MV(i),j+MV(j))} + \boldsymbol{r} \qquad (14)$$

where $MV(i,j)$ is the motion vector, $\boldsymbol{r}$ is the residual error and τ is the temporal lag to the reference frame n. Note that τ depends on visual properties of the video as well as the dimension of the prediction buffer. Moreover, regardless of the buffer size, the encoder selects the sparsest set of weights since only one reference submacroblock is taken into account. For B-frames and P-frames where weighted prediction is applied, two reference submacroblocks are utilized.

The estimation scheme of Section II can be straightforwardly extended in order to account for both temporal and spatial correlations. In this case, (3) could be seen as a generalization of (14). The stationary region $\mathcal{N}_{\boldsymbol{z}}$ will now not only comprise pixels from the current frame but also pixels from the previous frames. As in the case of SEC, the stationary 3D region is unknown and the whole support area $\mathcal{S}$ needs to be searched. We will set

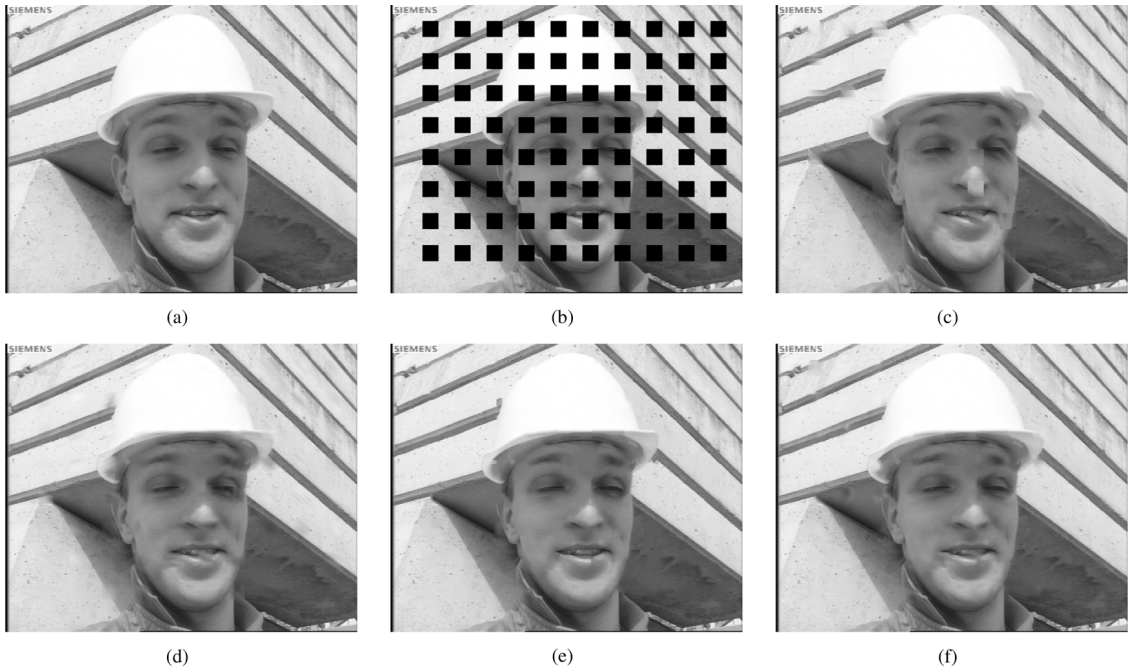


Fig. 6. SEC for the image of *Foreman* (a) Original image, (b) Received data, (c) Reconstruction using CAD ($\mathrm{PSNR} = 31.46$ dB, $\mathrm{MS-SSIM} = 97.54$), (d) FSE ($\mathrm{PSNR} = 34.17$ dB, $\mathrm{MS-SSIM} = 98.03$), (e) SLP-E ($\mathrm{PSNR} = 35.46$ dB, $\mathrm{MS-SSIM} = 98.73$), (f) SLP-C ($\mathrm{PSNR} = 35.48$ dB, $\mathrm{MS-SSIM} = 98.68$).

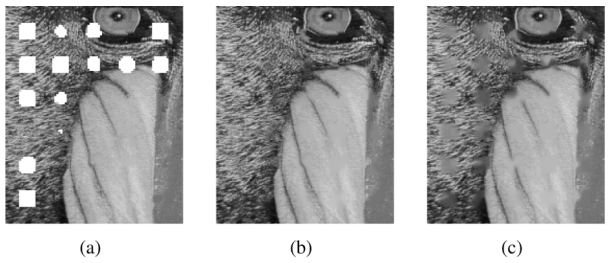


Fig. 7. EC with SLP-E for different values of σ^2: (a) $\sigma^2 = 0.5$, numerically unstable reconstructions are represented with white level, (b) $\sigma^2 = 10$, (c) $\sigma^2 = 50$.

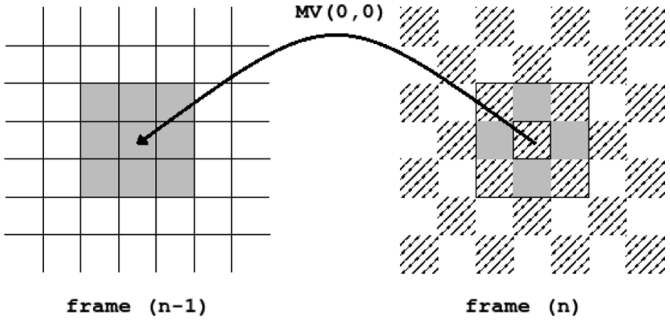


Fig. 8. Support area $\mathcal{S}$ (grey 16×16 macroblocks) for combined TEC/SEC. The striped macroblocks are lost.

the support area to include all the available neighbouring macroblocks from the current frame (as in the previous section) and all the corresponding macroblocks from the previous frame. For the sequences of *Foreman* and *Stefan*, more than 99% of MV have $\tau = 1$, so considering only the previous frame is a reasonable simplification. Fig. 4 illustrates an example where the corrupted frame utilizes dispersed slicing and the previous frame is received without errors.

In practice, the loss of a NALU implies that residual errors as well as motion vectors are lost (unless data partitioning is applied at the encoder side at the expense of a higher bit-rate) [3]. In order to obtain high quality predictions, the support area $\mathcal{S}$ should include all the motion compensated pixels located within the corrupt macroblock. For a standard frame rate of 30 fps, the motion vectors between two consecutive frames are likely to be moderate. In fact, Fig. 9 shows the histograms of motion vectors norm for four different 30-frame video sequences. It follows from the histograms that the support area composed as described above covers more than 95% of motion vectors. In other words, in less than 5% of cases the motion compensated macroblock lies (completely or partially) outside the support area (MV amplitude greater than 16). For the sake of computational simplicity, we assumed that the motion vectors were calculated using only the previous frame. The more motion vectors that are covered, the better reconstructions would be obtained as a more complete set of motion compensated pixels (useful for prediction) is used. However, the processing time increases with $|\mathcal{S}|$ so applying the proposed support area is a reasonable trade-off. Using this support area, the weights will be computed in the same way as in (12).

Note that pixels from the surroundings (within the current frame) of the missing macroblock are also included. Thus, the algorithm automatically decides whether to use SEC, TEC or combined concealment. This is the consequence of dynamically obtaining the LP coefficients and estimates the stationary area $\mathcal{N}_{\boldsymbol{z}}$, as discussed in Section II-B. For example, if the previous frame belongs to a different scene, all relevant weights calculated by (12) will most likely come from the current frame

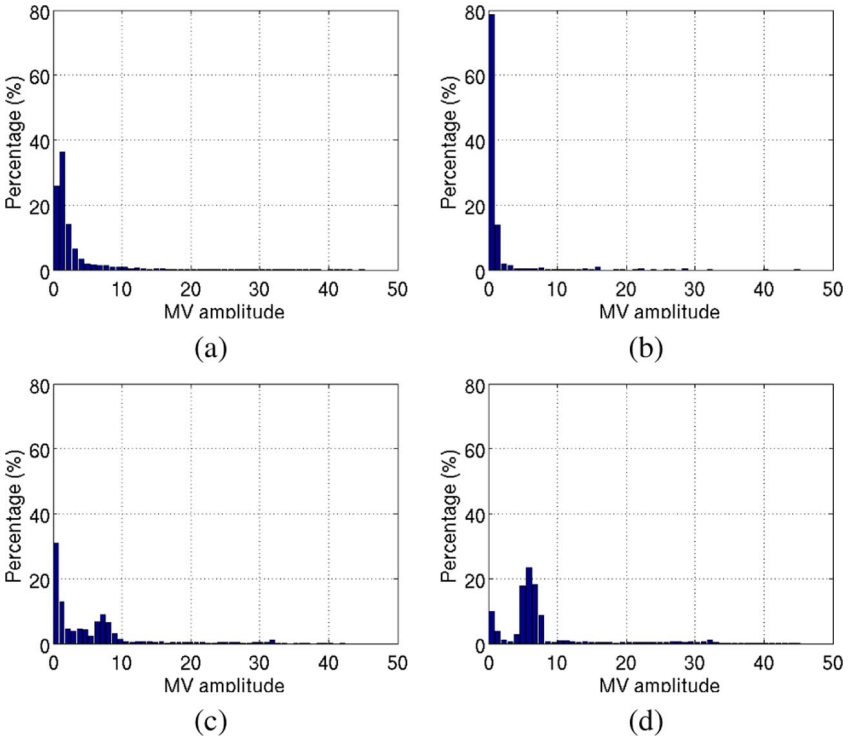


Fig. 9. Histogram of MV amplitudes for the video sequences of (a) *Foreman*, (b) *News*, (c) *Stefan* and (d) *Bus*. Motion vectors were obtained by minimizing the residual error and applying full range search.

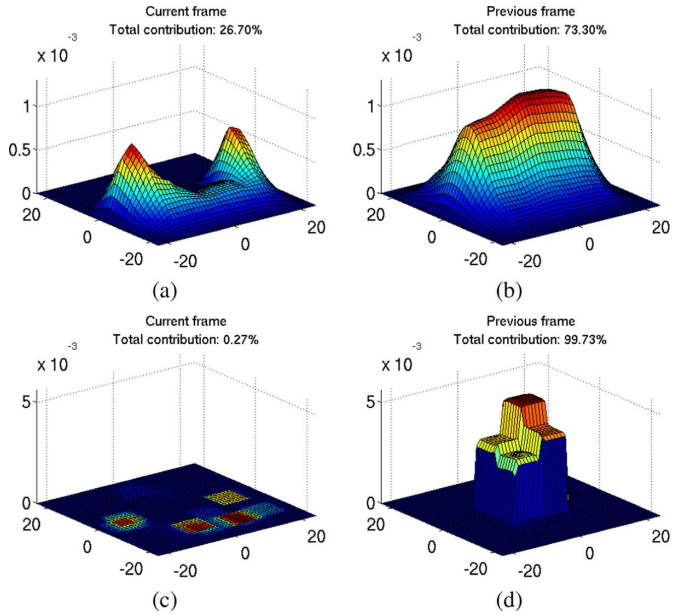


Fig. 10. Average weight per pixel from $\mathcal{S}$ for the sequence of *Stefan* (a)–(b) and *Waterfall* (c)–(d). The percentage indicates the total contribution from pixels from the current frame ((a) and (c)) and from the previous frame ((b) and (d)) to the final reconstruction.

and the contribution of pixels from the previous (uncorrelated) frame will be negligible. Nevertheless, temporal correlation is usually higher than the spatial one and this phenomenon is observed in the reconstruction process. Fig. 10 shows the average weight associated with each pixel within the support area for two different video sequences. We see that the contribution of

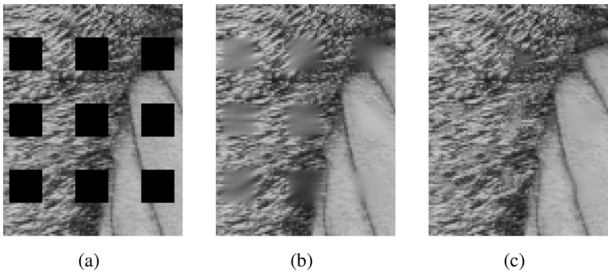


Fig. 11. Example of PSNR and MS-SSIM response to different image reconstructions. (a) Received image, (b) reconstructed by orientation adaptive interpolation (OAI) [16] ($\mathrm{PSNR} = 27,22, \mathrm{MS-SSIM} = 92,47$) (c) reconstructed using SLP-E (12) with $p = 2(\mathrm{PSNR} = 25,56, \mathrm{MS-SSIM} = 94,76)$.

pixels belonging to the previous frame is considerably higher than the contribution of those within the current frame. Simulations show that for standard video test samples, composed by a single shot, the amount of information (pixels) gathered from the previous frame is higher than 70%. Moreover, in some particular cases there will be almost no good template matches within the current frame, as shown in Figs. 10(c) and (d). Unlike the pure spatio-temporal hybrid algorithms, our proposal is applicable both for still images (or I-frames) and video. Since temporal correlations tend to be higher than spatial correlations, then smaller values of σ^2 are preferred. Moreover, due to the same reason, larger patches may be utilized to speed up the algorithm and obtain higher quality reconstructions. Here, σ^2 is set to 5 and 8×8 patches are employed.

Fig. 5 shows a comparison of our proposal using only spatial information, temporal information and a combination of

TABLE II
PSNR VALUES (IN DB) AND MS-SSIM INDICES (SCALED BY 100) FOR TEST IMAGES RECONSTRUCTED BY SEVERAL ALGORITHMS FOR BLOCK DIMENSIONS 16×16. THE BEST PERFORMANCES FOR EACH IMAGE ARE IN BOLD FACE

PSNR	BIL	POC	EXT	SHT	CAD	AVC	MRF	INP	BLF	OAI	FSE	SLP-E	SLP-C	ORA
Average	26.92	25.94	26.76	27.19	29.12	28.08	29.12	28.78	29.71	30.15	30.20	30.32	**30.90**	31.36
Lena	30.00	28.04	29.39	30.47	30.44	30.42	32.17	30.88	32.15	32.82	32.72	32.55	**32.85**	33.41
Goldhill	30.00	28.50	29.57	29.97	30.24	31.27	31.12	30.40	30.91	31.54	31.78	31.54	**32.07**	32.97
Foreman	27.12	28.49	29.26	28.34	31.46	29.11	32.98	33.87	34.75	35.03	34.18	35.46	**35.48**	37.38
Barbara	26.19	24.30	25.85	26.40	26.78	26.85	27.99	28.04	29.91	29.66	30.84	30.79	**31.91**	32.15
Office	27.54	27.56	27.32	27.54	29.43	29.99	29.77	29.64	30.06	31.77	31.33	31.30	**32.06**	32.68
Cameraman	25.96	23.66	24.82	26.16	26.51	26.14	26.67	25.45	26.03	27.27	**27.44**	27.24	27.27	27.28
Baboon	24.15	24.63	24.72	24.14	24.92	25.42	26.14	25.06	26.05	26.06	26.02	25.70	**26.21**	25.93
Clown	27.76	24.36	26.30	27.62	29.12	28.55	28.23	27.89	28.73	29.75	29.19	27.39	**30.79**	31.00
Tire	23.59	23.92	23.82	24.10	24.47	25.43	27.00	26.37	28.76	27.42	28.31	28.77	**29.32**	29.43
MS-SSIM														
Average	93.83	91.23	93.82	94.12	95.69	94.74	95.87	95.53	96.35	95.81	96.36	**97.04**	96.58	97.80
Lena	96.72	92.85	96.52	97.03	96.59	96.56	97.64	96.65	97.44	97.65	97.80	**97.97**	97.75	98.48
Goldhill	93.83	92.50	94.81	93.86	94.53	95.65	95.71	95.21	95.52	95.62	96.14	**96.43**	96.34	97.50
Foreman	95.16	93.09	97.16	95.65	97.58	96.87	98.10	98.22	98.15	98.68	97.92	**98.70**	98.65	99.19
Barbara	95.24	89.42	94.70	95.57	95.73	94.87	96.00	95.72	97.04	97.07	97.64	97.92	**98.12**	98.67
Office	93.93	93.24	94.84	93.92	95.66	95.77	96.21	96.35	96.12	97.27	96.90	**97.45**	97.39	98.32
Cameraman	93.38	87.22	93.14	93.47	94.87	93.72	94.95	93.96	95.97	93.87	94.31	**96.55**	92.93	96.83
Baboon	88.96	90.16	91.38	88.81	91.89	91.91	93.09	90.95	93.33	92.61	93.32	**93.42**	93.38	94.83
Clown	95.61	91.17	94.09	95.53	96.00	95.55	95.55	95.74	96.22	95.28	96.40	**97.19**	97.13	98.23
Tire	91.65	91.40	93.17	93.12	88.56	92.07	95.59	94.19	97.35	94.26	96.81	**97.70**	97.30	98.16

both with other techniques. In fact, it is observed, at both objective and subjective levels, that using only spatial information achieves poorer quality. The improvement of the combined method over the pure TEC is small, as can be also deduced from Fig. 10. Nevertheless, including spatial information may provide a noticeable visual improvement as can be observed comparing Fig. 5(h) and (i).

V. SIMULATION RESULTS

In order to better take into account the perceptual quality, the multi scale structural similarity (MS-SSIM) index [38] is used for comparison along with the PSNR measure. In the former case, the image is sequentially low-pass filtered and subsampled, so a set of images is obtained, including the original resolution. Then, the SSIM index is applied for every subimage within the set. The SSIM index aims at approximating the human visual system (HVS) response looking for similarities in luminance, contrast, and structure [39]. This index can be seen as a convolution of a fixed-sized mask with the residual error between the reference image and the concealed image [40]. A unique mask size is used for each of the images within the set. Therefore fine as well as coarse textures and objects are taken into account.

As shown in Fig. 11, the PSNR does not respond to perceptual visual quality as well as the MS-SSIM index does, since PSNR is a quality criterion merely based on the mean squared error. In spite of that, the weights $\boldsymbol{w}^*$ are obtained according to the squared error (12) since the SSIM index tends to marginalize the influence of changes in intensity [41]. This is a desirable behaviour when measuring the overall perceptual image quality but not when computing predictor coefficients. Thus, the squared error is used when computing the weights while the MS-SSIM index is preferred for an overall quality measure.[2]

The performance of our proposals in SEC mode is tested on the images of *Lena* (512×512), *Barbara* (512×512), *Baboon* (512×512), *Goldhill* (576×720), *Clown* (512×512), Matlab built-in images *Cameraman* (256×256), *Office* (592×896), *Tire* (192×224) and the first frame of *Foreman* (288×352) sequence. The test is carried out for 16×16 macroblocks and the rate of block loss is approximately 25%, corresponding to a single packet loss of a frame with dispersed slicing structure. We compare the performance with other SEC methods such as bilinear interpolation (BIL) [6], projections onto convex sets (POC) [8], directional extrapolation (EXT) [7], a Hough transform based SEC (SHT) [10], content adaptive technique (CAD) [13], non-normative SEC for H.264 (AVC) [42], Markov random fields approach (MRF) [11], inpainting (INP) [15], bilateral filtering (BLF) [18], frequency selective extrapolation (FSE) [19] and orientation adaptive interpolation (OAI) [16].[3] Both SLP via convex relaxation (SLP-C) and SLP with exponentially distributed weights (SLP-E) are tested. In the simulations, σ^2 is set to 10 and grey level images are used. Note that a pixel reconstructed by any of the aforementioned algorithms is usually real-valued and does not necessarily belong to Ψ. Thus, reconstructed pixels are rounded to the closest member of Ψ. A subjective comparison of the different algorithms is shown in Fig. 6. As can be seen in Table II, SLP-C provides the best PSNR results as expected, but SLP-E outperforms all the other technique for all the tested images in terms of MS-SSIM, leading so to higher perceptual quality reconstructions. Moreover, the average MS-SSIM and PSNR are superior to those of state-of-the-art algorithms. In addition, an oracle SPL-E (ORA) is included, where the best σ^2 (the value which provides the best reconstruction) is applied for every patch, and it represents the superior limit of the SPL-E performance.

[2]Note that the MS-SSIM index lies in $[-1; 1]$. In this section, we have scaled the index by 100 in order to better illustrate the differences.

[3]Implementations of most of these techniques, as well as the implementation of our algorithm, is available online at [43].

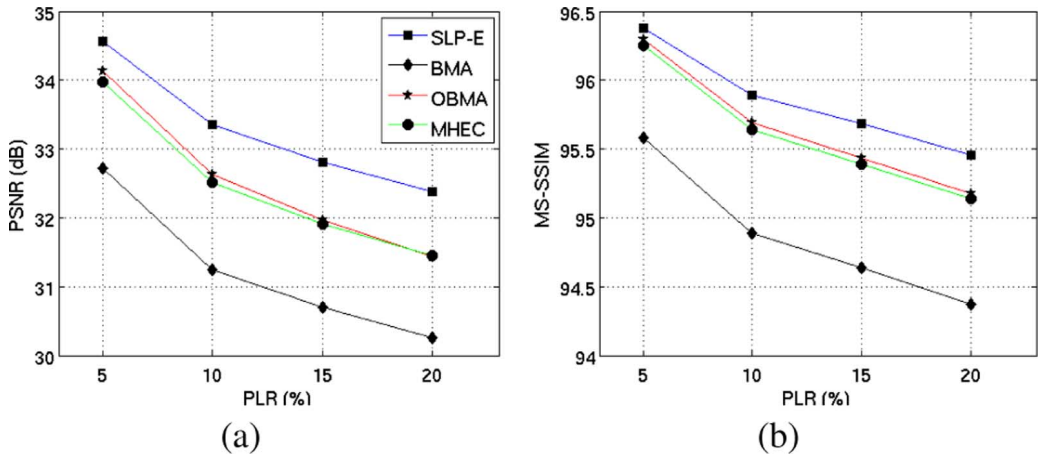


Fig. 12. Average PSNR (a) and MS-SSIM (b) values versus packet loss-rates averaged for all the tested video sequences. Tested procedures: SLP-E in the combined SEC/TEC mode, BMA, OBMA and MHEC.

The proposed SLP-E technique in the combined SEC/TEC mode is tested for H.264 coded video sequences of *Foreman*, *Stefan*, *Ice*, *Football*, *Bus*, *Irene*, *Flower* and *Highway*. All sequences employ the common intermediate format (CIF, 352×288) and they comprise 30 frames, where only the first frame is intracoded and the remaining frames are predictive coded. An aggressive block loss-rate is applied by utilizing a dispersed slicing structure with two slices per frame (the so-called chessboard structure, see Fig. 5(b)). In this scenario, a loss of one packet implies a loss of 50% of the macroblocks within a frame. However, note that our proposal can be easily extended to other slicing modes. The quantization parameter is set to 25 and the prediction buffer is one frame deep. Packet losses are randomly generated at rates of 5%, 10%, 15% and 20%. For each packet loss rate (PLR), the sequence is transmitted 20 times and the average PSNR and MS-SSIM values are calculated. The proposed technique is compared with other TEC algorithms, namely BMA [21], OBMA [21] and multi-hypothesis EC (MHEC) [22]. The search range for BMA, OBMA and MHEC is $[-16, 16]$ using the zero MV as the starting point, i.e., BMA, OBMA, MHEC and our proposal all work with the same information gathered from the previous frame. The proposed SLP-E outperforms the other techniques for all the tested sequences both in terms of PSNR and MS-SSIM. The results for half of the eight sequences are shown in Table III. PSNR and MS-SSIM values, averaged over all the tested sequences, are shown in Fig. 12. Finally, a subjective comparison is shown in Fig. 5.

Regarding the computational complexity, Table IV shows the processing time ratios of [15], [24] and SLP-E to BMA. We can observe that our proposal requires less processing time than some of the state-of-the-art techniques. Moreover, the average gains of [24] over BMA are approximately 2 dB for *Foreman* and 1 dB for *Irene* and it outperforms [15] for both cases. Utilizing the same simulation setup as in [24] (dispersed slicing, quantization parameter set to 25 and PLR of 3%, 5%, 10% and 20%), SLP-E achieves average gains over BMA of 2.55 dB and 1.20 dB, respectively. Thus, SLP-E outperforms both [24] and [15] with less computational burden. Due to the nature of our algorithm, the processing time per MB is approximately constant regardless of the sequence and its resolution, as has been confirmed by the simulations.

Finally, given a multi-scene sequence, the error may occur in the border frame (usually intracoded). In such a case, MV based techniques fail since they try to extract the concealment information from the previous, and therefore uncorrelated, frame. Modified BMA and OBMA are able to gather the information from the current frame although the reconstructions tend to be of poor quality since both algorithms seek the best match for

TABLE III
AVERAGE PSNR AND MS-SSIM VALUES FOR DIFFERENT PLR FOR VIDEO SEQUENCES OF *FOREMAN*, *STEFAN*, *FOOTBALL* AND *ICE*. TESTED PROCEDURES: BMA, OBMA, MHEC AND SLP-E. THE BEST PERFORMANCES FOR EACH SEQUENCE ARE IN BOLD FACE

Method	PSNR				MS-SSIM			
	PLR							
	5%	10%	15%	20%	5%	10%	15%	20%
	Foreman							
BMA	35.56	34.03	33.60	32.82	97.35	96.88	96.73	96.48
OBMA	37.69	36.33	35.61	34.90	97.81	97.49	97.30	97.13
MHEC	37.51	36.25	35.70	35.03	97.79	97.51	97.36	97.20
SLP-E	**38.12**	**37.33**	**37.05**	**36.53**	**97.82**	**97.67**	**97.63**	**97.52**
	Stefan							
BMA	29.59	28.18	28.01	27.37	93.84	92.89	92.80	92.37
OBMA	30.46	29.15	28.88	28.31	94.79	94.08	93.96	93.62
MHEC	30.61	29.25	28.98	28.42	94.87	94.14	94.05	93.65
SLP-E	**31.27**	**30.31**	**29.96**	**29.36**	**95.10**	**94.56**	**94.40**	**94.03**
	Football							
BMA	30.37	28.84	28.64	27.95	93.31	92.03	91.75	91.30
OBMA	31.07	29.34	28.93	28.16	93.86	92.44	91.91	91.29
MHEC	30.76	29.15	28.70	28.03	93.60	92.18	91.59	91.14
SLP-E	**31.53**	**30.06**	**29.79**	**29.08**	**94.11**	**93.04**	**92.71**	**92.27**
	Ice							
BMA	34.54	32.39	31.19	31.10	97.76	97.31	97.01	96.92
OBMA	35.25	32.85	31.55	31.16	98.09	97.57	97.23	97.04
MHEC	34.74	32.61	31.24	31.00	97.97	97.45	97.07	96.88
SLP-E	**35.48**	**33.57**	**32.53**	**32.40**	**98.11**	**97.76**	**97.52**	**97.44**

TABLE IV
AVERAGE ERROR CONCEALMENT TIME FOR A CORRUPTED FRAME COMPARED TO BMA

Sequence	[24]	[15]	SLP-E	BMA
Foreman	13.28	13.65	9.24	1.00
Irene	9.71	13.13	9.27	1.00

Finally, given a multi-scene sequence, the error may occur in the border frame (usually intracoded). In such a case, MV based techniques fail since they try to extract the concealment information from the previous, and therefore uncorrelated, frame. Modified BMA and OBMA are able to gather the information from the current frame although the reconstructions tend to be of poor quality since both algorithms seek the best match for

the entire missing macroblock and this approach usually does not lead to the lower residual energy. Note that the H.264/AVC codec overcome this problem by allowing submacroblock prediction. Moreover, OBMA cannot be applied for all the slicing modes, e.g., this method is unable to conceal the chessboard loss pattern utilizing only the spatial information. On the contrary, due to the sequential filling and the dynamic adaptation to the available information, none of the aforementioned scenarios is an issue for our proposal in the combined SEC/TEC mode.

VI. Conclusions

We have developed a sparse linear prediction estimator, which recovers lost regions in images by filling them sequentially with a weighted combination of patches that are extracted from the available neighborhood. The weights are obtained by solving a convex optimization problem that arises from a spatial image model. Moreover, we show that the weights can be approximated by an exponential function, so that the resulting method can be alternatively interpreted as a kernel-based Nadaraya-Watson regression. The proposed techniques automatically adapt themselves to SEC, TEC or a combined scenario and can be thus successfully applied to both still images and video sequences.

Our proposals achieve better PSNR and perceptual reconstruction quality than other state-of-the-art techniques. SLP-C is optimized for squared error so it achieves better PSNR than the approximated method. Simulations reveal, however, that SLP-E provides better MS-SSIM. Finally, by applying the approximated algorithm SLP-E the processing time is reduced in a factor of 100.

References

[1] *ITU-T*, ITU-T Recommendation H.264, International Telecommunication Union, 2010.

[2] T. Wiegand, G. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 3, pp. 560–576, Mar. 2003.

[3] I. Richardson, *The H.264 Advanced Video Compression Standard.* New York, NY, USA: Wiley, 2010.

[4] A. M. Gomez, A. Peinado, V. Sanchez, and A. Rubio, "Combining media specific FEC and error concealment for robust distributed speech recognition over loss-prone packet channels," *IEEE Trans. Multimedia*, vol. 8, no. 4, pp. 1228–1238, Nov. 2006.

[5] A. Peinado, V. Sánchez, and A. Gómez, "Error concealment based on MMSE estimation for multimedia wireless and IP applications," in *Proc. PIMRC*, Sep. 2008, pp. 1–5.

[6] P. Salama, N. Shroff, E. Coyle, and E. Delp, "Error concealment techniques for encoded video streams," in *Proc. ICIP*, 1995, pp. 9–12.

[7] Y. Zhao, H. Chen, X. Chi, and J. Jin, "Spatial error concealment using directional extrapolation," in *Proc. DICTA*, 2005, pp. 278–283.

[8] H. Sun and W. Kwok, "Concealment of damaged block transform coded images using projections onto convex sets," *IEEE Trans. Image Process.*, vol. 4, no. 4, pp. 470–477, Apr. 1995.

[9] D. Robie and R. Mersereau, "The use of Hough transforms in spatial error concealment," in *Proc. ICASSP*, 2000, vol. 4, pp. 2131–2134.

[10] H. Gharavi and S. Gao, "Spatial interpolation algorithm for error concealment," in *Proc. ICASSP*, Apr. 2008, pp. 1153–1156.

[11] S. Shirani, F. Kossentini, and R. Ward, "An adaptive Markov random field based error concealment method for video communication in error prone environment," in *Proc. ICIP*, 1999, vol. 6, pp. 3117–3120.

[12] W. Kung, C. Kim, and C. Kuo, "Spatial and temporal error concealment techniques for video transmission over noisy channels," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 7, pp. 789–802, Jul. 2006.

[13] Z. Rongfu, Z. Yuanhua, and H. Xiaodong, "Content-adaptive spatial error concealment for video communication," *IEEE Trans. Consumer Electron.*, vol. 50, pp. 335–341, Feb. 2004.

[14] P. Harrison, "Texture synthesis, texture transfer and plausible restoration," Ph.D. dissertation, Dept. Inf. Technol., Monash Univ., Victoria, Australia, 2005.

[15] A. Criminisi, P. Pérez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," *IEEE Trans. Image Process.*, vol. 13, no. 9, pp. 1200–1212, Sep. 2004.

[16] X. Li and M. Orchard, "Novel sequential error-concealment techniques using orientation adaptive interpolation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 8, pp. 857–864, Oct. 2002.

[17] G. Zhai, X. Yang, W. Lin, and W. Zhang, "Bayesian error concealment with DCT pyramid for images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 9, pp. 1224–1232, Sep. 2010.

[18] G. Zhai, J. Cai, W. Lin, X. Yang, and W. Zhang, "Image error-concealment via block-based bilateral filtering," in *Proc. ICME*, Jun. 2008, pp. 621–624.

[19] J. Seiler and A. Kaup, "Fast orthogonality deficiency compensation for improved frequency selective image extrapolation," in *Proc. ICASSP*, Mar. 2008, pp. 781–784.

[20] J. Model, "The JVT Reference Software for H.264/AVC." [Online]. Available: http://iphome.hhi.de/suehring/tml/download

[21] T. Thaipanich, P. Wu, and C. J. Kuo, "Video error concealment with outer and inner boundary matching algorithms," in *Proc. SPIE*, 2007.

[22] K. Song, T. Chung, C.-S. Kim, Y.-O. Park, Y. Kim, Y. Joo, and Y. Oh, "Efficient multi-hypothesis error concealment technique for H.264," in *Proc. ISCAS*, May 2007, pp. 973–976.

[23] J. Zhou, B. Yan, and H. Gharavi, "Efficient motion vector interpolation for error concealment of H.264/AVC," *IEEE Trans. Broadcasting*, vol. 57, no. 1, pp. 75–80, Mar. 2011.

[24] M. Ma, O. Au, S. G. Chan, and M. Sun, "Edge-directed error concealment," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 3, pp. 382–394, Mar. 2010.

[25] Y. Chen, Y. hu, O. Au, H. Li, and C. Chen, "Video error concealment using spatio-temporal boundary matching and partial differential equation," *IEEE Trans. Multimedia*, vol. 10, no. 1, pp. 2–11, Jan. 2008.

[26] S. Shirani, F. Kossentini, and R. Ward, "A concealment method for video communications in an error-prone environment," *IEEE J. Sel. Areas Commun.*, vol. 18, no. 6, pp. 1122–1128, Jun. 2000.

[27] D. Persson, T. Eriksson, and P. Hedelin, "Packet video error cocnealment with Gaussian mixture models," *IEEE Trans. Image Process.*, vol. 17, no. 1, pp. 145–154, Jan. 2008.

[28] D. Persson and T. Eriksson, "Mixture model-and least squares-based packet video error concealment," *IEEE Trans. Image Process.*, vol. 18, no. 5, pp. 1048–1054, May 2009.

[29] D. Nguyen, M. Dao, and T. Tran, "Video error concealment using sparse recovery and local dictionaries," in *Proc. ICASSP*, May 2011, pp. 1125–1128.

[30] Y. Zhang, X. Xiang, D. Zhao, S. Ma, and W. Gao, "Packet video error concealment with auto regressive model," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 1, pp. 12–27, Jan. 2012.

[31] J. Koloda, J. Østergaard, S. Jensen, A. Peinado, and V. Sánchez, "Sequential error concealment for video/images via weighted template matching," in *Proc. DCC*, 2012, pp. 159–170.

[32] D. Giacobello, M. Christensen, M. Murthi, S. Jensen, and M. Moonen, "Sparse linear prediction and its applications to speech processing," *IEEE Trans. Audio, Speech Language Process.*, vol. 20, no. 5, pp. 1644–1657, Jul. 2010.

[33] D. Donoho and Y. Tsaig, "Fast solution of L1-norm minimization problems when the solution may be sparse," *IEEE Trans. Inf. Theory*, vol. 54, no. 11, pp. 4789–4812, Nov. 2008.

[34] R. Little and D. Rubin, *Statistical Analysis With Missing Data*. New York, NY, USA: Wiley, 1987.

[35] J. Romberg, "Imaging via compressive sensing," *IEEE Signal Process. Mag.*, vol. 25, no. 2, Mar. 2008.

[36] L. Vandenberghe and S. Boyd, "Semidefinite programming," *Soc. Ind. Appl. Math.*, vol. 38, pp. 49–95, Mar. 1996.

[37] X. Zhang, Y. Zhang, D. Zhao, and S. M. Gao, "A high efficient error concealment scheme based on auto-regressive model for video coding," in *Proc. PCS*, 2009.

[38] Z. Wang, E. Simoncelli, and A. Bovik, "Multi-scale structural similarity for image quality assessment," *IEEE Signals, Syst. Comput.*, vol. 2, pp. 1398–1402, Nov. 2003.

[39] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assesment: From error visibility to structural visibility," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[40] J. Østergaard, M. Derpich, and S. Channappayya, "The high-resolution rate-distortion function under the structural similarity index," *EURASIP J. Adv. Signal Process.*, vol. 2011, 2011.

[41] A. Brooks, X. Zhao, and T. Pappas, "Structural similarity quality metrics in a coding context: Exploring the space of realistic distorsions," *IEEE Trans. Image Process.*, vol. 17, no. 8, pp. 1261–1273, Aug. 2008.

[42] V. Varsa and M. Hannuksela, "Non-normative error concealment algorithms," *ITU-T SG16, VCEG-N62*, vol. 50, Sep. 2001.

[43] [Online]. Available: http://dtstc.ugr.es/~jkoloda/download.html

Ján Koloda received the M.Sc. degree in telecommunications engineering from the University of Granada, Granada, Spain, in 2009. He is currently working towards the Ph.D. degree on error concealment algorithms for block-coded video. Since 2010, he has been with the Research Group on Signal Processing, Multimedia Transmission and Speech/Audio Technologies, Department of Signal Theory, Networking and Communications at the University of Granada, under a research grant. He has been a visiting researcher at Aalborg University, Aalborg, Denmark. His research interests are in the area of error concealment of block-coded video sequences, image and signal processing.

Jan Østergaard received the M.Sc. degree in Electrical Engineering from Aalborg University, Aalborg, Denmark, in 1999 and the Ph.D. degree (cum laude) from Delft University of Technology, Delft, The Netherlands, in 2007. From 1999 to 2002, he worked as an R&D Engineer at ETI A/S, Aalborg, Denmark, and from 2002 to 2003, he worked as an R&D Engineer at ETI Inc., Virginia, United States. Between September 2007 and June 2008, he worked as a post-doctoral researcher at The University of Newcastle, NSW, Australia. From June 2008 to March 2011, he worked as a post-doctoral researcher/Assistant Professor at Aalborg University and he is currently Associate Professor at Aalborg University. He has also been a visiting researcher at Tel Aviv University, Tel Aviv, Israel, and at Universidad Técnica Federico Santa María, Valparaíso, Chile. He has received a Danish Independent Research Council's Young Researcher's Award and a post-doctoral fellowship from the Danish Research Council for Technology and Production Sciences.

Søren Holdt Jensen received the M.Sc. degree in electrical engineering from Aalborg University, Aalborg, Denmark, in 1988, and the Ph.D. degree in signal processing from the Technical University of Denmark, Lyngby, Denmark, in 1995. Before joining the Department of Electronic Systems of Aalborg University, he was with the Telecommunications Laboratory of Telecom Denmark, Ltd, Copenhagen, Denmark; the Electronics Institute of the Technical University of Denmark; the Scientific Computing Group of Danish Computing Center for Research and Education, Lyngby; the Electrical Engineering Department of Katholieke Universiteit Leuven, Leuven, Belgium; and the Center for PersonKommunikation (CPK) of Aalborg University. He is Full Professor and is currently heading a research team working in the area of numerical algorithms, optimization, and signal processing for speech and audio processing, image and video processing, multimedia technologies, and digital communications. Prof. Jensen was an Associate Editor for the IEEE Transactions on Signal Processing and Elsevier Signal Processing, and is currently Associate Editor for the IEEE Transactions on Audio, Speech and Language Processing. He is a recipient of an European Community Marie Curie Fellowship, former Chairman of the IEEE Denmark Section and the IEEE Denmark Section's Signal Processing Chapter. He is member of the Danish Academy of Technical Sciences and was in January 2011 appointed as member of the Danish Council for Independent Research—Technology and Production Sciences by the Danish Minister for Science, Technology and Innovation.

Victoria Sánchez received the M.S. and the Ph.D. degrees from the University of Granada, Granada, Spain, in 1988 and 1995, respectively. In 1988, she joined the Signal Processing and Communications department of the University of Granada where she is currently a member of the research group on Signal Processing, Multimedia Transmission and Speech/Audio Technologies. During 1991, she was visiting with the Electrical Engineering Department, University of Sherbrooke, Canada. Since 1997, she is an Associate Professor at the University of Granada. Her research interests include speech and audio processing, multimedia transmission and speech recognition. She has authored over 60 journal articles and conference papers in these fields.

Antonio M. Peinado received the M.S. and the Ph.D. degrees in Physics from the University of Granada, Granada, Spain, in 1987 and 1994, respectively. Since 1988, he has been working at the University of Granada, where he has led several research projects related to signal processing and communications. In 1989, he was a Consultant at the Speech Research Department, AT&T Bell Labs. He earned the positions of Associate Professor (1996) and Full Professor (2010) in the Department of Signal Theory, Networking and Communications, University of Granada, and is currently director of the research group on Signal Processing, Multimedia Transmission and Speech/Audio Processing (SigMAT) at the same university. He is the author of numerous publications and coauthor of the book Speech Recognition over Digital Channels (Wiley, 2006), and has served as reviewer for international journals, conferences and project proposals. His current research interests are focused on robust speech recognition and transmission, robust video transmission, and ultrasound signal processing.

2.2.1.2 Speech Reconstruction by Sparse Linear Prediction

- J. Koloda, A.M. Peinado and V. Sánchez, "Speech Reconstruction by Sparse Linear Prediction", *IberSPEECH*, selected for publication in *Communications in Computer and Information Science (Springer)*, pp. 247-256, Madrid, Spain, November 2012.
    - Status: Published.

Speech Reconstruction by Sparse Linear Prediction*

Ján Koloda, Antonio M. Peinado,
and Victoria Sánchez

Dpt. Teoría de la Señal, Telemática y Comunicaciones,
Centro de Investigación en Tecnologías de la Información y de las Comunicaciones,
18071-Granada, Spain
{janko,amp,victoria}@ugr.es
http://tstc.ugr.es, http://citic.ugr.es

Abstract. This paper proposes a new variant of the least square autoregressive (LSAR) method for speech reconstruction, which can estimate via least squares a segment of missing samples by applying the linear prediction (LP) model of speech. First, we show that the use of a single high-order linear predictor can provide better results than the classic LSAR techniques based on short- and long-term predictors without the need of a pitch detector. However, this high-order predictor may reduce the reconstruction performance due to estimation errors, especially in the case of short pitch periods, and non-stationarity. In order to overcome these problems, we propose the use of a sparse linear predictor which resembles the classical speech model, based on short- and long-term correlations, where many LP coefficients are zero. The experimental results show the superiority of the proposed approach in both signal to noise ratio and perceptual performance.

Keywords: Speech reconstruction, error concealment, sparse linear prediction, least squares, autoregressive model.

1 Introduction

Speech Reconstruction is a subject that has been widely treated in the speech community and which has a number of applications. Thus, we can mention audio restoration, where short signal segments completely degraded must be recovered from adjacent segments as it frequently occurs in old recordings. Also, in Voice-over-IP (VoIP) systems based on intraframe codecs, the real time constraints imposed by the transmission protocols may cause a packet loss problem which finally results in the loss of speech segments.

In order to perform the reconstruction of a lost signal segment, some sort of sample interpolation or extrapolation using adjacent and correctly received samples must be applied [1]. This can be a difficult task. Fortunately, in the case of speech there exists a well-known signal production model based on linear prediction (LP)

* This work has been supported by the Spanish MEC/FEDER project TEC 2010-18009.

D.T. Toledano et al. (Eds.): IberSPEECH 2012, CCIS 328, pp. 247–256, 2012.

which is employed by many reconstruction methods. Thus, we have the least square autoregressive (LSAR) method [2], which carries out an iterative interpolation of the lost samples from the adjacent ones applying the LP model and a least squares (LS) estimation. Other methods also based on LP focus on the estimation of the LP excitation (LP residual) [3,4]. Also, the LP spectrum has been combined with sinusoidal models of the excitation for signal extrapolation [5].

In this paper we will focus on the class of LSAR signal interpolators, where the missing samples are LS-estimated according to a previous estimation of the LP model. Although the basic LSAR [2] just uses a short-term predictor, better results can be obtained when long-term prediction is also considered as it is common practice in speech coding [6]. A first drawback of this approach is that it requires the use of a pitch detector which may be affected by detection errors. This can be avoided using a single high-order predictor which accounts for both short- and long-term correlations. The prediction order must be large enough as to cover the longest possible correlations (due to the longest possible pitch). Although this approach increases the computational cost, we will show that it results in a better reconstruction performance.

The use of a single high-order predictor for LSAR is a simple and compact solution. However, it does not follow the classical speech model based on short- and long-term predictors. This involves that many LP coefficients that are forced to be zero by this speech model can have now non-zero values, which can be interpreted as a sort of estimation noise. Also, it must be considered that a high-order predictor may be more affected by non-stationarity. For example, and as it is shown later, this effect can degrade the performance for the case of relatively small pitch values since the LP order is likely much larger than necessary. This problem has been recently addressed by the application of sparse linear prediction (SLP) [7,8]. The SLP idea consists in the optimization of a single high-order linear predictor which maintains as much as possible the high sparsity level involved by the classical speech model. The underlying philosophy of SLP is that of predicting the missing samples by employing as few adjacent samples as possible. This idea has already been successfully applied by the authors to video packet loss concealment [9] and will be adapted here to speech reconstruction by LSAR methods.

The paper is organized as follows. Section 2 is devoted to the review and analysis of LSAR techniques. Then, the proposed SLP method is developed in Section 3 and the simulation results are shown and commented in Section 4. Finally, the main conclusions are summarized.

2 Least Square Autoregressive (LSAR) Interpolation

Let us review now the basic LSAR interpolation algorithm of reference [2]. According to the linear prediction model of speech signals, a sample $x(m)$ is modeled as,

$$x(m) = \sum_{k=1}^{P} a_k x(m-k) + e(m) \tag{1}$$

$$\begin{pmatrix} e(P) \\ e(P+1) \\ \vdots \\ e(k-1) \\ \hline e(k) \\ e(k+1) \\ e(k+2) \\ \vdots \\ e(k+M+P-2) \\ e(k+M+P-1) \\ \hline e(k+M+P) \\ e(k+M+P+1) \\ \vdots \\ e(N-1) \end{pmatrix} = \begin{pmatrix} x(P) \\ x(P+1) \\ \vdots \\ x(k-1) \\ \hline x_{Uk}(k) \\ x_{Uk}(k+1) \\ x_{Uk}(k+2) \\ \vdots \\ x(k+M+P-2) \\ x(k+M+P-1) \\ \hline x(k+M+P) \\ x(k+M+P+1) \\ \vdots \\ x(N-1) \end{pmatrix} - \begin{pmatrix} x(P-1) & x(P-2) & \dots & x(0) \\ x(P) & x(P-1) & \dots & x(1) \\ \vdots & \vdots & \ddots & \vdots \\ x(k-2) & x(k-3) & \dots & x(k-P-1) \\ \hline x(k-1) & x(k-2) & \dots & x(k-P) \\ x_{Uk}(k) & x(k-1) & \dots & x(k-P+1) \\ x_{Uk}(k+1) & x_{Uk}(k) & \dots & x(k-P+2) \\ \vdots & \vdots & \ddots & \vdots \\ x(k+M+P-3) & x(k+M+P-2) & \dots & x_{Uk}(k+M-2) \\ x(k+M+P-2) & x(k+M+P-1) & \dots & x_{Uk}(k+M-1) \\ \hline x(k+M+P-1) & x(k+M+P) & \dots & x(k+M) \\ x(k+M+P) & x(k+M+P+1) & \dots & x(k+M+1) \\ \vdots & \vdots & \ddots & \vdots \\ x(N-2) & x(N-3) & \dots & x(N-P-1) \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_P \end{pmatrix}$$

Fig. 1. Matrix form of the residual for the LSAR algorithm

where a_k are the model coefficients and $e(m)$ is a zero mean excitation signal.

Let us assume that a received signal segment $\boldsymbol{x} = (x(0), x(1), ..., x(N-1))^T$ contains a series of lost (unknown) samples $\boldsymbol{x}_{Uk} = (x(k), ..., x(k+M-1))$. The objective is to reconstruct the missing samples $\boldsymbol{x}_{Uk}$ using the remaining known samples and the LP model of the signal (1). Rearranging the LP model and expanding it to a matrix notation we obtain the formulation displayed in Fig. 1, which can be rewritten in a compact notation as,

$$\boldsymbol{e}(\boldsymbol{x}_{Uk}, \boldsymbol{a}) = \boldsymbol{x} - \boldsymbol{X}\boldsymbol{a} \tag{2}$$

The missing samples $\boldsymbol{x}_{Uk}$ are then reconstructed by minimizing the squared error expressed as

$$\varepsilon = \|\boldsymbol{e}\|_2^2 = \boldsymbol{e}^T\boldsymbol{e} = \boldsymbol{x}^T\boldsymbol{x} + \boldsymbol{a}^T\boldsymbol{R}_x\boldsymbol{a} - 2\boldsymbol{a}^T\boldsymbol{r}_x \tag{3}$$

where $\boldsymbol{R}_x = \boldsymbol{X}^T\boldsymbol{X}$ and $\boldsymbol{r}_x = \boldsymbol{X}^T\boldsymbol{x}$. Note that ε is a function of two unknown variables, the predictor coefficients $\boldsymbol{a}$ and the unknown segment $\boldsymbol{x}_{Uk}$, whose reconstruction is the objective of this problem. Since (3) involves unknown terms of fourth and cubic order, solving the problem by differentiating ε with respect to the unknown vectors $\boldsymbol{x}_{Uk}$ and $\boldsymbol{a}$ would be mathematically impractical. An estimation-maximization (EM) procedure is used instead. First, Eq. (2) is linearized by setting the unknown samples to zero (estimation). This makes the squared error $\boldsymbol{e}$ to be a function of the LP-coefficients $\boldsymbol{a}$ only. The coefficients are then computed by minimizing ε, that is, by solving the usual set of normal equations, which yields,

$$\hat{\boldsymbol{a}} = \boldsymbol{R}_x^{-1}\boldsymbol{r}_x. \tag{4}$$

Finally, the unknown samples are reconstructed using the estimated LP coefficients. This approach can be iterated several times, although in most cases very few iterations are needed.

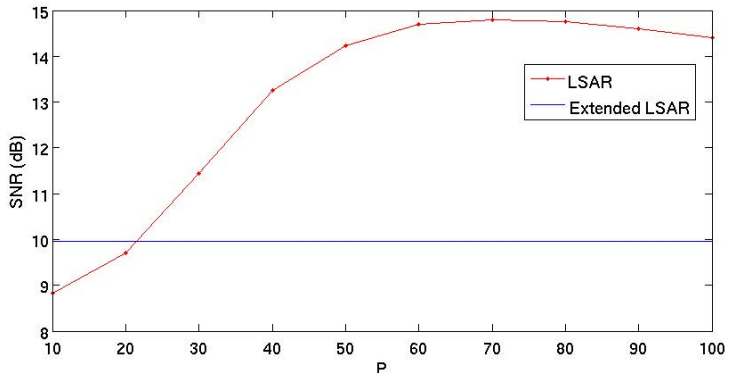


Fig. 2. SNR performance of LSAR (red) and Extended LSAR (blue). The Extended LSAR is applied with P = 10 and Q = 1.

Given that voiced speech signals are quasi-periodic, a speech sample is highly correlated with the neighboring ones as well as with the samples shifted by one (or several) pitch period. In order to exploit these longer correlations, a modification of the basic LSAR (Extended LSAR) which introduces a long-term predictor was proposed in [6]. The speech model involved by the Extended LSAR is,

$$x(m) = \sum_{k=1}^{P} a_k x(m-k) + \sum_{k=-Q}^{Q} p_k x(m-T-k) + e(m), \tag{5}$$

where Q is the order of the long term LP and T is the pitch period. This is the underlying speech model employed somehow by many speech codecs and can be solved again through the corresponding set of normal equations. An interesting feature of this model is that we can consider that equation (5) contains a single predictor with a high level of sparsity. This feature will be exploited in our proposal.

As mentioned in the introduction section, the long-term correlations can be also exploited by the basic LSAR if a prediction order P, large enough to cover the longest possible correlations, is used. The main advantage of this solution is that no pitch estimation is needed.

In order to asses both the basic LSAR and the Extended LSAR, Fig. 2 shows the average SNR values obtained by both techniques for gaps of 6 ms separated 30 ms. The corresponding experimental setup will be described in Section 4. The basic LSAR performance is plotted versus the LP order P, while the extended one is only shown for typical values ($P = 10$, $Q = 1$, 13 coefficients). A first comparison can be made for this typical LP orders. In this case, the Extended LSAR not only outperforms LSAR for $P = 10$, but also for 20 coefficients. This makes clear the need of including long-term correlations. However, it is also observed that the performance can be meaningfully increased with the basic LSAR by simply increasing the LP order. The order increase does not make sense for the Extended LSAR since this would simply imply that the short-term predictor would *absorb* the long-term one.

The main conclusion that can be extracted from the above discussion is that the basic LSAR must be employed if there are no strong computational constraints. However, it still has two problems:

1. Many LP coefficients, which are forced to be zero when the classical speech model is applied, can have now non-zero values. In principle, this may especially affect the inter-pitch and post-pitch coefficients and could be interpreted as a sort of estimation noise.
2. When a large order P is applied, the estimated coefficients can be more affected by the non-stationarity of the speech signal since more autocorrelation coefficients are used in (4).

The effect of these problems over the SNR plot of Fig. 2 is a SNR decay for the higher LP orders. In average, this decay starts after the average pitch value of the speech corpus (57.80 samples).

In this paper, we propose a modification of the LSAR algorithm oriented to mitigate the above problems by applying sparse linear prediction (SLP) for LP predictor estimation. We can consider that this proposal combines the best features of the basic LSAR with large P and Extended LSAR since it uses a single compact predictor which does not require pitch estimation and tries to keep the sparsity of Extended LSAR. SLP and the proposed modification to LSAR are presented in the next section.

3 LSAR by Sparse Linear Prediction

As discussed in the previous section, our goal is the development of a new variant of LSAR with a single large-order predictor which is, at the same time, highly sparse. Thus, we have to minimize the squared error in (3), with respect to $\boldsymbol{a}$, with a sparsity constraint, that is,

$$\begin{aligned} &\text{minimize } \epsilon(\boldsymbol{a}) = \left\|\boldsymbol{a}^T \boldsymbol{R}_x \boldsymbol{a} - 2\boldsymbol{a}^T \boldsymbol{r}_x\right\|_2^2 \\ &\text{subject to } \|\boldsymbol{a}\|_0 \leq \delta_0. \end{aligned} \tag{6}$$

where the term $\boldsymbol{x}^T\boldsymbol{x}$ is not included in the optimization procedure since it comprises the DC component of the squared error. The main problem that arises when solving (6) is that the ℓ_0-norm is non convex so that the global minimum is usually found by exhaustive search and is therefore computationally prohibitive. This problem has been thoroughly studied in compressive sensing theory and can be efficiently solved by applying convex relaxation [10], i.e.

$$\begin{aligned} &\text{minimize } \epsilon(\boldsymbol{a}) = \left\|\boldsymbol{a}^T \boldsymbol{R}_x \boldsymbol{a} - 2\boldsymbol{a}^T \boldsymbol{r}_x\right\|_2^2 \\ &\text{subject to } \|\boldsymbol{a}\|_1 \leq \delta_1. \end{aligned} \tag{7}$$

The objective function, as well as the constraints, are both convex and the optimization problem can be efficiently solved by a convex optimization algorithm. In our simulations, we apply the primal-dual interior point (IP) method [11].

The LP-coefficients obtained in the previous step are then used to re-estimate the unknown samples $\boldsymbol{x}_{Uk}$. This is carried out by inserting the obtained coefficients $\boldsymbol{a}$ into Eq.(2) and minimizing the squared error ε with respect to $\boldsymbol{x}_{Uk}$, which is the only unknown variable the squared error now depends on. Note that only the equations within the dashed lines in Fig. 1 are involved in the minimization since the remaining ones are constant with respect to $\boldsymbol{x}_{Uk}$. These equations can be rearranged so that the excitation signal is a combination of known and unknown samples:

$$\boldsymbol{e} = \boldsymbol{A}_1\boldsymbol{x}_{Uk} + \boldsymbol{A}_2\boldsymbol{x}_{Kn} \tag{8}$$

where the matrices $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ are both constructed using the LP-coefficients $\boldsymbol{a}$ and $\boldsymbol{x}_{Kn} = (x(k-P), \ldots, x(k+M+P-1))^T$ (see ref. [2] for more details). The total squared error is then given by,

$$\|\boldsymbol{e}\|_2^2 = \boldsymbol{e}^T\boldsymbol{e} = (\boldsymbol{A}_1\boldsymbol{x}_{Uk} + \boldsymbol{A}_2\boldsymbol{x}_{Kn})^T(\boldsymbol{A}_1\boldsymbol{x}_{Uk} + \boldsymbol{A}_2\boldsymbol{x}_{Kn}) \tag{9}$$

The unknown samples $\boldsymbol{x}_{Uk}$ that minimize the squared error are obtained by setting the derivative of the squared error function with respect to $\boldsymbol{x}_{Uk}$ to zero

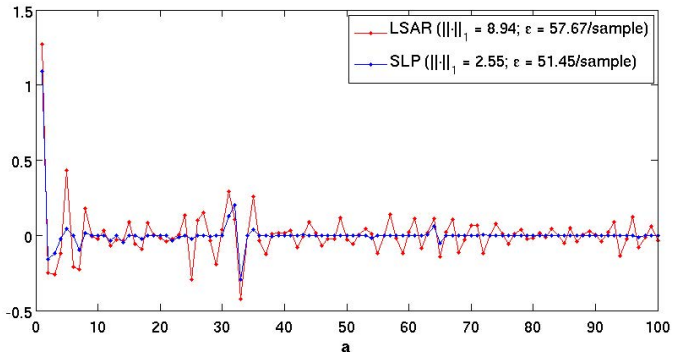
$$\frac{\partial \boldsymbol{e}^T\boldsymbol{e}}{\partial \boldsymbol{x}_{Uk}} = 2\boldsymbol{A}_1^T\boldsymbol{A}_1\boldsymbol{x}_{Kn} + 2\boldsymbol{A}_1^T\boldsymbol{A}_2\boldsymbol{x}_{Kn} \tag{10}$$

Finally, from Eq.(10) we have

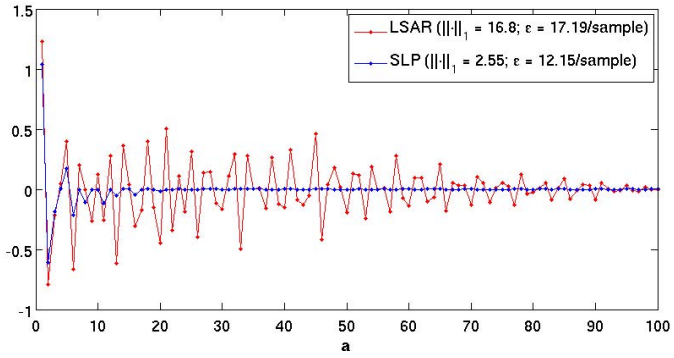
$$\hat{\boldsymbol{x}}_{Uk} = -(\boldsymbol{A}_1^T\boldsymbol{A}_1)^{-1}(\boldsymbol{A}_1^T\boldsymbol{A}_2)\boldsymbol{x}_{Kn} \tag{11}$$

The sparsity restriction does not make sense in this case since $\boldsymbol{x}_{Uk}$ is not sparse in general, although it can be solved via convex optimization with no restrictions.

In order to better illustrate the differences between the sparse approach and the classic LSAR, let us analyze two particular cases of missing segment reconstruction. The first case involves a voiced segment with pitch period equal to 32 samples. The pitch period is calculated over the clean (original) signal using the Yin pitch detector [12]. In the second case, an unvoiced segment is reconstructed. For both cases, we perform a reconstruction with 100 LP-coefficients using LSAR and the proposed technique. The results are shown in Fig. 3. Figure 3(a) shows the obtained coefficients for the voiced segment. As expected, the SLP-coefficients are much sparser than the coefficients obtained by LSAR while providing a reconstruction with lower squared error. Moreover, the significant elements of the LP-vector are concentrated around the position of 32, 64 and a small contribution around 96. Note that the pitch period of the original signal has been determined to be 32. Thus, the proposed SLP predictor adaptively encounters the pitch value. The case of the unvoiced segment reconstruction is shown in Fig. 3(b). Again, the LSAR coefficients vector is much less sparse while generating a reconstruction with larger squared error. In this case, SLP automatically concentrates the weights in the close proximity of the lost segment which is coherent with the assumption that in unvoiced segments the most correlated samples are the closest ones.



(a)



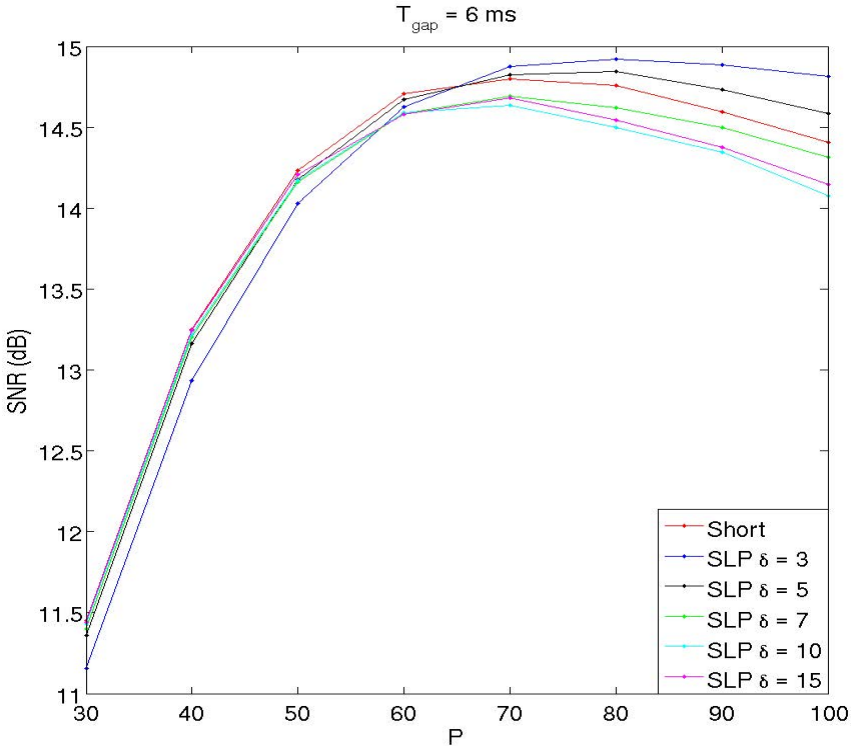
(b)

Fig. 3. Example of distribution of the LP-coefficients obtained by LSAR (red) and SLP (blue). (a) Voiced segment. (b) Unvoiced segment.

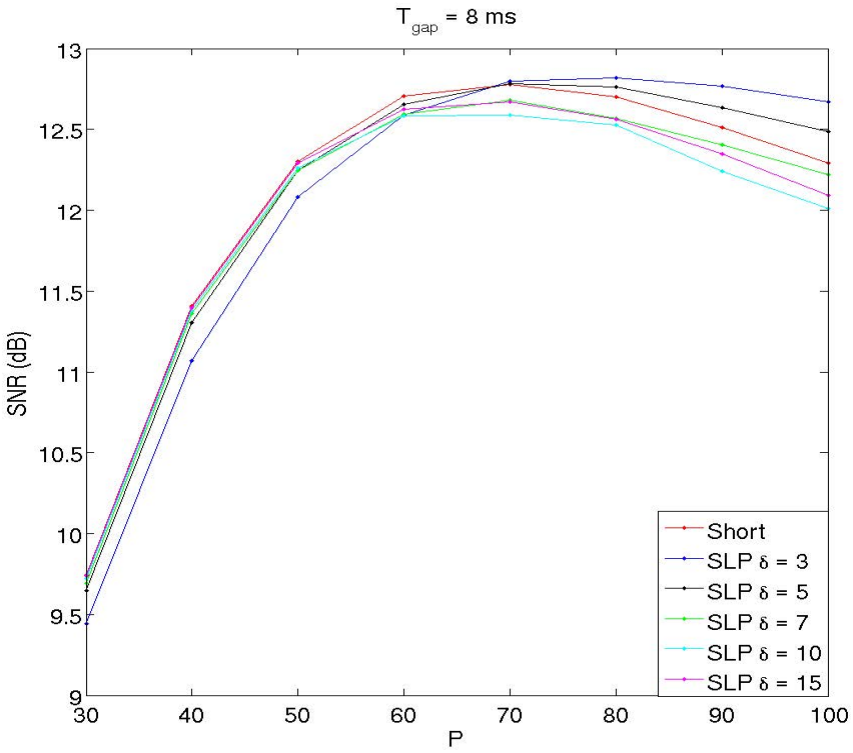
4 Simulation Results

The speech utterances used for testing comprise a subset of 400 sequences extracted from the geographic corpus of the *Albayzin* database [13]. All the speakers, used for recording the database, are also present in our tested subset. Figure 4 shows a comparison of LSAR and the proposed SLP-based algorithm in terms of SNR. The average SNR value over the 400 utterances is shown for different LP orders. The proposed SLP technique is also tested for different levels of sparsity (controlled with parameter δ_1). Moreover, the comparison is carried out for missing segment lengths (T_{gap}) of 6 ms and 8 ms. The losses are produced every 30 ms and a 32 ms window, centered over the missing segment, is used for estimating the predictor. The window is eventually extended (up to the required length) in cases where the sum of LP order and the duration of the gap is larger than the 32 ms window. Finally, two iterations are employed for both cases.

The simulations reveal that the larger the LP order, the sparser it should be in order to obtain better quality reconstructions. The average pitch period of the tested subset is 57,80 so, for shorter LP orders, there is no need to impose sparsity over the LP estimator. Note that weak sparsity restrictions approximate the LSAR behavior for low order LP estimators. For LP orders above the average pitch period, the performance of the LSAR technique starts to decay while



(a)



(b)

Fig. 4. Performance comparison, in terms of SNR, of LSAR and SLP with different values of δ. Both algorithms are tested for different values of LP order and two iterations are applied. (a) $T_{gap} = 6$ ms. (b) $T_{gap} = 8$ ms.

our proposal eventually rises and then practically maintains the reconstruction quality (with a very slight decay). This results confirm that the basic LSAR with large prediction order may be affected by noise estimation and non-stationarity and that the sparsity constraint helps to palliate these problems. Also, we can conclude that the sparsity parameter δ_1 could be set according to the LP order, although in this paper we focused on obtaining a fixed estimator suitable for the majority of pitch periods. Thus, a high order and highly sparse (small δ_1) LP estimator is preferred.

Table 1 shows the average values of SNR and PESQ (Perceptual Evaluation of Speech Quality) obtained for different lengths of lost segments. The LP order is set to 100 in order to include all possible pitch values in the database and 5 iterations are used. The proposed method outperforms the basic LSAR in all cases and the difference in perceptual quality has an increasing trend with the gap length.

Table 1. Average SNR and PESQ values for SLP and LSAR for different lost segment lengths. The simulation is carried out for P = 100 with 5 iterations.

SNR	$T_{gap} = 4$ms	$T_{gap} = 6$ms	$T_{gap} = 8$ms	$T_{gap} = 10$ms
SLP	18.10	15.13	13.03	11.38
LSAR	17.47	14.41	12.34	10.67
PESQ				
SLP	4.00	3.81	3.63	3.47
LSAR	3.91	3.72	3.51	3.34

5 Conclusions

We have proposed a modification of the LSAR speech reconstruction algorithm which uses sparse linear prediction. The proposed approach has several advantages as avoiding the use of pitch detectors, a better approximation to the sparse classical model employed in speech coding, a better behavior for large pitch values (reducing the estimation noise) and less sensitivity to non-stationarities. Applying convex relaxation allows to solve the minimization problem with sparsity constraint in a relatively efficient way. The proposed technique outperforms the classic LSAR both at objective and perceptual level. Future work includes the dynamic adaptation of the sparsity parameter δ_1 to the instantaneous pitch values and the LP order.

References

1. Vaseghi, S.: Multimedia signal processing. John Wiley (2007)
2. Janssen, A., Veldhuis, R., Vries, L.: Adaptive interpolation of discrete-time signals that can be modeled as AR processes. IEEE Transactions on Acoustics, Speech and Signal Processing, 317–330 (1986)

3. Jauppinen, I., Roth, K.: Audio signal restoration - theory and applications. In: Proceedings of the 5th Int. Conf. on Digital Audio Effects (2002)
4. Esquef, P., Biscainho, L.: An efficient model-based multirate method for reconstruction of audio signals along long gaps. IEEE Transactions on Speech and Audio Processing 14, 1391–1400 (2006)
5. Lindblom, J., Hedelin, P.: Packet loss concealment based on sinusoidal modelling. In: Proceedings of ICASSP 2002 (2002)
6. Vaseghi, S., Rayner, P.: Detection and suppression of impulsive noise in speech communication systems. IEE Proceedings 1, 38–46 (1990)
7. Giacobello, D., Christensen, M., Dahl, J., Jensen, S., Moonen, M.: Sparse linear prediction of speech. In: Proceedings of Interspeech 2008 (2008)
8. Giacobello, D., Christensen, M., Murthi, M., Jensen, S., Moonen, M.: Speech coding based on sparse linear prediction. In: Proceedings of Eusipco 2009 (2009)
9. Koloda, J., Østergaard, J., Jensen, S., Peinado, A., Sanchez, V.: Sequential error concealment of video/images via weighted template matching. In: Proceedings of DCC 2012 (2012)
10. Romberg, J.: Imaging via compressive sensing. IEEE Signal Processing Magazine 25 (March 2008)
11. Vandenberghe, L., Boyd, S.: Semidefinite programming. Society for Industrial and Applied Mathematics (1996)
12. Cheveigné, A., Kawahara, H.: Yin, a fundamental frequency estimator for speech and music. Journal of the Acoustical Society of America 111(4), 1917–1930 (2002)
13. Díaz-Verdejo, J., Peinado, A., Rubio, A., Segarra, E., Prieto, N., Casacuberta, F.: ALBAYZIN: a task-oriented spanish speech corpus. In: First International Conference on Language Resources and Evaluation, vol. 1, pp. 487–502 (May 1998)

2.2.2 Multimedia Signal Reconstruction by Kernel-based MMSE

2.2.2.1 On the Application of Multivariate Kernel Density Estimation to Image Error Concealment

- J. Koloda, A.M. Peinado and V. Sánchez, "On the Application of Multivariate Kernel Density Estimation to Image Error Concealment", *in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1330-1334, Vancouver, Canada, May 2013.
    - Status: Published.

ON THE APPLICATION OF MULTIVARIATE KERNEL DENSITY ESTIMATION TO IMAGE ERROR CONCEALMENT

Ján Koloda, Antonio M. Peinado, Victoria Sánchez

Dpt. Teoría de la Señal, Telemática y Comunicaciones - CITIC
Universidad de Granada, Spain
{janko,amp,victoria}@ugr.es

ABSTRACT

This paper proposes a methodology for the application of multivariate kernel density estimation (KDE) to MMSE-based image/video error concealment (EC). We show that the estimation of the kernel bandwidth matrix for EC must follow a criterion different from that of typical KDE problems. In particular, we propose a bandwidth built as the product of a structure matrix and a scale factor obtained with a minimum square error criterion. We show that our proposal can achieve average PSNR improvements larger than 1 dB with respect to other state-of-the-art techniques.

Index Terms— kernel estimation, error concealment

1. INTRODUCTION

Achieving high QoS in multimedia applications is a very challenging task since the transmission of multimedia contents over error prone channels may lead to errors or data losses. The most advanced and utilized image and video coding systems (JPEG, H.264/AVC, etc.) are block-based so these errors result in a loss of one or several macroblocks. In order to mitigate the effect of these losses, error concealment (EC) algorithms can be applied at the decoder. They take advantage of spatial and/or temporal correlations within the received stream to recover the missing data. For image communication or video transmission, when temporal information is not available or relevant, only spatial EC (SEC) is applicable.

A simple and common SEC technique is bilinear interpolation [1] which is defined as the default SEC method in the H.264/AVC codec. In order to better preserve important visual features, such as edges, a more advanced technique based on Markov random fields was proposed in [2]. In [3], a sequential pixel-wise method that draws on orientation adaptive interpolation was introduced. Bilateral filtering that exploits a pair of gaussian kernels is treated in [4]. A switching content adaptive SEC algorithm was proposed in [5]. Inpainting methods have also been successfully applied to EC problems [6]. A Hough transform based technique that aims at recovering edges based on their visual properties was proposed in [7]. Also, SEC techniques in a transformed domain have been recently proven to produce high-quality reconstructions [8].

In our previous paper [9] we proposed an EC technique which estimates a lost group of pixels (patch) through linear prediction (LP). This method provided better results than other state-of-the-art techniques such as [1]-[8]. The LP predictor is obtained by minimizing the square error between a context vector containing the available pixels around the missing patch and a linear combination of context vectors taken from the neighbourhood. This optimization was carried out under constraints of non-negativity and sparsity via convex relaxation. We also showed that the resulting estimation could be approximated by a multivariate Nadaraya-Watson regression with a Gaussian kernel [10]. This kernel-based view of sparse linear prediction offers a number of advantages. In particular, it can be interpreted as a minimum mean square error (MMSE) estimation where the required probabilities have been obtained through kernel density estimation (KDE) [11]. In this paper, we will exploit and generalize this new point of view which will allow us to apply powerful Bayesian tools to EC. Moreover, we will see that the goal of signal reconstruction is quite different from that of regression. Since the main problem in KDE is the estimation of the bandwidth matrix H, this means that H must be computed with a criterion different from the one usually applied for KDE or regression. Thus, we will propose a method to obtain the bandwidth which is specifically conceived for reconstruction.

The paper is organized as follows. The EC framework is detailed in Section 2. The proposed algorithm is described in Sections 3 and 4. Simulations results are discussed in Section 5. The last section is devoted to conclusions.

2. PREVIOUS WORK AND CONCEALMENT FRAMEWORK

The concealment framework used along this paper will be the same as that of reference [9]. In the following, we briefly summarize it. Let $\mathcal{L}$ be the set of missing pixels. Our goal

This work has been supported by an FPU grant from the Spanish Ministry of Education and by the MICINN TEC2010-18009 project.

is the prediction of a vector $\mathbf{z}_0 = (\mathbf{x}_0^t, \mathbf{y}_0^t)^t$, where $\mathbf{x}_0$ is a patch of lost pixels in $\mathcal{L}$ and $\mathbf{y}_0$ contains a set of (adjacent and available) context pixels. Let $\mathcal{S}$ be the set of available pixels which can be employed for prediction. We will consider all the possible vectors $\mathbf{z}_j$ ($j = 1, \ldots, M$) that can be built in $\mathcal{S}$ with the same shape and dimensionality as $\mathbf{z}_0$ (that is, $\mathbf{z}_j = (\mathbf{x}_j^t, \mathbf{y}_j^t)^t$). Then, the LP estimator for $\mathbf{x}_0$ can be written as,

$$\hat{\mathbf{x}}_0 = \sum_{j=1}^{M} w_j \mathbf{x}_j, \tag{1}$$

where $\mathbf{w} = (w_1, \ldots, w_M)^t$ is the vector of LP coefficients.

We consider a block-based codec where the missing region $\mathcal{L}$ is a 16×16 macroblock and the support area $\mathcal{S}$ comprises all the available pixels within the neighbouring macroblocks around $\mathcal{L}$. In this paper, we will employ an error pattern as shown in Fig. 3(a) which corresponds to a rate of block loss of approximately 25% with dispersed slicing structure [12]. Note, however, that our technique can be straightforwardly extended to other error patterns. We will also consider 2×2 patches $\mathbf{x}_0$ of missing pixels and the corresponding context $\mathbf{y}_0$ will comprise all the available pixels within the 6×6 pixel neighbourhood centred in $\mathbf{x}_0$. Vectors $\mathbf{z}_j$ replicate the shape of $\mathbf{z}_0$. These configurations are shown in Fig. 1(a). Moreover, macroblocks are concealed sequentially from the outer layer towards the centre (see Fig.1(b)). This filling order is based on a reliability parameter and it is detailed in [9].

In our previous work [9], the weights w_j of Eq. (1) are obtained by minimizing the following square error,

$$\epsilon_{\mathbf{y}}(\mathbf{w}; \mathbf{y}_0) = \left\| \mathbf{y}_0 - \sum_{j=1}^{M} w_j \mathbf{y}_j \right\|_2^2 \tag{2}$$

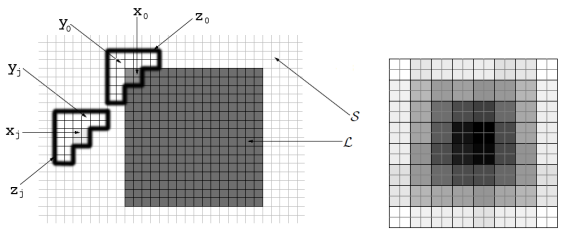
along with non-negativity ($\mathbf{w} \succeq 0$) and sparsity constraints. In [9], we also showed that these LP weights could be approximated through the following exponential function,

$$w_j = C \exp\left(-\frac{1}{2} \frac{\|\mathbf{y}_0 - \mathbf{y}_j\|^2}{m\sigma^2} \right), \tag{3}$$

where σ^2 is a decay factor ($\sigma^2 = 10$ in [9]), m is the dimensionality of the context vectors, and C is a normalization factor so that $\sum_j w_j = 1$. The resulting estimation can be viewed as a particular form of Nadaraya-Watson regression which employs a multivariate Gaussian kernel with a scalar bandwidth $h = \sqrt{m\sigma^2}$. This new kernel-based point of view is exploited in the following section.

3. THE KERNEL-BASED APPROACH

Kernel density estimation (KDE) is a non-parametric way for the estimation of the probability density function (pdf) associated to a given random process from a set of observations. In our case, we are interested in the pdf of $\mathbf{z} = (\mathbf{x}^t, \mathbf{y}^t)^t$ from the set of observations $\{\mathbf{z}_j; j = 1, \ldots, M\}$. The corresponding



(a) (b)

Fig. 1. (a) Example of configuration for the vectors $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$. $\mathcal{S}$ denotes the set of known pixels and $\mathcal{L}$ denotes the set of lost pixels. (b) Filling order for sequential reconstruction with 2×2 patches. The regions illustrated by brighter level are recovered first.

KDE estimate can be written as,

$$p(\mathbf{z}) = \frac{1}{M} \sum_{j=1}^{M} \frac{1}{|H|} K\left(H^{-1}(\mathbf{z} - \mathbf{z}_j)\right) = \frac{1}{M} \sum_{j=1}^{M} K_Z^{(j)}(\mathbf{z}). \tag{4}$$

where $K(\mathbf{u}) = \exp(-\mathbf{u}^t\mathbf{u}/2)/\sqrt{2\pi}$ is the (Gaussian) kernel employed and H is the bandwidth matrix. A more convenient form of the KDE estimator is given in the last part of Eq. (4), where $p(\mathbf{z})$ adopts the form of a Gaussian mixture model (GMM) and $K_Z^{(j)}(\mathbf{z})$ represents a multivariate Gaussian with mean $\mathbf{z}_j$ and covariance $\mathcal{H}$ which can be decomposed as,

$$\mathcal{H} = HH^t = \begin{pmatrix} \mathcal{H}_{XX} & \mathcal{H}_{XY} \\ \mathcal{H}_{YX} & \mathcal{H}_{YY} \end{pmatrix}. \tag{5}$$

In the following, we will also refer to $\mathcal{H}$ as bandwidth matrix.

Once $p(\mathbf{z})$ has been obtained, different Bayesian estimation techniques can be carried out. In particular, we are interested in the MMSE estimator of $\mathbf{x}_0$ given $\mathbf{y}_0$. Since the KDE estimate has the form of a GMM, we can adapt the well-known MMSE estimation formulae for GMM models [13], obtaining

$$\hat{\mathbf{x}}_0 = E[\mathbf{x}|\mathbf{y}_0] = \textstyle\sum_{j=1}^{M} w_j(\mathbf{y}_0) \boldsymbol{\mu}_{X|Y}^{(j)}(\mathbf{y}_0) \tag{6}$$

$$w_j(\mathbf{y}_0) = \frac{K_Y^{(j)}(\mathbf{y}_0)}{\sum_{i=1}^{M} K_Y^{(i)}(\mathbf{y}_0)} \tag{7}$$

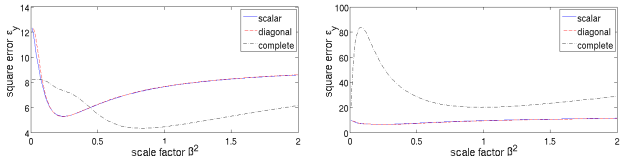
$$\boldsymbol{\mu}_{X|Y}^{(j)}(\mathbf{y}_0) = E[\mathbf{x}|\mathbf{y}_0, \mathbf{y}_j] = \mathbf{x}_j + \mathcal{H}_{XY}\mathcal{H}_{YY}^{-1}(\mathbf{y}_0 - \mathbf{y}_j) \tag{8}$$

where $K_Y^{(j)}(\mathbf{y})$ represents a multivariate Gaussian with mean $\mathbf{y}_j$ and covariance $\mathcal{H}_{YY}$. The estimator just derived can be interpreted as a multivariate generalization of the Nadaraya-Watson regressor defined by Eqs. (1) and (3). Note that, unlike the MMSE estimator in [13], our proposal does not require an off-line GMM training and can be easily applied on-line from the set of available vectors $\mathbf{z}_j (j = 1, ..., M)$.

4. BANDWIDTH ESTIMATION

4.1. Classical KDE estimation

The most important issue in KDE problems is the bandwidth estimation (BE). There exist several approaches for it. A pop-



(a) (b)

Fig. 2. Example of $\epsilon_{\mathbf{y}} = \epsilon_{\mathbf{y}}(\beta^2; \mathbf{y}_0)$ for two different patches using scalar, diagonal and complete bandwidths.

ular and usually recommended approach is that of the so-called plug-in methods [14]. The goal of these methods is the minimization of the asymptotic mean integrated squared error (AMISE).

In this paper, we will consider the plug-in method for multivariate KDE described in [15, 16]. In this case, it is considered that $\mathcal{H} = \beta^2 F_{ZZ}$, where β^2 is a scale factor and F_{ZZ} is a structure matrix. If F_{ZZ} is known, the problem is reduced to the estimation of the scale factor β^2. This requires a quite complex procedure whose details can be found in [15]. It is interesting to note that if we decompose F_{ZZ} in the same way as in (5) the conditional mean of Eq.(8) does not depend on β^2, that is,

$$\boldsymbol{\mu}_{X|Y}^{(j)}(\mathbf{y}_0) = E[\mathbf{x}|\mathbf{y}_0, \mathbf{y}_j] = \mathbf{x}_j + F_{XY} F_{YY}^{-1} (\mathbf{y}_0 - \mathbf{y}_j), \quad (9)$$

In [15], F_{ZZ} is approximated by the covariance matrix C_{ZZ} of the observed samples $\{\mathbf{z}_j; j = 1, \ldots, M\}$.

4.2. A minimum square error (MSE) approach

The classical methods for BE in KDE (or regression) problems try to estimate a pdf suitable for the whole space of observations [14]. However, the goal of reconstruction techniques is to obtain an estimate of a specific patch $\mathbf{x}_0$ given its known context $\mathbf{y}_0$. Thus, a BE procedure to be employed in signal recovery problems should be oriented to be as accurate as possible at the point of interest.

In this paper we propose that the criterion for BE should be the same as the one employed for the sparse linear prediction method in [9], that is, the minimization of the square error of Eq. (2). This minimization is now constrained to weights of the form given by Eq. (7). Since these weights only depend on the bandwidth $\mathcal{H}$, we can consider that the function to be minimized is $\epsilon_{\mathbf{y}} = \epsilon_{\mathbf{y}}(\mathcal{H}; \mathbf{y}_0)$.

In order to carry the minimization of $\epsilon_{\mathbf{y}}$ versus the bandwidth $\mathcal{H}$ we could apply some sort of optimization algorithm. Some preliminary experiments (with a steepest descent procedure) have revealed that this type of solution yields an unstable convergence and poor results due to the large number of parameters in matrix $\mathcal{H}$. Only in the case of considering a scalar bandwidth (that is, $\mathcal{H} = h^2 I$, I identity matrix), we could obtain acceptable results. However, even in this case, the steepest descent solution was not worthwhile either since the minimization of $\epsilon_{\mathbf{y}} = \epsilon_{\mathbf{y}}(h^2; \mathbf{y}_0)$ was even much more time-consuming than an exhaustive search within the typical range of variation of h^2.

In order to overcome these problems, in this paper we propose a BE procedure as follows:

1. We will adopt the same assumption as in the plug-in BE method described above based on the use of a scale factor β^2 and a known structure F_{ZZ}, that is, $\mathcal{H} = \beta^2 F_{ZZ}$.
2. Then, since F_{ZZ} is fixed, the weights are only functions of β^2 (that is, $w_j = w_j(\beta^2; \mathbf{y}_0)$), so that the square error to be minimized $\epsilon_{\mathbf{y}} = \epsilon_{\mathbf{y}}(\beta^2; \mathbf{y}_0)$ also depends only on the scale factor β^2. Therefore, the corresponding minimization is feasible by exhaustive search within the typical range of variation of β^2.

In order to carry out an efficient exhaustive search, we can define a set of auxiliary weights as follows,

$$\tilde{w}_j(\mathbf{y}_0) = \exp\left((\mathbf{y}_0 - \mathbf{y}_j)^t F_{YY}^{-1} (\mathbf{y}_0 - \mathbf{y}_j)\right). \quad (10)$$

These auxiliary weights do not depend on β^2 and can be precomputed. Then, during the exhaustive search, the weights (Eq.(7)) for every value of β^2 can be efficiently obtained as,

$$w_j(\beta^2; \mathbf{y}_0) = \frac{(\tilde{w}_j(\mathbf{y}_0))^{1/\beta^2}}{\sum_{i=1}^{M} (\tilde{w}_i(\mathbf{y}_0))^{1/\beta^2}}. \quad (11)$$

Finally, once the optimal value of β^2 and its corresponding weights have been obtained, the unknown patch $\mathbf{x}_0$ can be estimated through Eqs. (6) and (9).

Several approaches for BE are adopted depending on the selection of the structure matrix F_{ZZ} according to its level of complexity [17]:

1. A scalar bandwidth $F_{ZZ} = \sigma_Z^2 I$, where σ_Z^2 is the variance of the available pixels (in set $\mathcal{S}$). This approach can be reduced to the algorithm described in [9] by forcing $\beta^2 \sigma_Z^2 = 10m$.
2. A diagonal bandwidth $F_{ZZ} = diag(C_{ZZ})I$.
3. A complete bandwidth $F_{ZZ} = C_{ZZ}$, as in [15].

Figure 2(a) shows examples of the error curve $\epsilon_{\mathbf{y}}(\beta^2; \mathbf{y}_0)$, obtained during the minimization procedure, for all three approaches. Scalar and diagonal bandwidths produce almost identical results since $\boldsymbol{\mu}_{X|Y}^{(j)}(\mathbf{y}_0) = \mathbf{x}_j$ for both cases and the diagonal of the correlation matrix C_{ZZ} tends to be uniform. Simulations reveal that both configurations tend to smooth high frequency textures (see Fig. 3(b)). On the other hand, complete bandwidth matrices can recover fine textures with high accuracy (see Fig.3(c)), although sometimes they show an unexpected behaviour (see Fig. 2(b)). A possible explanation is that the minimization of $\epsilon_{\mathbf{y}}(\beta^2; \mathbf{y}_0)$ is equivalent to the maximization of the corresponding PSNR only if $F_{YY} = \sigma_Z^2 I$. Moreover, the scalar (and diagonal) approach is more robust against non-stationarity. In this case, an inaccurate selection of the structure matrix $F_{ZZ} = \sigma_Z^2 I$ can be corrected by modifying β^2, since $\mathcal{H} = \beta^2 \sigma_Z^2 I$. This, however, is not possible for complete structure matrices. Thus, in order to achieve a compromise between texture reconstruction and PSNR, we will also test a combination of scalar and

SEC	*Lena*		*Goldhill*		*Barbara*		Average	
	PSNR	MS-SSIM	PSNR	MS-SSIM	PSNR	MS-SSIM	PSNR	MS-SSIM
[1]	30.42	96.56	31.27	95.65	26.85	94.87	28.57	95.04
[5]	31.96	97.25	30.24	94.53	27.39	96.20	29.46	95.71
[2]	32.17	97.64	31.12	95.71	27.99	96.00	29.57	95.99
[4]	32.15	97.44	30.91	95.52	29.91	97.04	30.22	96.39
[6]	30.85	97.08	30.40	95.21	28.03	95.72	29.23	95.69
[7]	32.70	97.96	31.66	96.35	28.41	97.37	30.28	96.74
[3]	32.82	97.65	31.54	95.62	29.66	97.07	30.35	96.08
[8]	32.72	97.80	31.78	96.14	30.84	97.64	30.50	96.56
[9]	32.55	97.97	31.72	96.43	30.80	98.01	30.55	96.98
KD_S	32.22	98.02	31.43	96.40	30.84	98.11	30.51	97.00
MS_S	32.84	98.11	32.03	96.67	31.33	98.25	31.20	97.27
MS_D	32.87	98.11	32.02	96.66	31.35	98.26	31.21	97.28
MS_C	32.69	98.00	32.14	96.77	31.77	98.35	31.24	97.28
MS_X	**33.00**	**98.18**	**32.17**	**96.84**	**32.22**	**98.55**	**31.43**	**97.40**

Table 1. PSNR values (in dB) and MS-SSIM indices (scaled by 100) for test images reconstructed by several algorithms for block dimensions 16 × 16. The best performances for each image are in bold face.

complete bandwidths where the complete bandwidth matrix is employed to compute the conditional means $\boldsymbol{\mu}_{X|Y}^{(j)}(\mathbf{y}_0)$ (Eq. (8)) and the scalar bandwidth for the weights w_j (Eq. (7)).

5. EXPERIMENTAL RESULTS

In order to reflect the perceptual quality of the reconstructions, the multi-scale structural similarity (MS-SSIM) index [18] is used for comparison along with the objective PSNR measure. MS-SSIM is a weighted combination of SSIM indices computed over different image resolutions. Thus, coarse structures as well as fine textures are taken into account. SSIM index aims at approximating the human visual system response looking for similarities in structure, contrast and intensity [19].

The performance of our different proposals is tested on the images of *Lena* (512×512), *Goldhill* (720×576), *Foreman* (352×288), *Barbara* (512×512), *Baboon* (512×512), *Clown* (512×512), *Tire* (205×232), *Pirate* (1024×1024), *Boat* (512×512) and *Peppers* (384×512). We will use the framework described in Section 2. For our MSE approaches, β^2 is searched exhaustively within the range [0, 2] with steps of 0.01. First, we test the performance of the scalar bandwidth using the classical KDE (KD_S) of Section 4.1 and our MSE approach (MS_S) and compare them with our previous exponential sparse linear prediction (SLP) algorithm [9], which we will use as a reference (marked in Table 1). Table 1 shows that KD_S performs considerably worse than MS_S and it provides virtually no improvement over [9]. Thus, in the following, we focus on the MSE approach for scalar bandwidth as well as diagonal (MS_D) and complete (MS_C) bandwidth matrices. Moreover, we also use the combined scenario with scalar and complete bandwidths (MS_X) as described at the end of Section 4.2. We compare our proposals with other state-of-the-art SEC techniques [1]-[9].

Table 1 shows the results in terms of PSNR and MS-SSIM for the images of *Lena*, *Goldhill* and *Barbara* as well as the average performance over all ten tested images. The results confirm our hypothesis that the KD_S approach is not EC oriented. On the other hand, all of our MSE proposals outperform the other techniques, including our previous exponential SLP. In addition, scalar and diagonal bandwidths produce almost identical results. The complete bandwidth performs better on average although is inferior in some particular cases (e.g. *Lena*). Finally, the combination of the high quality reconstructions produced by complete bandwidth with the good behaviour of the scalar one produces the best result both on subjective and objective levels. A subjective comparison is shown in Fig.3.

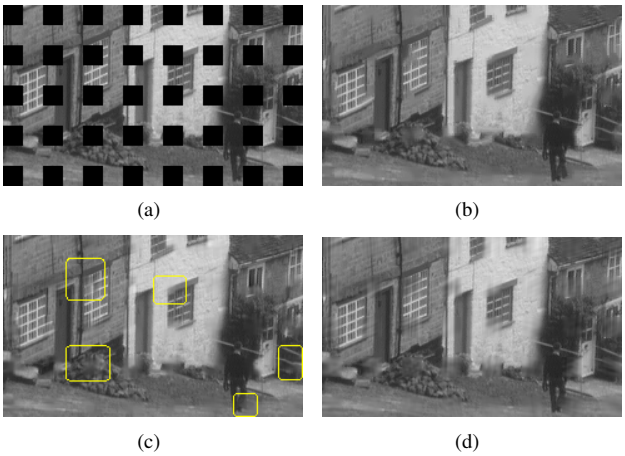


Fig. 3. Subjective comparison for a fraction of *Goldhill*. (a) Received data. (b) Reconstruction using scalar bandwidth (MS_S). (c) Reconstruction using complete bandwidth matrix (MS_C). (d) Reconstructed by [8].

Due to the efficient implementation of the exhaustive search, carried out utilizing precomputed weights (Eqs. (10) and (11)), the computational complexity is only moderately increased with respect to SLP for all the MSE proposals. This increment of complexity is reflected in the reconstruction quality which is improved in 0.9dB on average (for MS_X). The KD_S approach, on the other hand, requires up to half an hour per macroblock and therefore is computationally prohibitive for on-line applications.

6. CONCLUSIONS

We have proposed a framework for image EC based on a generalization of the Nadaraya-Watson estimator with an MSE-based bandwidth estimation. We have shown that this MSE criterion achieves a performance significantly better than that of classical KDE with pdf matching. Using a simple scalar bandwidth we achieve an average improvement over [9] of 0.7dB. This improvement is later incremented up to almost 1dB by combining the robustness of the scalar bandwidth with the accurate reconstructions of fine textures produced by complete bandwidth matrices. Ongoing work is focused on a more accurate selection of the bandwidth matrix structure.

7. REFERENCES

[1] T. Wiegand, G.J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, pp. 560–576, July 2003.

[2] S. Shirani, F. Kossentini, and R. Ward, "An adaptive Markov random field based error concealment method for video communication in error prone environment," in *Proceedings of ICIP*, 1999, vol. 6, pp. 3117–3120.

[3] X. Li and M.T. Orchard, "Novel sequential error-concealment techniques using orientation adaptive interpolation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, pp. 857–864, October 2002.

[4] G. Zhai, J. Cai, W. Lin, X. Yang, and W. Zhang, "Image error-concealment via block-based bilateral filtering," *IEEE International Conference on Multimedia and Expo*, pp. 621–624, June 2008.

[5] Z. Rongfu, Z. Yuanhua, and H. Xiaodong, "Content-adaptive spatial error concealment for video communication," *IEEE Transactions on Consumer Electronics*, vol. 50, pp. 335–341, February 2004.

[6] P.F. Harrison, *Texture Synthesis, Texture Transfer and Plausible Restoration*, Ph.D. thesis, Monash University, 2005.

[7] J. Koloda, V. Sánchez, and A. M. Peinado, "Spatial error concealment based on edge visual clearness for image/video communication," *Circuits, Systems and Signal Processing*, October 2012.

[8] A. Kaup, K. Meisinger, and T. Aach, "Frequency selective signal extrapolation with applications to error concealment in image communication," *AEUE - International Journal of Electronics and Communications*, vol. 59, pp. 147–156, 2005.

[9] J. Koloda, J. Østergaard, S. H. Jensen, V. Sánchez, and A. M. Peinado, "Sequential error concealment for video/images by sparse linear prediction," *IEEE Transactions on Multimedia*, In Press.

[10] E.A. Nadaraya, "On estimating regression," *Theory of Probability and its Applications*, vol. 9, pp. 141–142, September 1964.

[11] D.W. Scott, "Multivariate density estimation: Theory, practice, and visualization," Wiley, 1992.

[12] ITU-T, "ITU-T Recommendation H.264," International Telecommunication Union, 2010.

[13] D. Persson, T. Eriksson, and P. Hedelin, "Packet video error concealment with gaussian mixture models," *IEEE Transactions on Image Processing*, vol. 17, pp. 145–154, 2008.

[14] S.J. Sheather, "Density estimation," *Statistical Science*, vol. 19, no. 4, pp. 588–597, 2004.

[15] M. Kristan, A. Leonardis, and D. Skočaj, "Multivariate online kernel density estimation with gaussian kernels," *Pattern Recognition*, vol. 44, pp. 2630–2642, 2011.

[16] M.P. Wand and M.C. Jones, "Multivariate plug-in bandwidth selection," *Computational Statistics*, , no. 9, pp. 97–117, 1994.

[17] M.P. Wand and M.C. Jones, "Comparison of smoothing parameterizations in bivariate kernel density estimation," *Journal of the American Statistical Association*, vol. 88, no. 422, pp. 520–528, 1993.

[18] Z. Wang, E.P. Simoncelli, and A.C. Bovik, "Multiscale structural similarity for image quality assessment," *IEEE Signals, Systems and Computers*, vol. 2, pp. 1398–1402, November 2003.

[19] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assesment: From error visibility to structural visibility," *IEEE Transactions on Image Processing*, vol. 13, pp. 600–612, April 2004.

2.2.2.2 Kernel-based MMSE Multimedia Signal Reconstruction and its Application to Spatial Error Concealment

- J. Koloda, A.M. Peinado and V. Sánchez, "Kernel-based MMSE Multimedia Signal Reconstruction and its Application to Spatial Error Concealment", *IEEE Transactions on Multimedia*, accepted.
    - Status: Accepted.
    - Impact Factor (JCR 2012): 1.754
    - Subject Category: Computer Science, Information Systems. Ranking 24/132 (Q1).
    - Subject Category: Computer Science, Software Engineering. Ranking 15/105 (Q1).
    - Subject Category: Telecommunications. Ranking 14/78 (Q1).

Kernel-based MMSE multimedia signal reconstruction and its application to spatial error concealment

Ján Koloda, *Student Member, IEEE*, Antonio M. Peinado, *Senior Member, IEEE*, and Victoria Sánchez, *Member, IEEE*

***Abstract*—This paper proposes a novel approach for multimedia signal reconstruction based on kernel density estimation (KDE). We make use of a vector formalism in which vectors consist of a first subvector containing a set of missing samples and a second one containing a set of available context samples. The missing subvector is reconstructed by a minimum mean square error estimator which employs a probability density function (pdf) obtained by KDE. As in any kernel-based method, the main issue to deal with is the estimation of an appropriate kernel bandwidth. We propose an adaptive procedure for bandwidth estimation (BE) especially conceived for signal reconstruction. Thus, unlike general KDE or kernel-based regression, which try to obtain a general fit, the focus of this BE procedure is on the specific missing subvector. Also, in order to exploit local signal correlations, our BE proposal adopts a scaling approach in which the bandwidth is computed as the local covariance matrix scaled by two factors. These two scale factors are obtained by minimization of two different approximations to the reconstruction error. The resulting reconstruction methodology is tested on a spatial error concealment (EC) application in which intracoded images have been transmitted through an error prone channel. The experimental results show the superiority of the proposed approach over a wide range of existing EC techniques.**

***Index Terms*—kernel density estimation, bandwidth estimation, multimedia signal reconstruction, spatial error concealment**

I. Introduction

THE estimation of a probability density function (pdf) from a given data set is a necessary step in many multimedia signal processing applications [1]. A classical widespread approach is the one based on the assumption that the signal statistics can be appropriately modelled by a certain parametric pdf, so that the pdf estimation is reduced to the computation of the model parameters from training data. This approach has two critical issues which may make it inaccurate: the suitability of the selected model and the availability of sufficient data for model training. On the contrary, in non-parametric methods such as kernel density estimation (KDE), the available data set itself determines the structure of the pdf, avoiding the need for a model selection. Examples of multimedia applications, such as regression, classification, tracking, segmentation, super-resolution, reconstruction or, in particular, error concealment, where KDE has been successfully employed can be found in [1]–[6].

In KDE, a kernel function is repeated at every data point and the combination of the resulting replicas generates the desired pdf estimate. There are two issues that must be addressed in order to solve a KDE problem: the selection of the kernel function and the estimation of a suitable kernel width (usually known as bandwidth), being this last issue the most critical one [7]. While the typical approach involves the estimation of a single global bandwidth for the whole pdf [8], in many cases, e.g. multimedia signals, one global bandwidth is not sufficient since data statistics are not homogeneous and a global fit to the entire data set could be considerably inaccurate. Variable-bandwidth kernel estimation improves the performance of kernel estimators by focusing the kernel bandwidth to the local data statistics [3].

In this paper we will deal with the problem of signal reconstruction, that is, the estimation of a group of missing signal samples given its known context under a kernel-based point of view. In principle, signal reconstruction can be viewed as a regression problem, where the regressor can be expressed as an expectation over a KDE pdf estimate [9]. However, the objectives of signal reconstruction are significantly different from those of KDE or regression. First, the goal of reconstruction is the estimation of a specific group of samples rather than a global or even a local fitting like in KDE or regression. Second, an accurate reconstruction, able to reproduce fine signal details, must preserve local signal correlations. This issue is especially relevant in the case of multimedia signals. In order to achieve both goals, we will adopt a multidimensional framework where a group of adjacent signal samples are arranged in a multidimensional vector [10] [11]. This multidimensional formalism will allow a suitable signal correlation modelling as well as the interpretation of the reconstruction problem as the estimation of a single vector in the missing data subspace.

As for any KDE problem, the main issue of kernel-based reconstruction will also be the estimation of a suitable bandwidth, which, under the multidimensional formalism, becomes a matrix. Bandwidth estimation methods can be classified into two main categories [12]: quality-of-fit methods and plug-in methods. The first category employs cross validation. The bandwidth is estimated on a subset of samples and then validated on the remaining ones. Usually, a least squares criterion is followed. The plug-in techniques optimize the

The authors are with the Department of Signal Theory, Networking and Communications, University of Granada, Granada, Spain (e-mail: janko@ugr.es; amp@ugr.es; victoria@ugr.es).

This work has been supported by the Spanish MEC/FEDER project TEC 2010-18009.

fit between the real density function and its kernel-based approximation by minimizing the mean integrated square error (MISE). Other pdf error criteria can also be utilized [13].

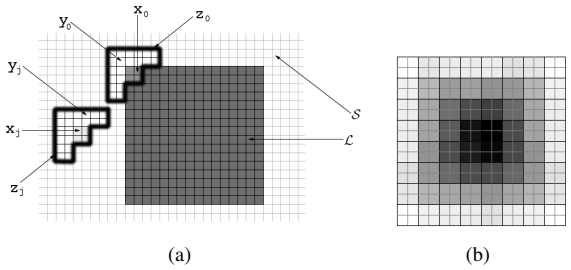
In spite of the extensive bibliography on bandwidth estimation, none of the above mentioned categories is oriented to multimedia signal reconstruction. In the following sections we will propose a methodology for signal reconstruction based on KDE with variable bandwidth which has been especially conceived for this task. This methodology can be summarized in three points. First, the signal reconstructor derives from a minimum mean square error (MMSE) estimator based on KDE. Second, the criteria for the estimation of the corresponding bandwidth matrix will be derived from a comparison with the linear MMSE (LMMSE) estimator. Finally, in order to mitigate the problem of the large number of parameters to be estimated in a full bandwidth matrix, we will propose a submatrix scaling approach where the bandwidth is obtained from a structure matrix (which accounts for the signal correlations) which is adapted by applying different scale factors at submatrix level.

Although the proposed methodology is general and applicable to any mono- or multi-dimensional signal, we will test its utility over an error concealment (EC) application and, in particular, we will consider the concealment of intra-coded images transmitted over a lossy channel with application to spatial EC in video signals. It is worth noticing that non-parametric methods and kernel regression have been scarcely researched within the image/video processing field [6]. Moreover, EC applications based on KDE are even scarcer. In this line, we can mention the bilateral kernel regression proposed in [14] which employs a pair of spatial and radiometric kernels that are used to filter the degraded image. This approach, however, neglects the correlations between the positions of the pixels and their values, and the reconstruction of periodic textures may be penalized as pointed out in [6] and [15], respectively. Also a sequential Bayesian approach using a DCT pyramid is treated in [11]. Here, fixed bandwidth values are applied regardless of the input data. Unlike our proposal, these EC techniques utilize scalar bandwidths that do not take into account the correlations between adjacent pixels so fine textures may be recovered inaccurately.

The paper is organized as follows. Section II describes the multidimensional framework that will be used throughout the paper and the experimental setup for the image EC application. Kernel-based MMSE estimation is introduced in Section III. In Section IV, a scaling approach for bandwidth estimation oriented to signal recovery is proposed. The simulation results are presented and discussed in Section V. The last Section is devoted to conclusions.

II. Multidimensional formalism and image error concealment

The utility of a multidimensional formalism for multimedia signal reconstruction was already justified in the introduction section and has been previously employed in other related work [10] [11] [15] [16]. In this section, this formalism is summarized along with its particularization to an EC application where groups of pixels in an intra-coded image, which have been lost due to an error-prone transmission channel, must be reconstructed.



(a) (b)

Fig. 1. (a) Example of configuration for the vectors $\boldsymbol{x}$, $\boldsymbol{y}$ and $\boldsymbol{z}$. (b) Filling order for sequential reconstruction with 2×2 patches. The regions illustrated by brighter level are recovered first.

We consider a set $\mathcal{R}$ of adjacent samples (e.g., a segment for 1D signals or a region for 2D signals). We also consider that this segment is degraded in some way, so it finally contains both missing and available samples. Let $\mathcal{L}$ and $\mathcal{S}$ be the sets of adjacent missing samples and available samples, respectively, so that $\mathcal{R} = \mathcal{L} \bigcup \mathcal{S}$.

Our goal is the estimation of a vector $\boldsymbol{x}_0 \subset \mathcal{L}$ of lost samples. In our scheme, we consider that $\boldsymbol{x}_0$ is part of a larger vector $\boldsymbol{z}_0 = (\boldsymbol{x}_0^t, \boldsymbol{y}_0^t)^t$, where $\boldsymbol{y}_0 \subset \mathcal{S}$ is a context vector which contains a set of available samples adjacent to those of $\boldsymbol{x}_0$. A 2D-signal example of configuration for $\boldsymbol{z}_0$ is illustrated in Fig. 1(a). Additionally, we will consider all the possible available vectors $\boldsymbol{z}_j$ $(j = 1, \ldots, M)$ that can be built in $\mathcal{S}$ with the same configuration, that is, the same shape and dimensionality as $\boldsymbol{z}_0$. Therefore, every $\boldsymbol{z}_j$ can be decomposed in the same way as $\boldsymbol{z}_0$, that is, $\boldsymbol{z}_j = (\boldsymbol{x}_j^t, \boldsymbol{y}_j^t)^t$. This is also illustrated in Fig. 1(a). In the following, $\boldsymbol{x}_j$ and $\boldsymbol{y}_j$ $(j = 1, \ldots, M)$ will be referred to as prototype vectors and context vectors, respectively.

As mentioned, the proposed kernel-based methodology for signal reconstruction will be tested over an EC application for intra-coded images transmitted over a lossy channel. The corresponding implementation details are the same as those described in [15] and are briefly summarized here. We consider a block-based codec where the missing region $\mathcal{L}$ is a 16×16 macroblock and the support area $\mathcal{S}$ comprises all the available pixels within the neighbouring 16×16 macroblocks around $\mathcal{L}$. We will employ an error pattern as shown in Fig. 8(a) which corresponds to a single packet loss of a frame with dispersed slicing structure, leading to a block loss rate of approximately 25% [15]. Moreover, random losses with the same rate will also be employed. Figure 5(a) shows an example of random loss. It must be pointed out that the techniques proposed along this paper can be straightforwardly extended to other error patterns or other applications, such as inpainting and, in general, signal recovery applications.

We will consider that the missing subvector $\boldsymbol{x}_0$ is a 2×2 patch of pixels, and its corresponding context vector $\boldsymbol{y}_0$ will comprise all the available pixels within the 6×6 pixel neighbourhood centred at $\boldsymbol{x}_0$, as described in [15]. This yields 4 dimensions for $\boldsymbol{x}_0$, and from 16 up to 32 dimensions for its context $\boldsymbol{y}_0$. Vectors $\boldsymbol{z}_j$ replicate the shape and dimensionality of $\boldsymbol{z}_0$. These configurations are also shown in Figure 1(a).

This figure also reveals that the inner samples in $\mathcal{L}$ cannot be directly estimated since there are no context samples around them. In order to overcome this problem, missing vectors in set $\mathcal{L}$ are reconstructed by applying a filling procedure (see [15] for details) where estimates are sequentially obtained from the outer layer of $\mathcal{L}$ towards its centre. Thus, samples already reconstructed can be considered available (that is, they are moved to set $\mathcal{S}$), allowing so the reconstruction of inner samples. This sequential filling is illustrated in Fig. 1(b).

III. Kernel-based MMSE estimation

As mentioned in the introduction section, our goal is to obtain an MMSE estimate $E[\boldsymbol{x}|\boldsymbol{y}_0]$ of the missing vector $\boldsymbol{x}_0$ given its context $\boldsymbol{y}_0$. In order to do so, a characterization of the signal statistical behaviour is required. In particular, we will consider the random vector variable $\boldsymbol{z} = (\boldsymbol{x}^t, \boldsymbol{y}^t)^t$ corresponding to the sample configuration defined by $\boldsymbol{x}_0$ and $\boldsymbol{y}_0$ as explained in Section II. If the probability density function (pdf) associated to $\boldsymbol{z}$ is available, then the desired MMSE estimate can be obtained from it. In this section, we explore the application of KDE for this task.

KDE provides an estimate of the pdf associated to the random variable $\boldsymbol{z}$ given a set of observed vectors $\{\boldsymbol{z}_j; j = 1, \ldots, M\}$ in a non parametric way, that is, avoiding any assumption about the original pdf. This is carried out by replicating a basic kernel function $K(\mathbf{u})$ at the observed vectors and summing as follows,

$$p(\boldsymbol{z}) = \frac{1}{M}\sum_{j=1}^{M}\frac{1}{|H|}K\left(H^{-1}(\boldsymbol{z}-\boldsymbol{z}_j)\right) = \frac{1}{M}\sum_{j=1}^{M}K_Z^{(j)}(\boldsymbol{z}). \tag{1}$$

Matrix H, commonly known as bandwidth, plays a key role in KDE methods since it controls the smoothness of the resulting pdf. We will assume a Gaussian kernel $K(\mathbf{u}) = \exp(-\mathbf{u}^t\mathbf{u}/2)/\sqrt{2\pi}$ in the rest of the paper. In Eq. (1) we have also introduced a simplified notation where $K_Z^{(j)}(\boldsymbol{z})$ is a multivariate Gaussian with mean $\boldsymbol{z}_j$ and covariance $\mathcal{H} = HH^t$ which will be also referred to as bandwidth for simplicity. This bandwidth can be decomposed as,

$$\mathcal{H} = \begin{pmatrix} \mathcal{H}_{XX} & \mathcal{H}_{XY} \\ \mathcal{H}_{YX} & \mathcal{H}_{YY} \end{pmatrix}. \tag{2}$$

The knowledge of $p(\boldsymbol{z})$ allows the application of Bayesian techniques and, in particular, MMSE estimation. In order to do that, it is convenient to observe that the pdf of Eq. (1) has the form of Gaussian mixture model (GMM) with equal *a priori* probabilities and covariance matrices ($1/M$ and $\mathcal{H}$, respectively). Then, we can straightforwardly adapt the well-known expressions for the MMSE estimator of $\boldsymbol{x}_0$ given $\boldsymbol{y}_0$ under GMM modelling [10], which for our case are [16],

$$\hat{\boldsymbol{x}}_0 = E[\boldsymbol{x}|\boldsymbol{y}_0] = \textstyle\sum_{j=1}^{M} w_j(\boldsymbol{y}_0)\boldsymbol{\mu}_{X|Y}^{(j)}(\boldsymbol{y}_0) \tag{3}$$

$$w_j(\boldsymbol{y}_0) = \frac{K_Y^{(j)}(\boldsymbol{y}_0)}{\sum_{i=1}^{M} K_Y^{(i)}(\boldsymbol{y}_0)} \tag{4}$$

$$\boldsymbol{\mu}_{X|Y}^{(j)}(\boldsymbol{y}_0) = E[\boldsymbol{x}|\boldsymbol{y}_0, \boldsymbol{y}_j] = \boldsymbol{x}_j + \mathcal{H}_{XY}\mathcal{H}_{YY}^{-1}\left(\boldsymbol{y}_0 - \boldsymbol{y}_j\right) \tag{5}$$

where $K_Y^{(j)}(\boldsymbol{y})$ is the marginal kernel for $\boldsymbol{y}$, with mean $\boldsymbol{y}_j$ and covariance $\mathcal{H}_{YY}$. In the following, we will refer to Eqs. (3)-(5) as K-MMSE estimator (kernel-based MMSE estimator). It is worth noticing that, unlike the MMSE estimator in [10], the K-MMSE estimator can be applied on-line from the set of observations (without the need of any *a priori* pdf model), although it will require an estimate of the bandwidth matrix $\mathcal{H}$. This issue is dealt with in the next section.

The K-MMSE estimator can be expressed more compactly as,

$$\hat{\boldsymbol{x}}_0 = \tilde{\boldsymbol{x}}_0 + \mathcal{H}_{XY}\mathcal{H}_{YY}^{-1}\left(\boldsymbol{y}_0 - \tilde{\boldsymbol{y}}_0\right) \tag{6}$$

where $\tilde{\boldsymbol{x}}_0$ and $\tilde{\boldsymbol{y}}_0$ are vectors linearly predicted from the sets of prototype and context vectors, respectively, that is,

$$\tilde{\boldsymbol{x}}_0 = \sum_{j=1}^{M} w_j(\boldsymbol{y}_0)\boldsymbol{x}_j \tag{7}$$

$$\tilde{\boldsymbol{y}}_0 = \sum_{j=1}^{M} w_j(\boldsymbol{y}_0)\boldsymbol{y}_j. \tag{8}$$

Note that when $\boldsymbol{x}$ and $\boldsymbol{y}$ are independent variables ($\mathcal{H}_{XY} = 0$), then the K-MMSE estimator is reduced to a multivariate Nadaraya-Watson (NW) regressor ($\hat{\boldsymbol{x}}_0 = \tilde{\boldsymbol{x}}_0$) with bandwidth matrix $\mathcal{H}_{YY}$. Therefore, we can say that the K-MMSE estimator consists of two terms. The first one, the predicted vector $\tilde{\boldsymbol{x}}_0$ of Eq. (7), is a NW estimate of $\boldsymbol{x}_0$. For a Gaussian kernel, the prediction weights w_j are computed as,

$$w_j(\boldsymbol{y}_0) = \frac{\exp\left(-\frac{1}{2}(\boldsymbol{y}_0 - \boldsymbol{y}_j)^t\mathcal{H}_{YY}^{-1}(\boldsymbol{y}_0 - \boldsymbol{y}_j)\right)}{\sum_{i=1}^{M}\exp\left(-\frac{1}{2}(\boldsymbol{y}_0 - \boldsymbol{y}_i)^t\mathcal{H}_{YY}^{-1}(\boldsymbol{y}_0 - \boldsymbol{y}_i)\right)}. \tag{9}$$

The second term of Eq. (6) is a correction vector where the unpredictable part of $\boldsymbol{y}_0$ is transformed into subspace $\boldsymbol{x}$.

Like in all kernel-based problems, the key point will be the estimation of the kernel bandwidth. In the following section we will focus on this issue and propose a bandwidth matrix estimation methodology especially oriented to multimedia signal reconstruction and able to dynamically adapt itself to the local characteristics of the data.

IV. Bandwidth estimation for signal reconstruction

The estimation of the bandwidth is the key issue in KDE (or kernel-based regression) problems, where a pdf suitable for the whole space of observations is usually desired [8]. However, the goal of signal reconstruction is to obtain an estimate of a specific vector $\boldsymbol{x}_0$ given its known context $\boldsymbol{y}_0$, where, in general, the signal will be non-stationary. Thus, a bandwidth estimation (BE) procedure for signal recovery should focus on the point of interest. This can be easily understood if we take into account that the weights w_j of Eq. (4) that control the contribution of every observation $\boldsymbol{z}_j$ can be alternatively written as,

$$w_j(\boldsymbol{y}_0) = \frac{K_Y^{(0)}(\boldsymbol{y}_j)}{\sum_{i=1}^{M} K_Y^{(0)}(\boldsymbol{y}_i)} \qquad (j = 1, ..., M) \tag{10}$$

where $K_Y^{(0)}(\boldsymbol{y})$ represents a multivariate Gaussian with mean $\boldsymbol{y}_0$ and covariance $\mathcal{H}_{YY}$. Then, instead of having a GMM

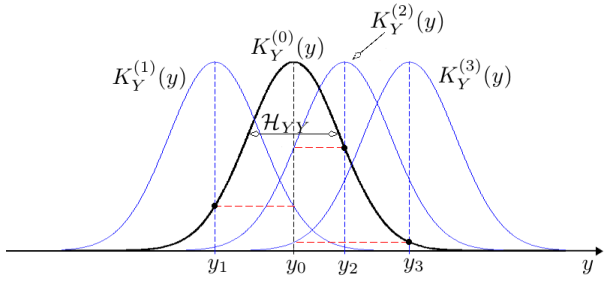


Fig. 2. Monodimensional example of prediction weight computation. The weights (marked with black points) can be obtained by evaluating kernel replicas $K_Y^{(j)}(y)$ at y_0 or, alternatively, a single kernel $K_Y^{(0)}(y)$ at y_j $(j = 1, ..., M)$.

model for the whole space of observations which is evaluated at point $\boldsymbol{y}_0$, we can consider that we have a single Gaussian centred at $\boldsymbol{y}_0$ and evaluated at every observed context $\boldsymbol{y}_j$. This is illustrated in Fig. 2 for the case of a monodimensional subvector $\boldsymbol{y}$. The figure makes it clear that the important issue here is the bandwidth of the kernel centred at $\boldsymbol{y}_0$, which controls the similarity of the observed contexts $\boldsymbol{y}_j$ $(j = 1, ..., M)$ with $\boldsymbol{y}_0$.

Also, we must point out the importance of using a multivariate kernel with a complete bandwidth matrix in signal reconstruction problems. Most of the literature on BE for KDE deals with the estimation of a scalar bandwidth. This type of bandwidth is enough for controlling the smoothness of the KDE-estimated pdf. Thus, too large bandwidths may produce significant bias whereas too small ones may cause large estimation variance [17]. However, in the case of multivariate kernels, we must also consider that a full bandwidth matrix, capable of capturing the correlations between the samples of pattern $\boldsymbol{z}$ in the neighbourhood of $\boldsymbol{z}_0$, may be useful to reconstruct fine signal textures.

In order to obtain the required full bandwidth matrix, we will see later in this section that reconstruction error criteria must be established. The corresponding error functions could be directly minimized versus $\mathcal{H}$ by means of an optimization algorithm. Some preliminary experiments with a steepest descent procedure have revealed that this type of solution yields an unstable convergence and poor results due to the large number of parameters in matrix $\mathcal{H}$, which, in our case, can have hundreds of different elements.

This problem of a large number of parameters to be estimated is commonly solved by *scaling* [8] [18] [19]. The underlying assumption of this approach is that the bandwidth can be expressed as $\mathcal{H} = \beta F$, where β is a scale factor and F is a structure matrix which represents the correlations between samples in $\boldsymbol{z}$. A suitable candidate for F [18] [19] is the covariance matrix C_{ZZ} of $\boldsymbol{z}$, which can be easily estimated from the observed data $\boldsymbol{z}_j$ $(j = 1, \ldots, M)$. Thus, the clear advantage of the scaling approach is that the only free parameter to be optimized is the scale factor. In this section we will propose a multivariate BE method which also adopts a scaling approach although introducing two different scale factors. This new scaling approach is based on the performance analysis developed in the following subsection.

A. K-MMSE estimator performance analysis

The performance of a given estimator is commonly evaluated by means of its mean square error (MSE). At the point of interest $\boldsymbol{y}_0$, the MSE is [20],

$$MSE(\boldsymbol{y}_0) = E[\|\boldsymbol{x} - \hat{\boldsymbol{x}}_0\|^2 | \boldsymbol{y}_0]. \tag{11}$$

The computation of this error would require the knowledge of the true local pdf $p(\boldsymbol{x}|\boldsymbol{y}_0)$. Obviously, this is not the case of our K-MMSE estimator, where, in fact, we are trying to estimate the local signal statistics. As an alternative, we will consider here the similarity of our K-MMSE estimator, as expressed in Eq. (6), with a linear MMSE (LMMSE) estimator [21] with mean vector $\tilde{\boldsymbol{z}}_0 = (\tilde{\boldsymbol{x}}_0^t, \tilde{\boldsymbol{y}}_0^t)^t$. If the true local second-order statistics around this mean vector are described by a covariance matrix Σ_{ZZ} and we decompose this matrix in the same way as in Eq. (2), then the LMMSE estimate of vector $\boldsymbol{x}$ given its context $\boldsymbol{y}$ is,

$$\boldsymbol{\mu}_L(\boldsymbol{y}) = \tilde{\boldsymbol{x}}_0 + \Sigma_{XY}\Sigma_{YY}^{-1}\left(\boldsymbol{y} - \tilde{\boldsymbol{y}}_0\right). \tag{12}$$

On the other hand, if we consider that the estimator of Eq. (6) can be extended to vectors $\boldsymbol{y}$ in the vicinity of $\boldsymbol{y}_0$, then we can express,

$$\hat{\boldsymbol{x}}(\boldsymbol{y}) = \tilde{\boldsymbol{x}}_0 + \mathcal{H}_{XY}\mathcal{H}_{YY}^{-1}\left(\boldsymbol{y} - \tilde{\boldsymbol{y}}_0\right). \tag{13}$$

We observe that the only difference between both estimators resides in the transformations applied to the difference vector $\boldsymbol{\Delta} = \boldsymbol{y} - \tilde{\boldsymbol{y}}_0$. These transformations will be noted as $T_\Sigma = \Sigma_{XY}\Sigma_{YY}^{-1}$ and $T_\mathcal{H} = \mathcal{H}_{XY}\mathcal{H}_{YY}^{-1}$. Since the LMMSE estimator is the optimal one for a given mean and covariance [21], an alternative criterion for the estimation of the bandwidth parameters could be the minimization of the following mean square error,

$$\mathcal{E}(\mathcal{H}) = E[\epsilon(\mathcal{H}; \boldsymbol{y})] \tag{14}$$

where

$$\epsilon(\mathcal{H}; \boldsymbol{y}) = \|\boldsymbol{\mu}_L(\boldsymbol{y}) - \hat{\boldsymbol{x}}(\boldsymbol{y})\|^2 = \|(T_\Sigma - T_\mathcal{H})\,\boldsymbol{\Delta}\|^2. \tag{15}$$

The minimization of $\mathcal{E}$ with respect to $\mathcal{H}$ is obviously infeasible since our problem is precisely that T_Σ is unknown (and it cannot be reliably estimated due to the likely lack of local samples). However, the square error of Eq.(15) still provides two useful hints about how the bandwidth should be obtained:

1) The first hint is obvious: $T_\mathcal{H}$ should be as similar as possible to T_Σ.
2) Since, in general, $T_\Sigma \neq T_\mathcal{H}$, the K-MMSE estimator only coincides with the LMMSE estimator when $\|\boldsymbol{\Delta}\|^2 = 0$, that is, when $\boldsymbol{y} = \tilde{\boldsymbol{y}}_0$. This fact suggests us that $\tilde{\boldsymbol{y}}_0$ (which is also a function of $\mathcal{H}$) should be as close as possible to the point of interest $\boldsymbol{y}_0$.

In the next subsection we propose a double scaling scheme which will allow us to exploit these hints even with an unknown T_Σ.

B. Covariance submatrix scaling

As previously mentioned, the estimation of matrix $\mathcal{H}$ involves a large number of parameters which makes the multivariate BE problem particularly complicated. While scaling a structure matrix by a single scale factor may provide a simple solution, the partition of the bandwidth matrix shown in Eq. (2), which divides $\mathcal{H}$ into the four submatrices employed by the K-MMSE estimator, suggests a new scaling scheme where different scale factors $(\beta_{XX}, \beta_{YY}, \beta_{XY}, \beta_{YX})$ are applied to different submatrices. Adopting the sample covariance as structure matrix, the bandwidth matrix is,

$$\mathcal{H} = \begin{pmatrix} \beta_{XX} C_{XX} & \beta_{XY} C_{XY} \\ \beta_{YX} C_{YX} & \beta_{YY} C_{YY} \end{pmatrix}. \tag{16}$$

Considering the symmetry constraint for $\mathcal{H}$, we have that $\beta_{XY} = \beta_{YX}$. Substituting (16) in Eq. (6), we obtain the following estimator,

$$\hat{\boldsymbol{x}}_0 = \tilde{\boldsymbol{x}}_0 + \alpha C_{XY} C_{YY}^{-1} (\boldsymbol{y}_0 - \tilde{\boldsymbol{y}}_0) \tag{17}$$

where $\alpha = \beta_{XY}/\beta_{YY}$. We can observe that this estimator only requires two bandwidth parameters, α, which appears explicitly in the estimator (Eq. (17)), and β_{YY}, which is required to compute the weights of Eq. (9) which, in turn, are required to compute the prediction vectors $\tilde{\boldsymbol{x}}_0$ and $\tilde{\boldsymbol{y}}_0$ (Eq. (7) and (8)). For the sake of simplicity, it is noted $\beta_{YY} = \beta$ in the following, so that $\tilde{\boldsymbol{x}}_0 = \tilde{\boldsymbol{x}}_0(\beta)$ and $\tilde{\boldsymbol{y}}_0 = \tilde{\boldsymbol{y}}_0(\beta)$.

Different bandwidths can be obtained depending on the selected type of structure matrix [22]. Under the scaling approach adopted here, the structure and, therefore, the bandwidth, is determined by the type of covariance matrix. The following options will be considered here:

1) A scalar bandwidth (K-MMSE$_S$): $C_{ZZ} = \sigma_Z^2 I$, where σ_Z^2 can be computed as the variance of the available samples in set $\mathcal{S}$.
2) A diagonal bandwidth (K-MMSE$_D$): C_{ZZ} is forced to be diagonal by retaining only the diagonal elements of the sample covariance matrix.
3) A complete bandwidth (K-MMSE$_C$): C_{ZZ} is employed as directly obtained from the observed data, as in [18].

Note that if the bandwidth is scalar or diagonal, then $\beta_{XY} = 0$, $\alpha = 0$, and the second term in Eq. (17) is also zero. Then, the K-MMSE estimator is reduced to a multivariate Nadaraya-Watson estimator. Therefore, only K-MMSE$_C$ will be able to exploit the cross-correlations between context $\boldsymbol{y}$ and vector $\boldsymbol{x}$. A particular case of K-MMSE$_C$ corresponds to the use of a single scale factor, that is, $\beta_{XY} = \beta_{YY} = \beta$, which involves $\alpha = 1$ [16].

C. Bandwidth parameter estimation

As mentioned above, the bandwidth scaling scheme proposed in Eq. (16) reduces our BE problem to the estimation of two parameters, α and β. Let us consider first the estimation of β. According to the LMMSE approximation criterion discussed in the previous subsection, $\tilde{\boldsymbol{y}}_0$ should be as close as possible to the point of interest $\boldsymbol{y}_0$. Since the prediction vector $\tilde{\boldsymbol{y}}_0$ only depends on the scale factor β (as deduced from Eq.(9) by making $\mathcal{H}_{YY} = \beta C_{YY}$), then this parameter can be obtained from the minimization of the following prediction square error,

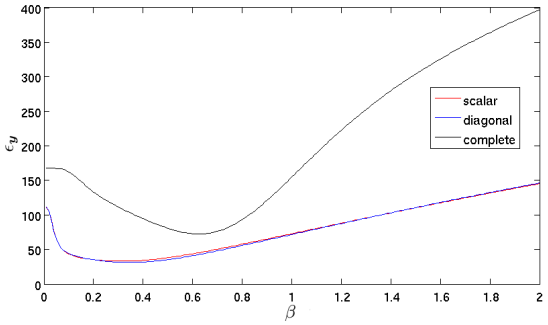


Fig. 3. Example of $\epsilon_{\boldsymbol{y}} = \epsilon_{\boldsymbol{y}}(\beta)$ using scalar, diagonal and complete bandwidths.

$$\epsilon_{\boldsymbol{y}}(\beta) = \|\boldsymbol{y}_0 - \tilde{\boldsymbol{y}}_0\|^2 = \left\| \boldsymbol{y}_0 - \sum_{j=1}^{M} w_j \boldsymbol{y}_j \right\|^2. \tag{18}$$

The square error function $\epsilon_{\boldsymbol{y}}(\beta)$ defined above is a non linear function whose minimization can be easily solved by any of the multiple optimization algorithms that can be found in the literature. However, if the expected range of variation of the scale factor β is small, an exhaustive search in this range will be more efficient. In particular, this will be the case of the image reconstruction experiments developed in the following section. Another argument in favour of an exhaustive search is that it can be efficiently implemented [16]. In order to do so, we can define a set of auxiliary weights as follows,

$$\tilde{w}_j(\boldsymbol{y}_0) = \exp\left(-\frac{1}{2} (\boldsymbol{y}_0 - \boldsymbol{y}_j)^t C_{YY}^{-1} (\boldsymbol{y}_0 - \boldsymbol{y}_j) \right). \tag{19}$$

These auxiliary weights correspond to those of Eq. (9) except for the contribution of the scale factor β, which has been removed. Therefore, they do not depend on β and can be precomputed. Then, during the exhaustive search, the weights (Eq.(9)) for every value of β can be efficiently obtained as,

$$w_j(\beta; \boldsymbol{y}_0) = \frac{(\tilde{w}_j(\boldsymbol{y}_0))^{1/\beta}}{\sum_{i=1}^{M} (\tilde{w}_i(\boldsymbol{y}_0))^{1/\beta}}. \tag{20}$$

Figure 3 shows an example of the error curve $\epsilon_{\boldsymbol{y}} = \epsilon_{\boldsymbol{y}}(\beta)$ for the three types of bandwidth. Scalar and diagonal bandwidths produce almost identical square errors since $\boldsymbol{\mu}_{X|Y}^{(j)}(\boldsymbol{y}_0) = \boldsymbol{x}_j$ for both cases and the elements in the diagonal of the correlation matrix C_{ZZ} tend to be equal. Typically, scalar and diagonal bandwidths will involve smaller minima than complete bandwidths as illustrated in the figure. This can be explained by the fact that, in this last case, weights w_j (Eq. (4)) are computed according to the Mahalanobis distance (through the complete bandwidth matrix) unlike the case of a scalar bandwidth, which involves an Euclidean distance, which, in fact, is the minimization criterion. In other words, scalar bandwidths are more coherent with the minimization of the square error $\epsilon_{\boldsymbol{y}}$ (which, in turn, is directly associated to a PSNR measure). However, as we will show later in section V-A, a scalar bandwidth tends to smooth high

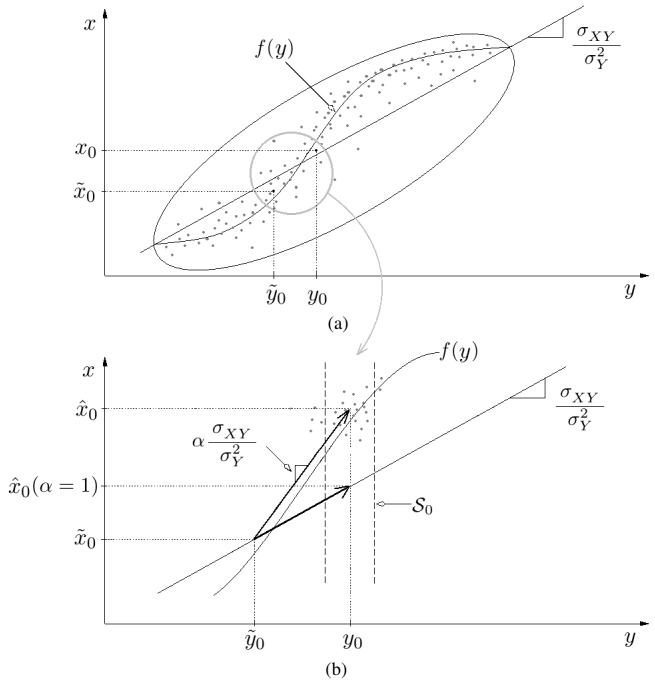


Fig. 4. Example of correction vectors corresponding to β and α. The situation in (a) is enlarged in (b). $\mathcal{S}_0$ is a set of observed vectors close to the original vector $\boldsymbol{z}_0$

frequency textures while complete bandwidth matrices will be capable of recovering fine textures through the exploitation of intercorrelations.

Once β and the prediction vectors, $\tilde{\boldsymbol{y}}_0$ and $\tilde{\boldsymbol{x}}_0$, have been obtained, let us consider the other bandwidth parameter, α. This parameter controls the contribution of the second term in Eq. (17). In order to illustrate the effect of this second term, let us consider the example of Fig. 4(a), where $\boldsymbol{x}$ and $\boldsymbol{y}$ are both monodimensional random variables related by a function $f(\cdot)$ plus a zero-mean noise ε, that is,

$$x = f(y) + \varepsilon. \tag{21}$$

The ellipse in Fig. 4(a) represents a Gaussian model of the observed data $\boldsymbol{z}_j$ $(j = 1, \ldots, M)$ with covariance C_{ZZ} defined by variances σ_X^2 and σ_Y^2, and cross-covariance σ_{XY}. Let us consider first that C_{ZZ} is scaled by a single scale factor β, that is, $\mathcal{H} = \beta C_{ZZ}$, so $\alpha = 1$. Then, according to Eq. (17), the initial NW estimate $(\tilde{x}_0, \tilde{y}_0)$ is corrected along a line with slope σ_{XY}/σ_Y^2. This is illustrated in Fig. 4(b). The approximation $\alpha = 1$ is, in general, inaccurate since the local slope of $f(y)$ around y_0 may be considerably different. Then, factor α allows a better approximation of $f(y)$ in the neighbourhood of the initial estimate $(\tilde{x}_0, \tilde{y}_0)$ (as illustrated in Fig. 4(b)) since it controls the (hyper)direction in which the linear correction term in Eq. (17) is applied.

In order to estimate α, we consider again the comparison with the LMMSE estimator made in section IV-A. In our submatrix scaling framework, we see that α can be employed to make $T_{\mathcal{H}} = \alpha C_{XY} C_{YY}^{-1}$ as close as possible to T_Σ. As mentioned previously, T_Σ is not known, so α cannot be directly optimized. However, since β and the prediction vectors have been already obtained, it is possible to estimate α by minimizing the following approximation to the MSE of Eq. (11) in the vicinity of $\boldsymbol{y}_0$,

$$\epsilon_{\boldsymbol{x}}(\alpha) = \frac{1}{|\mathcal{I}_0|} \sum_{i \in \mathcal{I}_0} \|\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i\|^2 = \frac{1}{|\mathcal{I}_0|} \sum_{i \in \mathcal{I}_0} (\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i)^t (\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i) \tag{22}$$

where

$$\hat{\boldsymbol{x}}_i = \tilde{\boldsymbol{x}}_0 + \alpha C_{XY} C_{YY}^{-1} (\boldsymbol{y}_i - \tilde{\boldsymbol{y}}_0) \tag{23}$$

and

$$\begin{array}{rcl} \mathcal{I}_0 & = & \{i \in \{1, ..., M\} | \boldsymbol{y}_i \in \mathcal{S}_0\}, \\ \mathcal{S}_0 & = & \{\boldsymbol{y}_i | \boldsymbol{y}_i \text{ close to } \boldsymbol{y}_0, i = 1, ..., M\}. \end{array} \tag{24}$$

Finally, solving $d\epsilon_{\boldsymbol{x}}/d\alpha = 0$, we obtain the following analytical solution for α,

$$\alpha = \frac{\sum_{i \in \mathcal{I}_0} (\boldsymbol{x}_i - \tilde{\boldsymbol{x}}_0)^t C_{XY} C_{YY}^{-1} (\boldsymbol{y}_i - \tilde{\boldsymbol{y}}_0)}{\sum_{i \in \mathcal{I}_0} (\boldsymbol{y}_i - \tilde{\boldsymbol{y}}_0)^t \left(C_{XY} C_{YY}^{-1}\right)^t \left(C_{XY} C_{YY}^{-1}\right) (\boldsymbol{y}_i - \tilde{\boldsymbol{y}}_0)}. \tag{25}$$

Defining the set $\mathcal{S}_0$ of neighbouring vectors is not a straightforward task since several criteria are possible. In the rest of the work we will consider $\mathcal{S}_0$ as the set of the N closest context vectors $\boldsymbol{y}_j$ to $\boldsymbol{y}_0$, according to the Euclidean distance. The selection of N will be discussed in the next section.

Finally, the reconstruction algorithm is summarized as follows:

1) Compute β by minimizing $\epsilon_{\boldsymbol{y}}(\beta)$ employing the auxiliary weights of Eq. (20).
2) Define set $\mathcal{S}_0$. The $N = |\mathcal{S}_0|$ closest neighbouring vectors are used for the estimation of α.
3) Compute α according to Eq. (25).
4) Finally, the missing sample vector $\boldsymbol{x}_0$ is reconstructed according to Eq. (17).

V. Implementation and experimental results for image error concealment

A. K-MMSE estimator implementation

In this section we will use and test the techniques developed in the previous sections for the particular application of image error concealment. As mentioned in section IV-B, we will consider three different types of sample covariance matrices (which have been adopted as structure matrices) which yield three different types of K-MMSE estimators. Thus, we have K-MMSE$_S$ for a scalar bandwidth, K-MMSE$_D$ for a diagonal covariance matrix, and K-MMSE$_C$ for a complete covariance matrix. K-MMSE$_S$ and K-MMSE$_D$ are two types of NW estimators since they do not use the second term of Eq. (17), so the only bandwidth parameter to be estimated is β. As also mentioned in section IV-B, only complete bandwidth matrices can exploit correlations among pixels inside vector $\boldsymbol{z}$. This fact will allow better reconstructions than those provided by scalar or diagonal bandwidths as illustrated in the example of Fig. 5.

First, let us analyze the selection of $\mathcal{S}_0$. On one hand, N should be as small as possible in order to select an homogeneous set of observations. On the other hand, the more vectors are employed the better is the average computed in Eq.(22).

(a) (b) (c)

Fig. 5. Subjective comparison for a fraction of *Goldhill*. (a) Received data. (b) Reconstruction using scalar bandwidth (K-MMSE$_S$). (c) Reconstruction using complete bandwidth matrix (K-MMSE$_C$). The most outstanding differences are marked with yellow boxes.

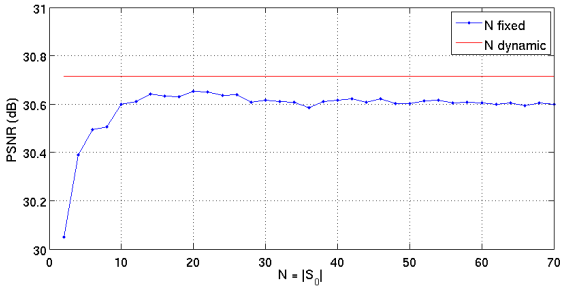


Fig. 6. Average PSNR (in dB) for the Kodak set using fixed and dynamic values of $N = |\mathcal{S}_0|$. Dispersed error pattern is employed.

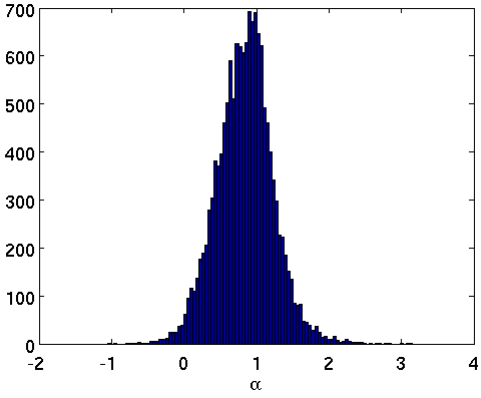


Fig. 7. Histogram of α for K-MMSE$_C$ obtained for the image of *Lena*

The selection of size N will require a compromise between these two facts. Let N_z, N_x and N_y be the dimensionality of $\boldsymbol{z}$, $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively ($N_x = 4$ and $N_y = 16$ to 32). The estimated value of $\boldsymbol{x}_0$ belongs to an N_x-dimensional subspace (within the N_z-dimensional space) defined by the known coordinates $\boldsymbol{y}_0$ (note that $N_z = N_x + N_y$). Thus, from a geometric point of view, since we need a single point in an N_z-dimensional space, we are looking for the intersection of this N_x-dimensional subspace with an N_y-dimensional subspace defined by the points within the set $\mathcal{S}_0$. The minimum number of points required to define an N_y-dimensional subspace is $N_y + 1$. Thus, in order to get an homogeneous set of observations we will dynamically employ the $N = N_y + 1$ closest observations to $\boldsymbol{y}_0$. Figure 6 shows the average PSNR computed over the validating set of 24 images by Kodak [23] using dispersed error pattern. We compare the performance when a fixed value of N is employed and our proposal of dynamic selection. It can be seen that defining $\mathcal{S}_0$ dynamically with $N = N_y + 1$ yields the best performance. It can be also observed that $N = 20$ achieves the best result when using a fixed value for N. It is worth mentioning that the average value of N_y over the employed images is 19.

Although the most general form proposed for K-MMSE$_C$ requires the estimation of the two bandwidth parameters, α and β, we will also consider the simplified case where α is not estimated by fixing it to 1. This approximation is supported by the histogram of Fig. 7 where we can observe that typical values of α are distributed around 1. This case will be referred to as K-MMSE$_C^1$ and it is equivalent to considering that the scale factor for C_{XY} is the same as that computed for C_{YY}, that is, $\beta_{XY} = \beta$ as already mentioned in section IV-B.

In subsection IV-C we discussed the mismatch problem that arose from the fact that, in the case of K-MMSE$_C$, the weights w_j were computed with a Mahalanobis distance criterion while we are evaluating performance through PSNR (which involves an Euclidean distance criterion). This mismatch only disappears in the case of scalar bandwidth, which is coherent with the Euclidean distance criterion. For the case of K-MMSE$_C$, this mismatch problem can be palliated by applying a combined procedure as follows. First, we compute $\tilde{\boldsymbol{x}}_0$ using a scalar bandwidth, that is, the weights w_j are computed according to the Euclidean distance, which is directly related to the square error criterion $\epsilon_{\boldsymbol{y}}$. Then, a complete bandwidth matrix is used just to compute the correction term of Eq. (17). This new approach will be referred to as K-MMSE$_+$, and K-MMSE$_+^1$ for the case of $\alpha = 1$ fixed.

As already mentioned, K-MMSE$_+$ is more PSNR oriented than K-MMSE$_C$. Therefore, it can be expected that its initial estimate $\tilde{\boldsymbol{y}}_0$ is considerably closer to $\boldsymbol{y}_0$. This means that the correction term in Eq. (17) is expected to be relatively small and, therefore, the use of α provides only minor improvements. This is not the case of K-MMSE$_C$, where the distance between $\tilde{\boldsymbol{y}}_0$ and $\boldsymbol{y}_0$ can be significant and the correction term should lead to more important improvements. In fact, our simulations show that the average power of the correction term (averaged over all the tested images, see Section V-B) in the case of K-MMSE$_C$ is more than four times larger than that of K-MMSE$_+$. An initial estimate $\tilde{\boldsymbol{y}}_0$ as close as possible to $\boldsymbol{y}_0$ was also one of the conditions deduced in section IV-A for an accurate K-MMSE estimate. However, this condition, directly related to PSNR, may lead to wasting useful correlations among pixels. Employing PSNR as an objective visual similarity criterion can neglect spatial correlations in order to minimize the square error. As anticipated in Fig. 5, this can yield poorer texture reconstructions. This issue will be analyzed in more detail in the next subsection.

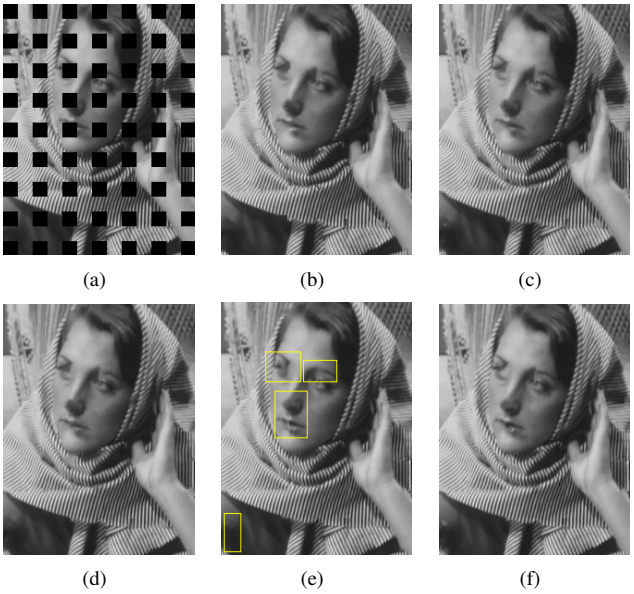


Fig. 8. Subjective comparison for a fraction of *Barbara*. (a) Received data. (b) Reconstruction by SLP-E. (c) Reconstruction by K-MMSE$_S$. (d) Reconstruction by K-MMSE$_D$. (e) Reconstruction by K-MMSE$_C$. (f) Reconstruction by K-MMSE$_+$.

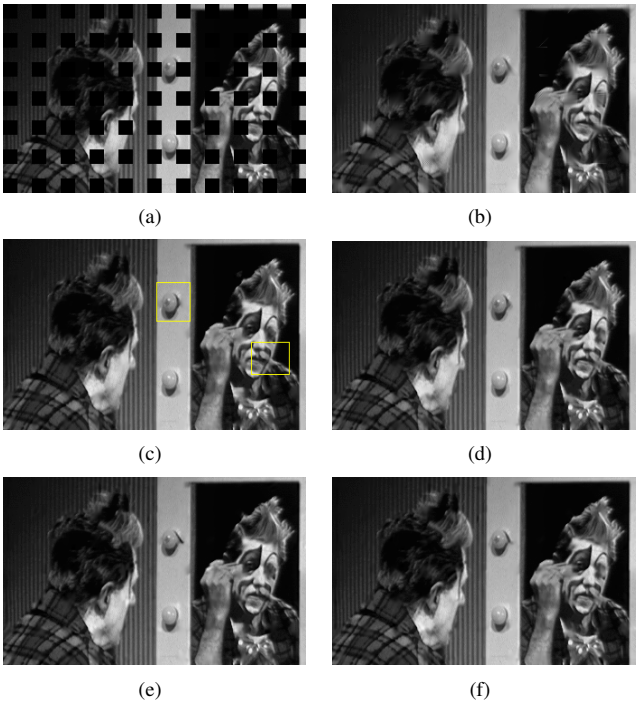


Fig. 9. Subjective comparison for a fraction of *Clown*. (a) Received data. (b) Reconstruction by OAI. (c) Reconstruction by K-MMSE$_C$. (d) Reconstruction by K-MMSE$_+$. (e) Reconstruction by K-MMSE^1_C. (f) Reconstruction by K-MMSE$^1_+$.

B. Experimental results

The performance of the reconstruction techniques presented in the previous sections will be evaluated over the spatial error concealment application described in Section II for block-coded images. The testing images are: *Lena* (512×512), *Goldhill* (720×576), *Foreman* (352×288), *Barbara* (512×512), *Baboon* (512×512), *Clown* (512×512), *Tire* (205×232), *Pirate* (1024×1024), *Boat* (512×512) and *Peppers* (384×512). In the simulations, β is searched exhaustively within the range $[0, 2]$ with steps of 0.01.

We compare the performance with other spatial EC methods, namely projections onto convex sets (POC) [24], directional extrapolation (EXT) [25], a Hough transform based SEC (HTS) [26], content adaptive technique (CAD) [27], non-normative SEC for H.264/AVC (AVC) [28], multi-dimensional adaptive SEC (MDA) [29], Markov random fields approach (MRF) [30], inpainting (INP) [31], bilateral filtering (BLF) [14], edge recovery technique based on visual clearness (EVC) [32], orientation adaptive interpolation (OAI) [33], frequency selective extrapolation (FSE) [34] and our previous exponential approximation to sparse linear prediction (SLP-E) [15].

Additionally, we are interested in showing two important advantages of our approach for signal reconstruction. First, that our BE proposal is superior to a classical adaptive plug-in BE approach. Thus, we have also tested our K-MMSE estimator with the BE technique described in reference [18] (labelled as KDE in Table I). Also, we want to demonstrate that our non-parametric K-MMSE approach can capture local statistics more suitably (and provide better results) than a classical MMSE estimator based on off-line statistics. Thus, we have adapted the GMM-based MMSE estimator described in reference [10] to the multidimensional formalism described in Section II: vectors have 6×6 dimensions which are adapted to every specific estimate by marginalization [10]. Since we are interested in checking whether this GMM/MMSE estimator could surpass our K-MMSE proposal independently of the suitability of the training data, the GMM has been trained with 529937 vectors obtained from the original (not corrupted) testing images, so that we can consider it a sort of oracle GMM model. From our experiments, we have observed that the performance of this method converges for 1024 Gaussians. Also, note that this number approximates the number of observations employed by our K-MMSE methods. Thus, this is the number of Gaussians employed for comparison purposes. The resulting technique will be referred to as GMM in the following. [1]

Table I shows a PSNR comparison of the tested techniques for dispersed and random losses. Figures 8 and 9 show some reconstruction examples in order to assess the subjective performance of our proposals. It can be observed that our K-MMSE proposals outperform other state-of-the art techniques in terms of PSNR. The comparison with KDE and GMM is particularly interesting since it confirms that K-MMSE can better capture the local signal statistics. Also, as expected, we observe that K-MMSE$_S$ and K-MMSE$_D$ lead to almost identical results and that the best performance is achieved by K-MMSE$_+$ and K-MMSE$_C$.

Since K-MMSE$_S$ provides the initial estimate $(\tilde{\boldsymbol{x}}_0, \tilde{\boldsymbol{y}}_0)$ for K-MMSE$_+$, we can say that the application of the correction term in Eq. (17) helps to improve the reconstruction quality

[1] Implementations of most of these techniques, as well as the implementation of our algorithm, is available online at [35].

SEC	*Lena*		*Goldhill*		*Foreman*		*Barbara*		*Baboon*		*Clown*		*Tire*		*Pirate*		*Boat*		*Peppers*		Average	
	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)
POC	28.04	-	28.50	-	28.49	-	24.30	-	24.63	-	24.36	-	23.92	-	26.42	-	26.05	-	28.68	-	26.34	-
EXT	29.10	-	29.57	-	29.59	-	25.62	-	24.76	-	26.30	-	23.40	-	27.57	-	26.63	-	30.76	-	27.33	-
HTS	30.55	-	29.97	-	28.09	-	26.41	-	24.11	-	27.62	-	24.25	-	28.12	-	27.20	-	29.72	-	27.60	-
CAD	31.96	-	30.24	-	34.85	-	27.39	-	25.16	-	29.12	-	27.53	-	28.44	-	27.73	-	32.20	-	29.46	-
AVC	30.42	-	31.27	-	29.11	-	26.85	-	25.42	-	28.55	-	24.99	-	28.74	-	28.25	-	32.13	-	28.57	-
MDA	32.80	-	31.56	-	31.79	-	27.81	-	25.77	-	28.80	-	26.36	-	29.36	-	28.62	-	33.43	-	29.63	-
MRF	32.17	31.10	31.12	29.81	32.98	31.07	27.99	27.67	26.14	26.12	28.23	27.47	27.00	26.09	29.52	28.13	27.91	27.00	32.59	28.27	29.57	28.27
INP	30.85	30.01	30.40	29.06	34.44	31.65	28.03	27.86	25.10	24.97	27.89	27.14	27.33	25.23	28.44	27.35	27.79	26.72	32.13	27.35	29.23	27.94
BLF	32.15	30.59	30.91	29.40	34.75	28.72	29.91	28.19	26.05	25.68	28.73	26.83	28.77	26.17	29.36	27.38	28.37	26.92	33.17	27.38	30.22	27.91
EVC	32.70	-	31.66	-	35.09	-	28.41	-	26.00	-	29.51	-	27.58	-	29.90	-	28.66	-	33.29	-	30.28	-
OAI	32.82	30.47	31.54	28.59	35.03	27.05	29.66	27.51	26.06	24.39	29.75	27.26	27.42	26.75	29.90	27.30	29.50	26.69	34.84	30.38	30.35	27.64
FSE	32.72	31.32	31.78	30.26	34.18	31.70	30.84	29.47	26.02	25.88	29.19	27.76	28.31	27.38	29.64	28.01	28.87	27.61	33.48	30.56	30.50	29.00
SLP-E	32.36	31.61	31.72	30.29	35.78	32.75	30.80	30.13	26.02	25.49	29.70	27.88	28.33	26.42	29.63	28.44	28.54	27.70	33.94	30.61	30.68	29.13
GMM	31.70	30.74	31.82	30.43	31.82	30.49	27.46	27.64	25.60	25.50	30.12	28.33	28.56	27.23	30.37	28.37	28.94	28.08	34.64	**32.57**	30.10	28.99
KDE	32.22	31.27	31.43	30.01	35.57	31.09	30.84	29.02	24.68	25.96	29.98	27.45	28.22	26.20	29.51	28.20	28.77	27.39	33.86	29.93	30.51	28.65
K-MMSE$_S$	32.84	32.22	32.03	30.47	36.16	32.48	31.33	30.97	26.15	26.14	30.79	28.12	28.62	28.82	30.15	28.77	29.34	28.44	34.59	31.27	31.20	29.77
K-MMSE$_D$	32.87	32.23	32.01	30.44	36.18	32.69	31.35	30.87	26.14	26.13	30.83	28.13	28.58	28.52	30.16	28.76	29.36	28.47	34.55	31.23	31.21	29.78
K-MMSE$_C$	**33.08**	32.60	**32.17**	30.54	**36.21**	**33.71**	32.00	**31.28**	26.31	26.16	**31.28**	28.38	29.00	28.30	30.17	**28.97**	**29.90**	28.39	34.70	32.20	**31.48**	**30.06**
K-MMSE$_+$	32.96	**32.70**	32.08	30.37	36.18	33.71	**32.25**	31.11	26.37	26.20	31.00	28.32	**29.13**	27.98	**30.39**	28.90	29.50	**28.48**	35.00	**32.52**	**31.48**	30.04
K-MMSE$_C^1$	32.69	31.86	32.14	**30.70**	35.69	33.79	31.77	31.01	26.26	26.31	30.82	**28.41**	28.55	**28.86**	30.12	28.68	29.70	28.17	34.68	32.07	31.24	29.93
K-MMSE$_+^1$	33.00	32.32	**32.16**	30.58	35.89	33.95	32.22	30.79	**26.38**	**26.36**	30.99	27.91	28.54	27.83	**30.39**	28.90	29.64	28.37	**35.07**	32.39	31.43	29.94

TABLE I
PSNR VALUES (IN DB) FOR TEST IMAGES RECONSTRUCTED BY SEVERAL ALGORITHMS FOR BLOCK DIMENSIONS 16×16. DISPERSED ERROR PATTERN (A) AND RANDOM LOSSES (B) ARE APPLIED. THE BEST PERFORMANCES FOR EACH IMAGE ARE IN BOLD FACE.

in this case. On the other hand, K-MMSE$_C$ can achieve similar or even better results while providing higher quality reconstructions of complicated textures and fine details as shown in Figs. 5(c) or 8(e) and as already remarked in the previous subsection.

Figure 9 shows that assuming $\alpha = 1$ can be a reasonable simplification for the case of K-MMSE$_+$ since the reconstruction quality and the PSNR are deteriorated negligibly. However, this approximation yields a more prejudicial effect over K-MMSE$_C$, as reflected also in Table I.

SLP-E	GMM	KDE	K-MMSE$_S$	K-MMSE$_C$	K-MMSE$_C^1$
1.00	9.53	695.46	3.56	4.04	3.87

TABLE II
AVERAGE ERROR CONCEALMENT TIME PER MACROBLOCK COMPARED TO SLP-E.

Finally, regarding the computational complexity, Table II shows the processing time ratios of GMM, KDE, K-MMSE$_S$ (the simplest proposal), K-MMSE$_C$ (the most complex one) and K-MMSE$_C^1$ to SLP-E. It follows that our proposals are computationally more expensive than SLP-E since bandwidth estimation is involved. This complexity increase is reflected on the reconstruction quality which is improved by up to 1dB on average, as shown in Table I. On the other hand, our proposals require less than half the processing time with respect to GMM. Classic KDE is, by far, the most time consuming technique and even impractical for real applications.

VI. CONCLUSIONS

In this paper, we have proposed a multidimensional kernel-based MMSE technique for multimedia signal reconstruction. The most important issue in any KDE problem is the selection of a suitable bandwidth. Thus, a procedure for bandwidth estimation especially oriented to this reconstruction task has been proposed. Our proposal introduces a bandwidth matrix which is built by scaling a structure matrix which, in our case, is the covariance of the locally available data. The autocovariance and the cross-covariance submatrices of this structure matrix are scaled independently, and each scale factor is estimated by minimizing a specific type of reconstruction error. The corresponding optimization criteria have been derived from a comparison of the proposed K-MMSE estimator with an LMMSE estimator. Our proposal has been tested over an image error concealment application. An average improvement of up to 1dB is achieved with respect to a classical plug-in bandwidth estimation. The improvement is even larger with respect to a classical GMM-based MMSE reconstruction. Finally, several simplifications, affecting the structure matrix and/or the bandwidth estimation procedure, have been applied. Some of these simplifications have been shown to provide PSNR results similar to the complete formulation, although with a noticeable degradation in the reconstruction of fine textures.

Ongoing work is focused on the application of the proposed kernel-based MMSE estimation to video reconstruction.

REFERENCES

[1] R. Duda, P. Hart, and D. Stork, "Pattern classification," John Wiley, 2000.

[2] A. Elgammal, R. Duraiswami, D. Harwood, and L. Davis, "Background and foreground modeling using nonparametric kernel density estimation for visual surveillance," in *Proceedings of IEEE*, vol. 90, 2002, pp. 1151–1163.

[3] D. Comaniciu, "An algorithm for data-driven bandwidth selection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, pp. 281–288, February 2003.

[4] V. Katkovnik, "A new method for varying adaptive bandwidth selection," *IEEE Transactions on Signal Processing*, vol. 47, pp. 2567–2571, September 1999.

[5] K. Zhang, X. Gao, D. Tao, and X. Li, "Single image super-resolution with-non local means and steering kernel regression," *IEEE Transactions on Image Processing*, vol. 21, pp. 4544–4556, November 2012.

[6] H. Takeda, S. Farsiu, and P. Milanfar, "Kernel regression for image processing and reconstruction," *IEEE Transactions on Image Processing*, vol. 16, pp. 349–366, February 2007.

[7] B. Silverman, "Density estimation for statistics and data analysis," Chapman & Hall, 1986.
[8] S. Sheather, "Density estimation," *Statistical Science*, vol. 19, no. 4, pp. 588–597, 2004.
[9] D. Scott, "Multivariate density estimation: Theory, practice, and visualization," Wiley, 1992.
[10] D. Persson, T. Eriksson, and P. Hedelin, "Packet video error concealment with gaussian mixture models," *IEEE Transactions on Image Processing*, vol. 17, pp. 145–154, 2008.
[11] G. Zhai, X. Yang, W. Lin, and W. Zhang, "Bayesian error concealment with DCT pyramid for images," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, pp. 1224–1232, September 2010.
[12] A. Bors and N. Nasios, "Kernel bandwidth estimation for nonparametric modeling," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 39, pp. 1543–1555, December 2009.
[13] J. Marron and A. Tsybakov, "Visual error criteria for qualitative smoothing," *Journal of the American Statistical Association*, vol. 90, pp. 499–507, June 1995.
[14] G. Zhai, J. Cai, W. Lin, X. Yang, and W. Zhang, "Image error-concealment via block-based bilateral filtering," *IEEE International Conference on Multimedia and Expo*, pp. 621–624, June 2008.
[15] J. Koloda, J. Østergaard, S. H. Jensen, V. Sánchez, and A. M. Peinado, "Sequential error concealment for video/images by sparse linear prediction," *IEEE Transactions on Multimedia*, vol. 15, pp. 957 – 969, June 2013.
[16] J. Koloda, A. Peinado, and V. Sanchez, "On the application of multivariate kernel density estimation to image error concealment," in *Proceedings of ICASSP*, May 2013.
[17] L. Yang and R. Tschernig, "Multivariate bandwidth selection for local linear regression," *Journal of the Royal Statistical Society*, vol. 61, pp. 793–815, 1999.
[18] M. Kristan, A. Leonardis, and D. Skočaj, "Multivariate online kernel density estimation with gaussian kernels," *Pattern Recognition*, vol. 44, pp. 2630–2642, 2011.
[19] M. Schimek, "Smoothing and regression: Approaches, computation and application," Wiley, 2000.
[20] J. Flam, S. Chatterjee, K. Kansanen, and T. Ekman, "On MMSE estimation: A linear model under gaussian mixture statistics," *IEEE Transactions on Signal Processing*, vol. 60, no. 7, pp. 3840 – 3845, July 2012.
[21] S. Kay, "Fundamentals of statistical signal processing," Prentice-Hall, 1998.
[22] M. Wand and M. Jones, "Comparison of smoothing parameterizations in bivariate kernel density estimation," *Journal of the American Statistical Association*, vol. 88, no. 422, pp. 520–528, 1993.
[23] "Kodak test images," http://r0k.us/graphics/kodak/.
[24] H. Sun and W. Kwok, "Concealment of damaged block transform coded images using projections onto convex sets," *IEEE Transactions on Image Processing*, vol. 4, no. 4, pp. 470–477, April 1995.
[25] Y. Zhao, H. Chen, X. Chi, and J. Jin, "Spatial error concealment using directional extrapolation," in *Proceedings of DICTA*, 2005, pp. 278–283.
[26] H. Gharavi and S. Gao, "Spatial interpolation algorithm for error concealment," in *Proceedings of ICASSP*, April 2008, pp. 1153–1156.
[27] Z. Rongfu, Z. Yuanhua, and H. Xiaodong, "Content-adaptive spatial error concealment for video communication," *IEEE Transactions on Consumer Electronics*, vol. 50, pp. 335–341, February 2004.
[28] V. Varsa and M. Hannuksela, "Non-normative error concealment algorithms," *ITU-T SG16, VCEG-N62*, vol. 50, September 2001.
[29] H. Asheri, H. Robiee, N. Pourdamghani, and M. Ghanbari, "Multi-directional spatial error concealment using adaptive edge thresholding," *IEEE Transactions on Consumer Electronics*, vol. 58, pp. 880–885, August 2012.
[30] S. Shirani, F. Kossentini, and R. Ward, "An adaptive Markov random field based error concealment method for video communication in error prone environment," in *Proceedings of ICIP*, vol. 6, 1999, pp. 3117–3120.
[31] P. Harrison, "Texture synthesis, texture transfer and plausible restoration," Ph.D. dissertation, Monash University, 2005.
[32] J. Koloda, V. Sánchez, and A. M. Peinado, "Spatial error concealment based on edge visual clearness for image/video communication," *Circuits, Systems and Signal Processing*, vol. 32, pp. 815–824, April 2013.
[33] X. Li and M. Orchard, "Novel sequential error-concealment techniques using orientation adaptive interpolation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, pp. 857–864, October 2002.
[34] J. Seiler and A. Kaup, "Fast orthogonality deficiency compensation for improved frequency selective image extrapolation," in *Proceedings of ICASSP*, March 2008, pp. 781–784.
[35] [Online], *Available: http://dtstc.ugr.es/~jkoloda/download.html*.

Ján Koloda received the M.Sc. degree in telecommunications engineering from the University of Granada, Granada, Spain, in 2009. He is currently working towards the Ph.D. degree on error concealment algorithms for block-coded video. Since 2010, he has been with the Research Group on Signal Processing, Multimedia Transmission and Speech/Audio Technologies, Department of Signal Theory, Networking and Communications at the University of Granada, under a research grant. He has been a visiting researcher at Aalborg University, Aalborg, Denmark and at Friedrich-Alexander University, Erlangen, Germany. His research interests are in the area of error concealment of block-coded video sequences, image and signal processing.

Antonio M. Peinado (M'95-SM'05) received the M.S. and the Ph.D. degrees in Physics from the University of Granada, Granada, Spain, in 1987 and 1994, respectively. Since 1988, he has been working at the University of Granada, where he has led several research projects related to signal processing and transmission. In 1989, he was a Consultant at the Speech Research Department, AT&T Bell Labs. He earned the positions of Associate Professor (1996) and Full Professor (2010) in the Department of Signal Theory, Networking and Communications, University of Granada, and is currently head of the research group on Signal Processing, Multimedia Transmission and Speech/Audio Technologies (SigMAT) at the same university. He is the author of numerous publications and coauthor of the book Speech Recognition over Digital Channels (Wiley, 2006), and has served as reviewer for several international journals, conferences and project proposals. His current research interests are focused on robust speech recognition and transmission, robust image/video transmission, and ultrasound signal processing.

Victoria Sánchez (M'95) received the M.S. and the Ph.D. degrees from the University of Granada, Granada, Spain, in 1988 and 1995, respectively. In 1988, she joined the Signal Processing and Communications department of the University of Granada where she is currently a member of the research group on Signal Processing, Multimedia Transmission and Speech/Audio Technologies. During 1991, she was visiting with the Electrical Engineering Department, University of Sherbrooke, Canada. Since 1997, she is an Associate Professor at the University of Granada. Her research interests include speech and audio processing, multimedia transmission and speech recognition. She has authored over 60 journal articles and conference papers in these fields.

2.3 EC in transformed domain

The paper associated to this part is:

2.3.1 Frequency Selective Extrapolation with Residual Filtering for Image Error Concealment

- J. Koloda, J. Seiler, A. Kaup, V. Sánchez and A.M. Peinado, "Frequency Selective Extrapolation with Residual Filtering for Image Error Concealment", *in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1995-1999, Florence, Italy, May 2014.
    - Status: Published.

FREQUENCY SELECTIVE EXTRAPOLATION WITH RESIDUAL FILTERING FOR IMAGE ERROR CONCEALMENT

Ján Koloda†, Jürgen Seiler‡, André Kaup‡, Victoria Sánchez† and Antonio M. Peinado†

†Dpt. of Signal Theory, Networking and Communications
Universidad de Granada, Spain
{janko, victoria, amp}@ugr.es

‡Chair of Multimedia Communications and Signal Processing
University of Erlangen-Nuremberg, Germany
{seiler, kaup}@LNT.de

ABSTRACT

The purpose of signal extrapolation is to estimate unknown signal parts from known samples. This task is especially important for error concealment in image and video communication. For obtaining a high quality reconstruction, assumptions have to be made about the underlying signal in order to solve this underdetermined problem. Among existent reconstruction algorithms, frequency selective extrapolation (FSE) achieves high performance by assuming that image signals can be sparsely represented in the frequency domain. However, FSE does not take into account the low-pass behaviour of natural images. In this paper, we propose a modified FSE that takes this prior knowledge into account for the modelling, yielding significant PSNR gains.

Index Terms— Image processing, error concealment

1. INTRODUCTION

Signal reconstruction is a very challenging task for many multimedia applications where the quality of the received data is of utmost importance. A common example is the transmission of image/video signals over error prone channels which may yield block losses. The lost areas need to be concealed employing the information provided by the correctly received data. There are several examples of efficient error concealment (EC) techniques applied to image communication. The EC algorithm proposed in [1] is based on Markov random fields and focuses on preserving visually important features, such as edges. Bilateral filtering that exploits a pair of gaussian kernels is treated in [2]. In [3], the lost region is recovered through sparse linear prediction. Moreover, inpainting [4] can also be employed for concealment purposes.

An alternative approach to image EC is the frequency selective extrapolation (FSE) proposed in [5]. In particular, the complex-valued FSE implementation [6] can provide high quality reconstructions with a low computational burden. This technique develops a signal model from the set of Fourier basis functions which can be used to replace the unknown samples. Although this FSE algorithm basically consists in determining frequency components, it does not exploit any a priori knowledge regarding the typical spectrum of natural images, which may result in high-frequency artifacts. In this paper, we propose the introduction of a low-pass filtering in the FSE iterative procedure which can efficiently account for this fact, increasing the FSE performance while maintaining a low computational cost.

The paper is organized as follows. In Section 2, we provide a short review of the FSE algorithm. Our proposal, based on residual filtering, is described in Section 3. Experimental results are discussed in Section 4. The last section is devoted to conclusions.

2. FREQUENCY SELECTIVE EXTRAPOLATION

Our proposal is a modification of the complex-valued implementation of FSE [6]. This approach is able to robustly reconstruct various image contents at very high quality [6, 5, 7]. We briefly summarize it in this section.

During the extrapolation process of FSE, the image is divided into blocks of equal size. Besides the block actually containing areas to be reconstructed, neighbouring samples belonging to adjacent blocks are taken into account, as well. All the considered samples make up the so called extrapolation area $\mathcal{L}$ (an example is shown in Fig. 1). The size of area $\mathcal{L}$ is $M \times N$ samples and the signals in this area are indexed by spatial variables m and n. All samples in area $\mathcal{L}$ belong to one of the three following groups: the known samples built up support area $\mathcal{A}$, all unknown samples belong to the loss area $\mathcal{B}$ (located at the centre of $\mathcal{L}$) and all samples from neighbouring blocks that have been extrapolated before belong to the reconstructed area $\mathcal{R}$.

FSE extrapolation is carried out from a parametric model

$$g(m,n) = \sum_{(k,l)\in\mathcal{K}} c_{k,l}\varphi_{k,l}(m,n). \qquad (1)$$

This is a weighted superposition of two-dimensional basis functions $\varphi_{k,l}\,(m,n)$ with weights $c_{k,l}$. In this work we will

This work has been supported by an FPU grant from the Spanish Ministry of Education and by the MICINN TEC2010-18009 project.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

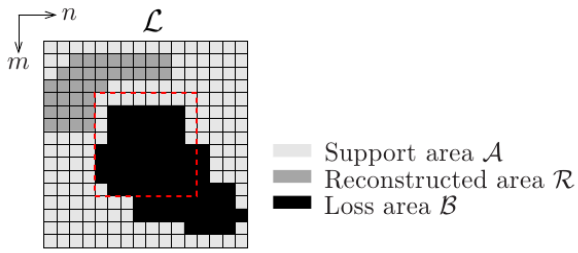


Fig. 1. Extrapolation area $\mathcal{L}$ as union of support area $\mathcal{A}$, reconstructed area $\mathcal{R}$, and loss area $\mathcal{B}$. The currently processed block (marked by the red dashed line) is located in the centre.

employ Fourier functions,

$$\varphi_{k,l}(m,n) = \frac{1}{MN} \mathrm{e}^{\frac{2\pi \mathrm{j}}{M} km} \mathrm{e}^{\frac{2\pi \mathrm{j}}{N} ln}. \tag{2}$$

As described in detail in [6], the model generation is performed iteratively, with the initial model $g^{(0)}(m,n)$ being 0, which involves that coefficients $c_{k,l}^{(0)}$ are also set to 0. At every iteration, one of the possible basis functions is selected. After estimating the corresponding weight, it is added to the model that has been generated so far. In order to determine the best basis function and its weight at every iteration ν, the residual

$$r^{(\nu)}(m,n) = \Big(s(m,n) - g^{(\nu)}(m,n)\Big) \cdot b(m,n) \tag{3}$$

between the available signal $s(m,n)$ and the current model $g^{(\nu)}(m,n)$ generated so far is regarded. Window $b(m,n)$ is zero for $(m,n) \in \mathcal{B}$ and one otherwise in order to ensure that unknown pixels are not used.

The best function $\varphi_{u,v}(m,n)$ at this iteration, conveniently weighted by a factor $\Delta c_{u,v}$, will be the one which can better approximate this residual. Let us suppose that we already know this function. Then, the corresponding model coefficient will be updated as

$$c_{u,v}^{(\nu+1)} = c_{u,v}^{(\nu)} + \gamma \Delta c_{u,v} \tag{4}$$

and the residual for the next iteration will be

$$r_{u,v}^{(\nu+1)}(m,n) = \Big(r^{(\nu)}(m,n) - \Delta c_{u,v} \varphi_{u,v}(m,n)\Big) \cdot b(m,n). \tag{5}$$

Factor γ in Eq.(4) is introduced to compensate the orthogonality deficiency of the proposed framework [7]. Coefficient $\Delta c_{u,v}$ is estimated by minimizing a weighted square error obtained from this last residual as

$$E_{u,v}^{(\nu+1)} = \sum_{(m,n)\in\mathcal{L}} w(m,n) \left| r_{u,v}^{(\nu+1)}(m,n) \right|^2. \tag{6}$$

Finally, the desired coefficient is

$$\Delta c_{u,v} = \frac{\displaystyle\sum_{(m,n)\in\mathcal{L}} r^{(\nu)}(m,n) \varphi_{u,v}^*(m,n) w(m,n)}{\displaystyle\sum_{(m,n)\in\mathcal{L}} \varphi_{u,v}^*(m,n) w(m,n) \varphi_{u,v}(m,n)} \tag{7}$$

which can be interpreted as a weighted projection coefficient of $r^{(\nu)}(m,n)$ on $\varphi_{u,v}(m,n)$, The weighting function $w(m,n)$ can be defined as [5]

$$w(m,n) = \begin{cases} \hat{\rho}^{\sqrt{\left(m-\frac{M-1}{2}\right)^2 + \left(n-\frac{N-1}{2}\right)^2}} & \forall\,(m,n) \in \mathcal{A} \\ \delta \hat{\rho}^{\sqrt{\left(m-\frac{M-1}{2}\right)^2 + \left(n-\frac{N-1}{2}\right)^2}} & \forall\,(m,n) \in \mathcal{R} \\ 0 & \forall\,(m,n) \in \mathcal{B} \end{cases}. \tag{8}$$

Using this function, the influence of each sample on the model generation can be controlled according to its position. This is also the reason why the weighting function is divided into three different parts. As all unknown samples cannot contribute to the model generation, they have to be excluded from the calculations. Accordingly, their weight in area $\mathcal{B}$ is set to 0. For the known samples, an exponentially decaying weight is used for reducing their influence with increasing distance to the area to be extrapolated in the current block. Parameter $\hat{\rho}$ controls the speed of the decay. As samples from neighbouring blocks that are originally not known but have been extrapolated before are not as reliable as originally available samples, the influence for these samples is weighted by an extra factor $\delta \in [0,1]$.

The remaining issue is the determination of the best function $\varphi_{u,v}(m,n)$. In order to do so, we must consider that, in fact, the projection coefficient and the square error can be computed for every basis function $\varphi_{k,l}(m,n)$. Furthermore, considering the orthogonality principle, every square error $E_{k,l}^{(\nu+1)}$ can be decomposed as the square error determined for the previous iteration $E^{(\nu)}$ minus the achieved decrease of square error, which is defined as [5],

$$\Delta E_{k,l}^{(\nu)} = |\Delta c_{k,l}|^2 \sum_{(m,n)\in\mathcal{L}} \varphi_{k,l}^*(m,n) w(m,n) \varphi_{k,l}(m,n). \tag{9}$$

The basis function can be selected now as the one which maximizes this decrease, that is,

$$(u,v) = \underset{(k,l)}{\operatorname{argmax}}\, \Delta E_{k,l}^{(\nu)}. \tag{10}$$

After the model generation has finished, all the samples that are originally not known are taken from the model and inserted at the corresponding positions of the incomplete original signal.

3. FSE WITH RESIDUAL FILTERING

It is well known that low frequencies are likely to yield larger Fourier coefficients than high ones in natural images [8, 9]. This is an a priori knowledge not considered in the original FSE algorithm which could be incorporated into it in order to improve both reconstruction quality and robustness. Thus, in the same way as the knowledge about spatial influence is controlled with weights $w(m,n)$, we propose here the use of

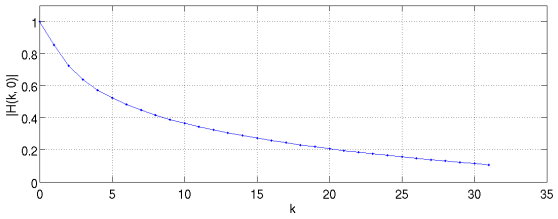


Fig. 2. One-dimensional profile of the filter H of size 64×64 with $f_0 = 0.0098$ and $G = 292.9$.

a frequency weighting (filtering) which, applied to the residuals, can exploit this a priori knowledge about frequency importance. In order to do so, it is convenient to express both residuals and square errors in the frequency domain.

3.1. FSE in the frequency domain

FSE can be efficiently implemented and easily viewed in the frequency domain [6]. Let us consider the spatially-weighted version of the residual,

$$r_w^{(\nu)}(m,n) = w(m,n) r^{(\nu)}(m,n). \tag{11}$$

Then, from Eq. (7), the projection coefficient for function $\varphi_{k,l}(m,n)$ can be expressed as

$$\Delta c_{k,l} = MN \frac{R_w^{(\nu)}(k,l)}{W(0,0)}, \tag{12}$$

where $R_w^{(\nu)}(k,l)$ and $W(k,l)$ are the DFTs of $r_w^{(\nu)}(m,n)$ and $w(m,n)$, respectively. Also, the decrease of square error can be expressed as,

$$\Delta E_{k,l}^{(\nu)} = \frac{|R_w^{(\nu)}(k,l)|^2}{W(0,0)}. \tag{13}$$

Finally, from Eq. (5), it is easily deduced that

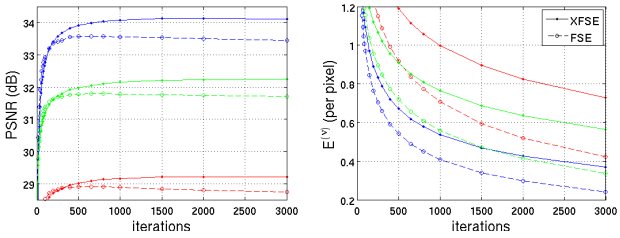
$$R_w^{(\nu+1)}(k,l) = R_w^{(\nu)}(k,l) - \frac{1}{MN} \Delta c_{u,v} W(k-u, l-v), \tag{14}$$

which provides the weighted residual required for the next iteration directly in the frequency domain. Equations (12)-(14) provide an efficient implementation of FSE, since it can be fully carried out in the frequency domain.

3.2. Filtering the weighted residual (XFSE)

We can see that the evolution of the iterative procedure relies on the computation carried out in Eqs. (12) and (13), that is, on the weighted residual $R_w^{(\nu)}(k,l)$. Therefore, a possible way of incorporating the a priori knowledge about the low-pass behaviour of natural images can be the low-pass filtering of the residual in these equations, that is,

$$\Delta c_{k,l} = MN \frac{R_w^{(\nu)}(k,l) H(k,l)}{W(0,0)}, \tag{15}$$



(a) (b)

Fig. 3. Performance overview in terms of (a) PSNR and (b) residual energy $E^{(\nu)}$ of FSE and XFSE for the images of *Peppers* (blue), *Boat* (red) and *Goldhill* (green). Dispersed error pattern is employed.

$$\Delta E_{k,l}^{(\nu)} = \frac{|R_w^{(\nu)}(k,l) H(k,l)|^2}{W(0,0)}, \tag{16}$$

where $H(k,l)$ is even, real-valued and non-negative, and represents the frequency response of the applied low-pass filter. The rest of the procedure can be kept unaltered. The resulting procedure will be referred to as XFSE in the following.

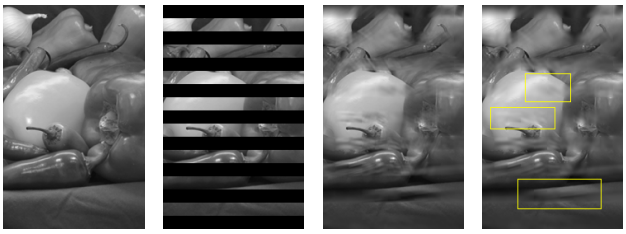
The main issue to be addressed now is the low-pass filter selection. After some preliminary experiments, we have applied a filter with the following circularly symmetric frequency response,

$$H(k,l) = \frac{\log\left[G \frac{f_0}{2\pi} \frac{1}{\left[f_0^2 + \left(\frac{k}{M}\right)^2 + \left(\frac{l}{N}\right)^2\right]^{3/2}}\right]}{\log\left(\frac{G}{2\pi f_0^2}\right)}. \tag{17}$$

This filter is inspired on the average power spectral density of natural (isotropic) images given in [8], modified with a gain factor G, smoothed by logarithm, and normalized to provide $H(0,0) = 1$. Parameter f_0 controls the bandwidth. A one-dimensional profile of this filter is shown in Fig. 2.

Let us analyze now the effect of this filter over the square error decrease. At every FSE iteration, the basis function that produces the largest decrease in the residual energy $\Delta E_{k,l}^{(\nu)}$ is selected. However, this may lead to overfitting since the reconstruction quality decreases once a critical number of iterations is achieved [7], while the weighted residual error $E^{(\nu)}$ keeps falling (see Fig.3(a)). In order to prevent this overfitting, when several basis functions yield a comparable (maximum) decrease $\Delta E_{k,l}^{(\nu)}$, the introduced filtering favours the lowest frequencies. This is illustrated in Fig. 3(b), where we can see that XFSE yields higher weighted residual error $E^{(\nu)}$ but, however, improves the reconstruction quality.

Regarding the projection coefficient $\Delta c_{u,v}$ for the selected function, since $H(k,l) \leq 1$, the filter acts as a weighting factor that reduces the contribution of high frequencies to the reconstructed signal. This does not mean that high frequencies are avoided, since if a high frequency is a clear candidate to be included in the signal model, this frequency will appear again in subsequent iterations. However, if it is



(a) (b) (c) (d)

Fig. 4. Subjective comparison for a fraction of *Peppers*. (a) Original image. (b) Received data. (c) Reconstruction by FSE. (d) Reconstruction by XFSE.

not, it will only appear spuriously, and its contribution to the final signal model will be negligible.

Since $H(u,v) \leq 1$, we can alternatively see our filtering as a dynamic reduction of the orthogonality deficiency compensation factor γ. As shown in [7], smaller compensation factors yield a better convergence (slower performance decrease after a certain number of iterations) although more iterations are required to achieve maximum performance. Although we will frequently find that $H(u,v)\gamma \ll \gamma$, during the first iterations low frequencies with high $H(k,l)$ tend to be selected, so there will be only little penalization in reconstruction quality. On the other hand, in later iterations higher frequencies are selected in order to tune fine details. For these frequencies, the effective orthogonality deficiency compensation factor $H(u,v)\gamma$ is smaller and the convergence is improved as remarked above. This is shown in the next section.

4. EXPERIMENTAL RESULTS

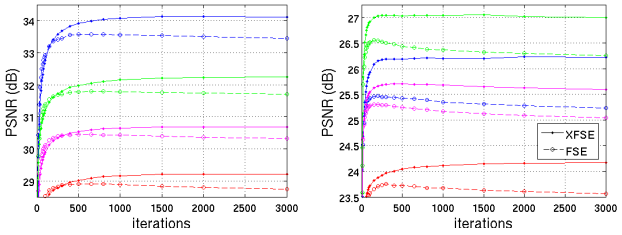
The performance of our proposal is tested on the images of *Peppers* (384×512), *Boat* (512×512) and *Goldhill* (720×576). In addition, the set of 24 images (768×512) by Kodak [10] is also used. We will employ a dispersed error pattern with a block loss rate of around 25% (see [3] for details). In addition, consecutive block losses (50% loss rate) will also be considered (see Fig. 4(b)). The blocks are considered to have dimensions of 16×16 pixels and the size of $\mathcal{L}$ is 48×48. We compare the performance with other spatial EC techniques, namely EC based on Markov random field (MRF) [1], inpainting (INP) [4], bilateral filtering (BLF) [2] and sparse linear prediction (SLP) [3].

To set up the filter, the gain factor G has been heuristically set to 292.9 in order to guarantee that the filter frequency response is always positive. On the other hand, the filter bandwidth is usually expressed as $f_0 = \alpha/2\pi$ and the value of α is around 0.06 [8] leading to $f_0 = 0.0098$ which involves a 3dB-cutoff bin of 2.17 for $N = M = 64$. The remaining FSE parameters are set according to [6], with $\gamma = 0.25$.

A comparison of XFSE and FSE is shown in Fig.4. By applying the residual filtering, the performance is improved on

		MRF	INP	BLF	SLP	FSE_{max}	$XFSE_{fse}$	$XFSE_{max}$
Peppers	(a)	32.59	33.13	33.17	33.94	33.58	33.91	**34.13**
	(b)	25.04	25.28	25.43	24.64	25.47	26.18	**26.24**
Boat	(a)	27.91	27.79	28.37	28.54	28.90	29.02	**29.22**
	(b)	23.07	22.69	22.85	22.48	23.75	23.97	**24.16**
Goldhill	(a)	31.12	30.40	30.91	31.72	31.79	32.10	**32.24**
	(b)	26.09	25.82	24.49	26.19	26.56	27.00	**27.05**
Kodak	(a)	29.61	28.76	29.64	29.92	30.45	30.54	**30.69**
	(b)	24.76	24.38	24.83	24.84	25.30	25.63	**25.71**

Table 1. PSNR values (in dB, whole images) for test images reconstructed by several algorithms. The average PSNR for the Kodak set is also included. Dispersed error pattern (a) and consecutive losses (b) are applied. The best performances are in bold face.



(a) (b)

Fig. 5. Performance comparison for (a) dispersed and (b) consecutive losses. The PSNR for *Peppers* (blue), *Boat* (red), *Goldhill* (green) and the average PSNR for the Kodak set (magenta) are shown.

average by approximately 0.4dB. This improvement is even higher when consecutive block losses are considered. Also, it is observed that the performance decrease with high a number of iterations is alleviated. Note that although XFSE achieves the maximum performance using more iterations than FSE, XFSE already outperforms FSE at the number of iterations for which FSE reaches its maximum PSNR. This behaviour is also reflected in Table 1.

Table 1 shows a PSNR comparison of the tested techniques for dispersed and consecutive losses. The best performance of FSE (FSE_{max}) is compared to the best performance of XFSE ($XFSE_{max}$) as well as to XFSE using the same number of iterations as FSE_{max} ($XFSE_{fse}$). Our proposal outperforms other state-of-the-art techniques and improves the reconstruction quality with respect to FSE by up to 0.5dB for dispersed losses and 0.7dB for consecutive losses. Finally, simulations reveal that the processing time is increased by approximately only 13%.

5. CONCLUSIONS

We have proposed the introduction of the prior knowledge about the natural image spectra into the FSE algorithm. This is achieved by filtering the residual error by a specifically designed low-pass filter. Better convergence and gains of up to 0.7dB with respect to the original FSE are achieved with a marginal additional computational cost.

6. REFERENCES

[1] S. Shirani, F. Kossentini, and R. Ward, "An adaptive Markov random field based error concealment method for video communication in error prone environment," in *Proceedings of ICIP*, 1999, vol. 6, pp. 3117–3120.

[2] G. Zhai, J. Cai, W. Lin, X. Yang, and W. Zhang, "Image error-concealment via block-based bilateral filtering," *IEEE International Conference on Multimedia and Expo*, pp. 621–624, June 2008.

[3] J. Koloda, J. Østergaard, S. H. Jensen, V. Sánchez, and A. M. Peinado, "Sequential error concealment for video/images by sparse linear prediction," *IEEE Transactions on Multimedia*, pp. 957–969, June 2013.

[4] P.F. Harrison, *Texture Synthesis, Texture Transfer and Plausible Restoration*, Ph.D. thesis, Monash University, 2005.

[5] A. Kaup, K. Meisinger, and T. Aach, "Frequency selective signal extrapolation with applications to error concealment in image communication," *AEUE - International Journal of Electronics and Communications*, vol. 59, pp. 147–156, 2005.

[6] J. Seiler and A. Kaup, "Complex-valued frequency selective extrapolation for fast image and video signal extrapolation," *IEEE Signal Processing Letters*, vol. 17, pp. 949–952, 2010.

[7] J. Seiler and A. Kaup, "Fast orthogonality deficiency compensation for improved frequency selective image extrapolation," in *Proceedings of ICASSP*, April 2008, pp. 781–784.

[8] J.B. O'Neal and T.R. Natarajan, "Coding isotropic images," *IEEE Transactions on Information Theory*, vol. 23, pp. 697–707, 1977.

[9] ITU-T, "ITU-T Recommendation H.264," International Telecommunication Union, 2010.

[10] "Kodak test images," http://r0k.us/graphics/kodak/.

2.4 Improved filling order

The papers associated to this part are:

2.4.1 Sequential Error Concealment for Video/Images by Weighted Template Matching

- J. Koloda, J. Østergaard, S.H. Jensen, A.M. Peinado and V. Sánchez, "Sequential Error Concealment for Video/Images by Weighted Template Matching", *in Proceedings of IEEE Data Compression Conference (DCC)*, pp. 159-168, Snowbird, Utah (USA), April 2012.
    - Status: Published.

2012 Data Compression Conference

Sequential Error Concealment for Video/Images by Weighted Template Matching

Ján Koloda[†], Jan Østergaard[‡], Søren H. Jensen[‡]

Antonio M. Peinado[†] and Victoria Sanchez[†]

[†]Dpt. Signal Theory, Telematics and Communications
Universidad de Granada, Spain

[‡]Multimedia Information and Signal Processing
Dpt. of Electronic Systems
Aalborg University, Denmark

Abstract

In this paper we propose a novel spatial error concealment algorithm for video and images based on convex optimization. Block-based coding schemes in packet loss environment are considered. Missing macroblocks are sequentially reconstructed by filling them with a weighted set of templates extracted from the available neighbourhood. Moreover, a fast approximation of the optimization method is proposed. The technique produces high quality reconstructions that outperforms the state-of-the-art algorithms both in terms of PSNR and MS-SSIM.

1 Introduction

Block-based video coding standards, such as MPEG-4 or H.264/AVC, are widely used in recent multimedia applications. Video signals are split into macroblocks that are coded using inter- or intraframe prediction. Quantization is carried out in the DCT domain and lossless arithmetic compression is applied [1]. This leads to low distortions at moderate bitrates. However, achieving high quality reception is a challenging task since data streams are usually transmitted over error-prone channels.

The H.264/AVC standard has introduced several error resilience tools, such as arbitrary slice order or flexible macroblock ordering. Macroblocks within a frame are split into one or more slices. A slice forms the payload of a network abstraction layer unit (NALU), which is a data sequence that can be decoded independently [1]. Video streams are packetized by NALUs so the loss of a packet would lead to the loss of, at least, one macroblock.

Error concealment (EC) techniques form a very challenging field, since QoS is of utmost importance for the users. In many cases, retransmission of lost data is not possible due to real-time constraints of the application or lack of bandwidth. In contrast to error resilience, which is carried out at the encoder, EC is applied at the decoder. EC algorithms can be classified into two categories: Spatial EC (SEC), which relies on the information provided within the current frame and Temporal EC (TEC), that utilizes temporal information such as motion vectors (MV) and previous/future frames. Both categories exploit redundancy due to high spatial and temporal correlation within a video sequence. Temporal correlation tends to be higher than the spatial correlation, so TEC techniques usually

This work has been supported by the Spanish MEC/FEDER project TEC 2010-18009.

1068-0314/12 $26.00 © 2012 IEEE
DOI 10.1109/DCC.2012.24

provide better results. This would be the straightforward choice when concealing a P/B-frame (intercoded). However, utilizing temporal information for the recovery of I-frames (intracoded) is not always possible, since these are inserted mainly to reset the prediction error when a change of scene occurs. Thus, when all the available temporal information belongs to different scene or there is no temporal information available, SEC algorithms are preferred. Every I/P-frame in the video sequence usually serves as a prediction template for, at least, one intercoded frame. Thus, high quality concealment is required since any reconstruction error will be propagated until the next I-frame arrives.

Several SEC techniques have been proposed for block-coded video/images. In [2], a simple spatial interpolation is used. In [3] a directional extrapolation algorithm was proposed, which exploits the fact that high frequencies and especially edges are visually the most relevant features. More advanced techniques including edge detectors combined with a Hough transform were utilized in [4]. Modelling natural images as Markov random fields for error concealment purposes was treated in [5]. The authors in [6] combined edge recovery and selective directional interpolation in order to achieve a more visually pleasing texture reconstruction. Patch-based texture recovery was introduced in [7]. Inpainting-based methods can also be adopted for SEC purposes [8] [9]. Sequential pixel-wise recovery based on orientation adaptive interpolation is treated in [10]. In [11], sequential Bayesian restoration is combined with DCT pyarmid decomposition. Recently, SEC techniques in transform domains have shown promising results [12].

In this paper we propose a spatial error concealment technique, where the lost regions are recovered sequentially using templates that are extracted from the available neighbourhood and combined according to a proper set of weights. First, we formulate the problem as a convex optimization problem and then we derive a fast approximation. We compare our proposals to the existing state-of-the-art algorithms on a wide selection of images and show that both in terms of PSNR and MS-SSIM our proposals provide better results.

The paper is organized as follows. In Section 2 we formulate the problem. The proposed algorithm is described in Section 3. Simulations results and comparisons with other SEC techniques are presented in Section 4. The last section is devoted to conclusions.

2 Problem Formulation

Our aim is to conceal the lost region, $\mathcal{L}$, by exploiting only the correctly received pixels in the neighbouring support area, $\mathcal{S}$.

2.1 Spatial model of an image

We define the image as a quasi-stationary signal that is locally generated by means of a stationary AR process [13]. Thus, the pixel $z(i,j)$ located at position (i,j) is generated as a linear combination of those in its neighbourhood: [1]

$$z(i,j) \;\; = \;\; \sum_{(k,l)\in\mathcal{N}} w_{(k,l)} z(k,l) + \nu(i,j), \tag{1}$$

where $\mathcal{N}$ is the stationary surrounding area of $z(i,j)$ and $\nu(i,j)$ is the residual error, which is often modelled as independent and identically distributed Gaussian noise $N(0,\sigma_N^2)$. In

[1] Note that the intracoding scheme in H.264/AVC [1] can be seen as particular case of (1).

this work, however, we do not impose any specific type of distribution upon the noise.

We will assume that the pixel values are in the range $[0; 255]$ for each colour space component. Thus, the autocorrelation is non-negative regardless of the lag, i.e. $w_{(k,l)} \geq 0$, where $w_{(k,l)}$ denotes the scalar weight associated with pixel $z(k, l)$, and that will be used for reconstruction of the pixel $z(i, j)$.

2.2 Weighted template matching (WTM)

Let $\boldsymbol{z}$ be an arbitrarily shaped group of stationary pixels. Writing $\boldsymbol{z}$ as a column vector we have $\boldsymbol{z} \in \Psi^n$, where Ψ is the $[0; 255]$ subset of integers. We will now extend the AR formulation of (1) to vectors $\boldsymbol{z}$. First, let $\mathcal{Z}$ be the set of all shifted versions of $\boldsymbol{z}$ within the stationary neighbourhood so that $\bigcup_{j=1}^{|\mathcal{Z}|} \boldsymbol{z}_j = \mathcal{N}$ and $|\boldsymbol{z}_j \bigcap \boldsymbol{z}_k| < |\boldsymbol{z}_j|$ for any $j \neq k$, i.e. no two $\boldsymbol{z}$'s completely overlap. Thus, a group of lost pixels $\boldsymbol{z}_i$ can be expressed as a linear combination of neighbouring vectors as,

$$\boldsymbol{z}_i \;=\; \sum_{j=1}^{|\mathcal{Z}|} w_j \boldsymbol{z}_j + \boldsymbol{\nu}_i. \tag{2}$$

The (unpredictable) residual error $\boldsymbol{\nu}_i$ is independent of the surrounding pixels and it is therefore not possible to exactly recover $\boldsymbol{z}_i$. In order to perform the concealment, the vector $\boldsymbol{w} = [w_1, ..., w_{|\mathcal{Z}|}]^T$ of scalar weights must first be computed in a way the residual error, $\boldsymbol{\nu}_i$, is minimized. Since the image is, in general, non-stationary, the support area $\mathcal{S}$ may include more pixels than those contained in the stationary neighbourhood $\mathcal{N}$ of $\boldsymbol{z}_i$, so the Yule-Walker equations cannot be directly applied. However, the missing block, $\boldsymbol{z}_i$, depends only on the set of $\boldsymbol{z}_j$'s belonging to the stationary neighbourhood $\mathcal{N}$. Since the location of these $\boldsymbol{z}_j$'s is unknown the entire support area needs to be searched. The set $\mathcal{Z}$ is therefore extended, so $\bigcup \boldsymbol{z}_j = \mathcal{S}$. Including more information than required for the generating AR process will, by definition, not lead to better results. Thus, given the search area $\mathcal{S}$, we are seeking the sparsest solution among all the $\boldsymbol{w}$'s that minimizes the squared prediction error. Specifically, we need to solve the following optimization problem:

$$\begin{aligned} f(\delta) = \quad & \min_{\boldsymbol{w} \in \mathcal{W}(\delta)} \|\boldsymbol{z}_i - \sum_{j=1}^{|\mathcal{Z}|} w_j \boldsymbol{z}_j\|_2^2 \\ \mathcal{W}(\delta) = \quad & \{\boldsymbol{w} |\, \|\boldsymbol{w}\|_0 \leq \delta \quad and \quad \boldsymbol{w} \succeq 0\}, \end{aligned} \tag{3}$$

where $\delta \in \mathbb{N}$ is the sparsity. The sparsity level of the sparsest solution is then given by

$$\delta^* \;=\; \min\{\operatorname*{argmin}_{\delta} f(\delta)\}. \tag{4}$$

The sparsest solution $\boldsymbol{w}^*$ is the one corresponding to δ^*.

3 Proposed Algorithms

In this section, we first propose a SEC technique based on convex optimization. Then, we derive a computationally less expensive algorithm by applying several approximations.

3.1 WTM via convex relaxation

The optimization problem defined by (3) and (4) can be rewritten as

$$\begin{aligned} &minimize \quad \|\boldsymbol{w}\|_0 \\ &subject\,to \quad \|\boldsymbol{z}_i - \sum_{j=1}^{|\mathcal{Z}|} w_j \boldsymbol{z}_j\|_2^2 = \gamma \quad and \quad \boldsymbol{w} \succeq 0, \end{aligned} \tag{5}$$

where γ is the squared ℓ_2-norm of the residual error from (2). The minimization over the ℓ_0-"norm" usually requires exhaustive search and is therefore computationally expensive. Using convex relaxation [14], (5) can be written in terms of the ℓ_1-norm

$$\begin{aligned} &minimize \quad \|\boldsymbol{w}\|_1 \\ &subject\,to \quad \|\boldsymbol{z}_i - \sum_{j=1}^{|\mathcal{Z}|} w_j \boldsymbol{z}_j\|_2^2 = \gamma \quad and \quad \boldsymbol{w} \succeq 0. \end{aligned} \tag{6}$$

The residual energy γ is usually not known in advance, so we rewrite (6) as a joint minimization problem

$$\begin{aligned} f(\delta) = &\min_{\boldsymbol{w} \in \mathcal{W}(\delta)} \|\boldsymbol{z}_i - \sum_{j=1}^{|\mathcal{Z}|} w_j \boldsymbol{z}_j\|_2^2 \\ \mathcal{W}(\delta) = &\{\boldsymbol{w} |\, \|\boldsymbol{w}\|_1 \leq \delta \quad and \quad \boldsymbol{w} \succeq 0\}, \end{aligned} \tag{7}$$

$$\delta^* = \min\{\operatorname*{argmin}_{\delta} f(\delta)\}. \tag{8}$$

Eqs. (7) and (8), however, cannot be applied directly as $\boldsymbol{z}_i$ is unknown, since it constitutes the lost region. Let us consider, without loss of generality, $\boldsymbol{z}_i$ to be the group of pixels as shown in Fig. 1(a). Note that $\boldsymbol{z}_i$ can be split into two subsets: $\boldsymbol{x}_i$, which contains only the missing pixels and $\boldsymbol{y}_i$, which is formed only by received and correctly decoded pixels and can be seen as the spatial context of $\boldsymbol{x}_i$. All $\boldsymbol{z}_j \in \mathcal{Z}$ are split in a similar way, as shown in Fig. 1(a). Since $\boldsymbol{z}_i$ is (locally) stationary and $\boldsymbol{y}_i \subset \boldsymbol{z}_i$ then the weights obtained in (7) and (8) will be identical to the weights obtained by

$$\begin{aligned} g(\delta) = &\min_{\boldsymbol{w}} \|\boldsymbol{y}_i - \sum_{j=1}^{|\mathcal{Z}|} w_j \boldsymbol{y}_j\|_2^2 \\ \mathcal{W}(\delta) = &\{\boldsymbol{w} |\, \|\boldsymbol{w}\|_1 \leq \delta \quad and \quad \boldsymbol{w} \succeq 0\}. \end{aligned} \tag{9}$$

$$\delta^* = \min\{\operatorname*{argmin}_{\delta} g(\delta)\} = \min\{\operatorname*{argmin}_{\delta} f(\delta)\}. \tag{10}$$

Finally, according to (2) the concealed group of pixels, $\hat{\boldsymbol{x}}_i$, can be approximated by a linear combination of blocks within its stationary neighbourhood

$$\hat{\boldsymbol{x}}_i = \sum_{j=1}^{|\mathcal{Z}|} w_j^* \boldsymbol{x}_j. \tag{11}$$

Computing the sparsity level δ^* of $\boldsymbol{w}^*$ (10) for every $\boldsymbol{x}_i$ is computationally expensive. Instead, we estimate the sparsity by assuming smoothness in the visual features of an image.

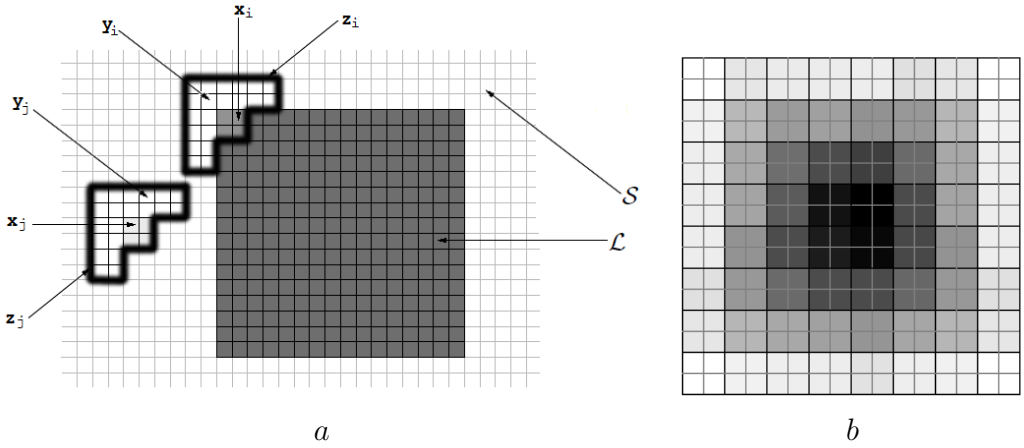


Figure 1: *(a) Example of configuration for the vectors* $\boldsymbol{x}$*,* $\boldsymbol{y}$ *and* $\boldsymbol{z}$*.* $\mathcal{S}$ *denotes the set of known pixels and* $\mathcal{L}$ *denotes the set of lost pixels. (b) Filling order for sequential reconstruction with 2×2 patches (*$p = 2$*). The regions illustrated by brighter level are recovered first.*

In the particular case of luma, it means that a reconstructed pixel could not be brighter (darker) than the brightest (darkest) pixel in $\mathcal{S}$. This requires that (11) must be a convex combination and it implies that $\delta = 1$. Finally, (9) and (10) can be reduced to

$$\begin{aligned} &\underset{\boldsymbol{w}}{minimize} \quad \|\boldsymbol{y}_i - \sum_{j=1}^{|\mathcal{Z}|} w_j \boldsymbol{y}_j\|_2^2 \\ &subject\ to \quad \|\boldsymbol{w}\|_1 \leq 1 \quad and \quad \boldsymbol{w} \succeq 0. \end{aligned} \tag{12}$$

The H.264/AVC coder packetizes the stream by slices so a loss of one packet implies a loss of, at least, one 16×16 macroblock. Applying (11) for $\boldsymbol{x}_i \in \Psi^{16\times16}$ would lead to significant imprecisions as well as blurring since it is often not possible to find a combination of $\boldsymbol{x}_j$'s that matches $\boldsymbol{x}_i$ well enough in the entire $\Psi^{16\times16}$ space. It implies that the residual error from (2) carries significant energy. In order to manage with this problem, we introduce sequential recovery. Thus, the macroblock is recovered using a set of square patches $\hat{\boldsymbol{x}}_i \in \Psi^{p\times p}$ with $1 \leq p \leq 16$. Pixel-wise reconstructions, as in [10], may introduce considerable blurring when high frequencies are involved (Fig. 4b). Using larger templates, the correlation within a template is better preserved and so is the texture (Fig. 4c). Let us consider, without loss of generality, $p = 2$ and let $\boldsymbol{y}_i$ include all the received or already recovered pixels within the 6×6 pixel neighbourhood of $\boldsymbol{x}_i$, as shown in Fig. 1(a). The macroblock is recovered sequentially by filling it with $\hat{\boldsymbol{x}}_i$ obtained by applying (12) and (11). The filling order is crucial for a high quality recovery and we base it on the reliability of $\boldsymbol{y}_i$. We define the reliability of the context $\boldsymbol{y}_i$, ρ_i, as the sum of reliabilities of its pixels. The reliability of a pixel is set to 1 if it has been correctly received and decoded. Missing pixels have zero reliability. When a pixel $x \in \boldsymbol{x}_i$ is concealed, its reliability is set to $\alpha\rho_i/m$, where $0 < \alpha < 1$ and m is the number of pixels contained in $\boldsymbol{y}_i$. We use $\alpha = 0.9$ in our simulations. The lost region $\hat{\boldsymbol{x}}_i$, whose context $\boldsymbol{y}_i$ produces the highest reliability, is recovered first. The reliability is non-increasing and the reconstruction evolves from the outer layer towards the centre of the corrupt macroblock. Fig. 1(b) shows the filling order of a 16×16 macroblock using 2×2 templates. Note that the first squares to be concealed

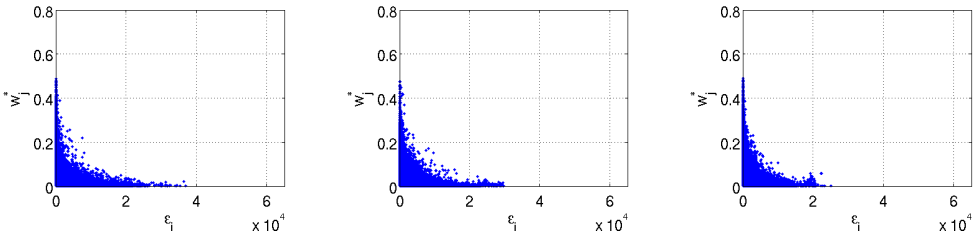
are the corners as their contexts are the largest, providing thus more reliable information which leads to a more accurate estimate of the weights.

3.2 WTM with exponentially distributed weights

Although there are efficient algorithms for solving convex optimization problems, the processing time remains high. In this section we develop a fast approximation for solving the minimization problem in (12). Specifically, we show that the weights $\boldsymbol{w}^*$ can be well modelled by an exponential distribution.

According to (12), every context $\boldsymbol{y}_j$ provides a weight w_j^*. Due to the high spatial correlation of an image, it is likely that contexts that produce smaller square error, ϵ_j, would generate larger weights, where

$$\epsilon_j \;=\; \frac{\|\boldsymbol{y}_i - \boldsymbol{y}_j\|_2^2}{m}. \tag{13}$$



a *b* *c*

Figure 2: *Distribution of weights as a function of ϵ for (a) "Peppers", (b) "Cameraman" and (c) "Barbara".*

The weights w_j^*, obtained by (12), for three different concealed images are shown in Fig. 2 as a function of ϵ_j. The error pattern applied is the one shown in Fig. 3(b). Note that the weights appear to be exponentially decaying. There is, however, a cluster of relatively low weights associated with small quadratic errors. The mild sparsity constraint in (12) implies that given two candidate templates with similar quadratic error, the optimization algorithm picks one and suppresses the other instead of using them both. Since the contribution of templates with very small weights to the final reconstruction is negligible, we will approximate the distribution of the weights by an exponential distribution, i.e.,

$$\hat{w}_j \;=\; \exp(-\frac{1}{2}\frac{\epsilon_j}{\sigma^2}), \tag{14}$$

where $\hat{w}_j$ is the approximated w_j^* and σ^2 is the variance of the distribution and can be estimated for each macroblock (or each patch). For the sake of computational simplicity a fixed value of σ^2 is used. In order to find the appropriate σ^2, we minimize the following penalty function:

$$\pi(\sigma^2) \;=\; \sum_{j=1}^{|\mathcal{Z}|}(\hat{w}_j - w_j^*)^2 w_j^* \;=\; \sum_{j=1}^{|\mathcal{Z}|}(\exp(-\frac{1}{2}\frac{\epsilon_j}{\sigma^2}) - w_j^*)^2 w_j^*. \tag{15}$$

Note that each term in the sum is scaled by w_j^* in order to reduce the influence of the small weights. Given an image, we obtain a set of estimated variances, one per patch, by minimizing the penalty function in (15). Table 1 shows the mean and the median values for the majority of the tested images. For natural images, values around 10 lead to visually good results (Fig. 5b). Larger values of σ^2 lead to oversmoothing (Fig. 5c) while small values can lead to numerical instability and should be avoided (Fig. 5a) (unless there are extremely good candidate templates).

Finally, the weights are normalized and the missing area $\boldsymbol{x}_i$ is estimated as

$$\hat{\boldsymbol{x}}_i \;=\; \frac{1}{\sum_{j=1}^{|\mathcal{Z}|} \hat{w}_j} \sum_{j=1}^{|\mathcal{Z}|} \hat{w}_j \boldsymbol{x}_j. \tag{16}$$

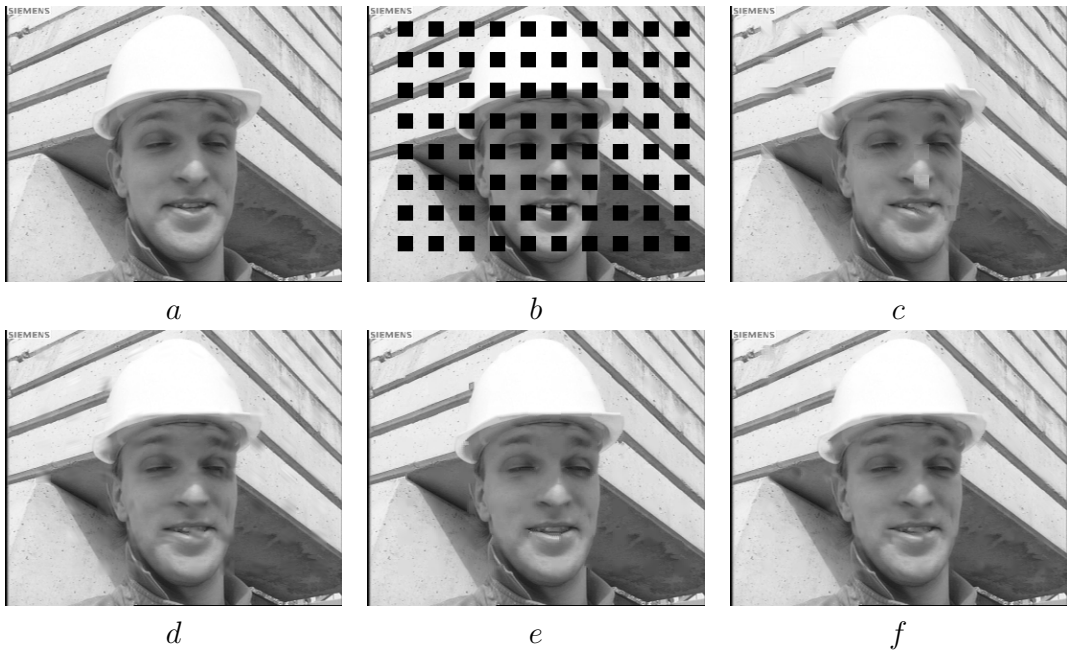


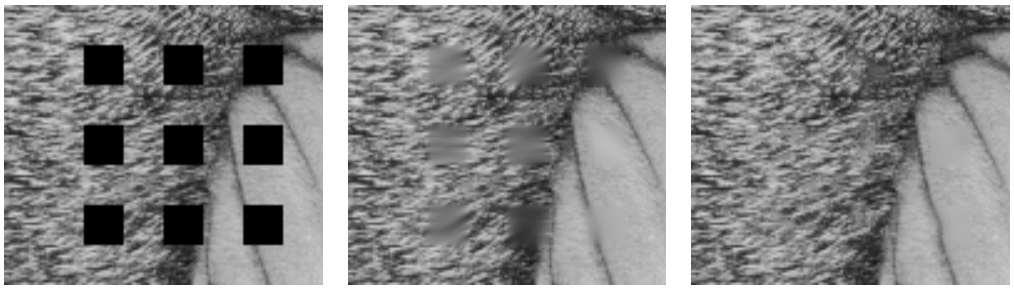
Figure 3: *SEC for the image of "Foreman" (a) Original image, (b) Received data, (c) Reconstruction using CAD (PSNR = 31.46dB, MS-SSIM = 97.54), (d) FSE (PSNR = 34.17dB, MS-SSIM = 98.03), (e) WTE (PSNR = 35.46dB, MS-SSIM = 98.73), (f) WTC (PSNR = 35.48dB, MS-SSIM = 98.68).*

σ^2	Barbara	Lena	Office	Peppers	Foreman	Cameraman	Average
mean	13.59	10.53	12.13	5.52	6.80	12.59	10.19
median	5.00	5.00	3.00	1.00	3.00	2.00	4.00

Table 1: *Estimated variance for tested images*

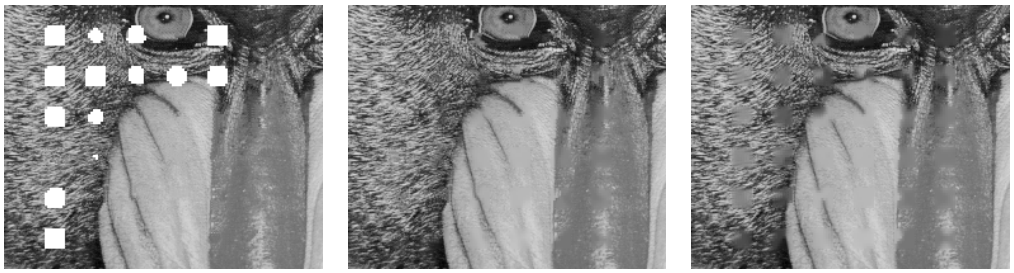
4 Simulation Results

In order to better take into account the perceptual quality, the multi scale structural similarity (MS-SSIM) index [15] is used for comparison along with the PSNR measure. In the



a b c

Figure 4: *(a) Received image, (b) reconstructed by OAI (PSNR = 27,22, MS-SSIM = 92,47) (c) reconstructed using (14) and (16) with $p = 2$ (PSNR = 25,56, MS-SSIM = 94,76).*



a b c

Figure 5: *EC for different variances σ^2: (a) σ^2 = 0.5, numerically unstable reconstructions are represented with white level, (b) σ^2 = 10, (c) σ^2 = 50.*

	BIL	EXT	SHT	CAD	AVC	MRF	INP	FSE	OAI	WTE	WTC
Average (PSNR)	26.75	27.41	27.01	28.09	27.77	29.37	28.94	30.48	30.28	**30.71**	31.18
Average (MS-SSIM)	93.93	95.03	94.29	94.84	95.14	96.31	95.42	96.69	96.28	**97.32**	96.86
Lena (PSNR)	30.00	29.39	30.47	30.44	30.42	31.89	30.88	32.72	**32.83**	32.55	32.85
Lena (MS-SSIM)	96.66	96.47	96.97	96.59	96.57	97.68	96.60	97.79	97.63	**97.94**	97.74
Peppers (PSNR)	29.47	30.24	29.68	31.26	32.13	32.76	32.16	33.48	**34.53**	33.95	34.35
Peppers (MS-SSIM)	94.90	96.86	95.26	96.84	97.26	97.86	97.29	97.99	98.39	**98.42**	98.23
Foreman (PSNR)	27.12	29.26	28.34	31.46	29.11	33.87	33.87	34.17	34.87	**35.46**	35.48
Foreman (MS-SSIM)	95.24	97.17	95.77	97.54	96.96	98.36	98.25	98.03	98.66	**98.73**	98.68
Barbara (PSNR)	26.19	26.85	25.85	26.78	26.40	27.74	28.04	**30.84**	29.37	30.79	31.91
Barbara (MS-SSIM)	95.26	94.88	95.57	95.77	95.10	96.29	95.81	97.72	96.81	**97.95**	98.19
Office (PSNR)	27.54	30.00	27.32	29.43	27.55	29.76	29.64	31.32	30.39	**31.34**	32.06
Office (MS-SSIM)	94.00	94.94	94.00	95.73	95.96	96.64	96.38	96.98	96.12	**97.43**	97.41
Baboon (PSNR)	24.15	24.72	24.14	24.92	25.42	25.17	25.06	26.02	**26.15**	26.02	26.21
Baboon (MS-SSIM)	88.84	91.35	88.69	91.80	91.92	91.94	90.90	93.27	92.69	**93.33**	93.39
Cameraman (PSNR)	25.96	25.05	26.16	25.96	26.14	26.41	25.96	**27.44**	27.00	27.24	27.26
Cameraman (MS-SSIM)	93.91	94.47	94.09	94.84	94.04	95.10	93.18	94.84	94.79	**96.92**	93.76
Tire (PSNR)	23.59	23.82	24.10	24.47	24.99	27.37	25.93	27.87	27.06	**28.33**	29.32
Tire (MS-SSIM)	92.63	94.09	93.97	89.59	93.28	96.60	94.92	96.91	95.11	**97.80**	97.46

Table 2: *PSNR values (in dB) and MS-SSIM indices (scaled by 100) for test images reconstructed by several algorithms for block dimensions 16 × 16. The best performances for each image are put in bold face (excluding the results for the more computationally expensive algorithm WTC).*

former case, the image is sequentially low-pass filtered and subsampled so a set of images is obtained, including the original resolution. Then, the SSIM index is applied for every subimage within the set. The SSIM index aims at approximating the human visual system

(HVS) response looking for similarities in luminance, contrast, and structure [15]. This index can be seen as a convolution of a fixed-sized mask with the residual error between the reference image and the concealed image [16]. A unique mask size is used for each of the images within the set so fine as well as coarse textures and objects are taken into account.

As shown in Fig. 4, the PSNR does not respond to perceptual visual quality as well as the MS-SSIM index does. In spite of that, the weights $\boldsymbol{w}^*$ are obtained according to the squared error (14) since the SSIM index tends to marginalize the influence of changes in intensity [15]. This is a desirable behaviour when measuring the perceptual image quality but not when finding candidate templates. Thus, the squared error is used when computing the weights while the MS-SSIM index is preferred for an overall quality measure. [2]

The performance of the proposed algorithm is tested on the images of "Lena" (512×512), "Barbara" (512×512), "Baboon" (512×512), Matlab built-in images "Peppers" (384×512), "Office" (592×896), "Cameraman" (256×256), "Tire" (192×224) and the first frame of "Foreman" (288×352) sequence. The test is carried out for macroblock dimensions of 16×16 and the rate of block loss is approximately 25%, corresponding to a single packet loss of a frame with dispersed slicing structure. We compare the performance with other SEC methods such as bilinear interpolation (BIL) [2], directional extrapolation (EXT) [3], a Hough transform based SEC (SHT) [4], content adaptive technique (CAD) [7], non-normative SEC for H.264 (AVC) [17], Markov random fields approach (MRF) [5], inpainting (INP) [9], frequency selective extrapolation (FSE) [12] and orientation adaptive interpolation (OAI) [10]. Both WTM via convex relaxation (WTC) and WTM with exponentially distributed weights (WTE) are tested. In the simulations, σ^2 is set to 10 and grey level images are used. Note that a pixel reconstructed by any of the aforementioned algorithms is usually real-valued and does not necessarily belong to Ψ. Thus, for comparison purposes, reconstructed pixels are rounded to the closest member of Ψ. Subjective comparison of different algorithms is shown in Fig. 3. As can be seen in Table 2, the proposed technique outperforms the others for all the tested images in terms of MS-SSIM. Moreover, the average MS-SSIM and PSNR are superior to those of state-of-the-art algorithms.

5 Conclusions

We have developed a weighted template matching algorithm, which recovers lost regions in images by filling them sequentially with a weighted combination of templates that are extracted from the available neighbourhood. The weights are obtained by solving a convex optimization problem that arises from a spatial image model. Alternatively, we show that the weights can be approximated by an exponential distribution. Our proposals achieve better PSNR and perceptual reconstruction quality than other state-of-the-art techniques. WTC is optimized for squared error so it achieves better PSNR than the approximated method. Simulations reveal, however, that WTE provides better MS-SSIM. Finally, by applying the approximated algorithm the processing time is reduced in a factor of 100.

Ongoing research is devoted to the extension of our algorithm into error concealment problems in the temporal domain.

[2]Note that the MS-SSIM index lies between [-1; 1]. In this section, we have scaled the index by 100 in order to better illustrate the differences.

References

[1] ITU-T, "ITU-T Recommendation H.264," International Telecommunication Union, 2005.

[2] P. Salama, N. Shroff, E. Coyle, and E. Delp, "Error concealment techniques for encoded video streams," in *Proceedings of ICIP*, pp. 9–12, 1995.

[3] Y. Zhao, H. Chen, X. Chi, and J. Jin, "Spatial error concealment using directional extrapolation," in *Proceedings of DICTA*, pp. 278–283, 2005.

[4] H. Gharavi and S. Gao, "Spatial interpolation algorithm for error concealment," in *Proceedings of ICASSP*, pp. 1153–1156, April 2008.

[5] S. Shirani, F. Kossentini, and R. Ward, "An adaptive Markov random field based error concealment method for video communication in error prone environment," in *Proceedings of ICIP*, vol. 6, pp. 3117–3120, 1999.

[6] W. Kung, C. Kim, and C. Kuo, "Spatial and temporal error concealment techniques for video transmission over noisy channels," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, pp. 789–802, July 2006.

[7] Z. Rongfu, Z. Yuanhua, and H. Xiaodong, "Content-adaptive spatial error concealment for video communication," *IEEE Transactions on Consumer Electronics*, vol. 50, pp. 335–341, February 2004.

[8] P. Harrison, "Texture synthesis, texture transfer and plausible restoration," *PhD. Thesis, Monash University*, 2005.

[9] A. Criminisi, P. Pérez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," *IEEE Transactions on Image Processing*, vol. 13, pp. 1200–1212, September 2004.

[10] X. Li and M. Orchard, "Novel sequential error-concealment techniques using orientation adaptive interpolation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, pp. 857–864, October 2002.

[11] G. Zhai, X. Yang, W. Lin, and W. Zhang, "Bayesian error concealment with DCT pyramid for images," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, pp. 1224–1232, September 2010.

[12] J. Seiler and A. Kaup, "Fast orthogonality deficiency compensation for improved frequency selective image extrapolation," in *Proceedings of ICASSP*, pp. 781–784, March 2008.

[13] N. Jayant and P. Noll, "Digital coding of waveforms," Prentice Hall, 1984.

[14] J. Romberg, "Imaging via compressive sensing," *IEEE Signal Processing Magazine*, vol. 25, March 2008.

[15] Z. Wang, E. Simoncelli, and A. Bovik, "Multi-scale structural similarity for image quality assessment," *IEEE Signals, Systems and Computers*, vol. 2, pp. 1398–1402, November 2003.

[16] J. Østergaard, M. Derpich, and S. Channappayya, "The high-resolution rate-distortion function under the structural similarity index," *EURASIP Journal on Advances in Signal Processing*, 2011.

[17] V. Varsa and M. Hannuksela, "Non-normative error concealment algorithms," *ITU-T SG16, VCEG-N62*, vol. 50, September 2001.

2.4.2 An Error-based Recursive Filling Ordering for Image Error Concealment

- J. Koloda, J. Seiler, A. Kaup, V. Sánchez and A.M. Peinado, "An Error-based Recursive Filling Ordering for Image Error Concealment", *in Proceedings of IEEE Internationl Conference on Image Processing (ICIP)*, Paris, France, October 2014.
    - Status: Accepted.

AN ERROR-BASED RECURSIVE FILLING ORDERING FOR IMAGE ERROR CONCEALMENT

Ján Koloda†, Jürgen Seiler‡, André Kaup‡, Victoria Sánchez† and Antonio M. Peinado†

†Dpt. Signal Theory, Networking and Communications
Universidad de Granada, Spain
{janko, victoria, amp}@ugr.es

‡Chair of Multimedia Communications and Signal Processing
University of Erlangen-Nuremberg, Germany
{seiler, kaup}@LNT.de

ABSTRACT

Many image error concealment (EC) algorithms can be employed to reconstruct a lost region by dividing it into a set of (sub)blocks which are estimated from a set of known context pixels. In turn, these former pixels may have been previously obtained by estimation. In this situation, the order in which the lost region is recursively filled will clearly condition the resulting reconstruction. This paper proposes a novel filling order aimed to improve the performance of EC algorithms that are applied recursively over the lost area. It takes into account the reconstruction quality of the already concealed blocks in order to determine the filling order. Blocks surrounded by known pixels or by high quality reconstructions are prioritized. The proposed technique is applicable to a wide range of EC algorithms and achieves an improvement of up to 1dB (in PSNR) with negligible additional computational load.

Index Terms— Error Concealment, Recursive Recovery, Filling Order

1. INTRODUCTION

Signal reconstruction is a very challenging task for many multimedia applications where the quality of the received data is of utmost importance. A common example is the transmission of image/video signals over error prone channels which may yield block losses. These lost areas need to be concealed employing the information provided by the correctly received data. Examples of efficient error concealment (EC) techniques applied in image communication can be found in [1, 2, 3, 4].

Previous work has shown that applying recursive recovery yields better reconstruction quality [2, 5]. The lost region is divided into smaller blocks that are estimated one by one following a certain order. The estimation, in turn, also relies on pixels that have been already reconstructed. In general, the filling order for image reconstruction is crucial [6] since errors can be propagated throughout the lost area. Several filling order techniques have been proposed [6]. Raster scan or concentric layer filling use a fixed filling order. The method in [5] is based on the amount of the correctly received pixels around the missing block. The more pixels there are, the higher is the block priority. Although this recursive concealment performs considerably better than a single step recovery, it does not distinguish between correctly received pixels and already extrapolated pixels. This issue is dealt with in [6] by introducing a confidence parameter. This technique, however, is based on isophotes and therefore prioritizes linear structures which may lead to considerable error propagation. In [7], the confidence term is combined with a parameter based on fractional derivative. This approach prioritizes strong edges which serves well for inpainting purposes but may not be convenient for EC tasks. In fact, it is worth noticing that the majority of the work on filling order is related to inpainting and the reconstruction error is neglected.

In this paper, we will adopt the framework described in [2], which proposes a filling order specifically designed for EC. It makes use of a pixel reliability parameter which is then employed to compute block priorities. This reliability-based sequential filling will be referred to as RSF in the following. In this work, we now propose to employ the reconstruction error to determine the filling order. Regions that yield better reconstructions will be prioritized in order to reduce error propagation and achieve better overall reconstruction quality.

The paper is organized as follows. The recursive filling framework is detailed in Section 2. The proposed technique is described in Section 3. Simulations results are discussed in Section 4. The last section is devoted to conclusions.

2. PREVIOUS WORK AND RECURSIVE FILLING FRAMEWORK

Let us consider an image region $\mathcal{R}$ as shown in Fig. 1 which comprises the region of lost pixels, $\mathcal{L}$, and the support area of available pixels, $\mathcal{S}$. Several state-of-the-art techniques conceal the lost area $\mathcal{L}$ by reconstructing the entire image region $\mathcal{R}$ and then cutting out the pixels of interest (see [2, 3, 8, 9, 10], among others). These concealment techniques are usu-

This work has been supported by an FPU grant from the Spanish Ministry of Education and by the MICINN TEC2010-18009 project.

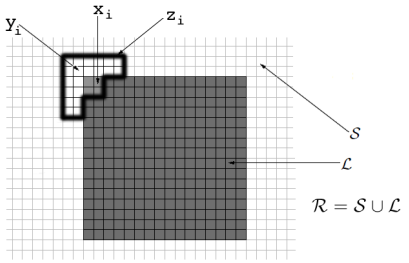


Fig. 1. Example of configuration for the vectors $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$. $\mathcal{S}$ denotes the set of known pixels and $\mathcal{L}$ denotes the set of lost pixels.

ally carried out by minimizing a reconstruction error over the support area $\mathcal{S}$ and then, exploiting the high spatial correlation within natural images, it is assumed that the reconstruction error over the lost area $\mathcal{L}$ is also minimized.

In order to achieve better reconstruction quality, the lost area $\mathcal{L}$ can be concealed recursively using smaller blocks $\mathbf{x}_i$ such that $\mathcal{L} = \bigcup_{i=1}^{N-1} \mathbf{x}_i$, where N is the number of blocks that comprise the lost area $\mathcal{L}$. We can also associate a group of pixels, called spatial context, $\mathbf{y}_i \in \mathcal{S}$ to every block $\mathbf{x}_i$. Many EC algorithms reconstruct the whole area $\mathbf{z}_i = [\mathbf{x}_i, \mathbf{y}_i]$ and then the region of interest, $\mathbf{x}_i$, is cut out. An example of such a configuration, with $\mathbf{y}_i$ comprising all the available pixels within the 6×6 neighbourhood centred in $\mathbf{x}_i$, is shown in Fig. 1. As shown in [2], this size of $\mathbf{y}_i$ is well suited for error concealment. Employing a recursive approach yields better reconstruction quality [2, 5]. It is due to the fact that the recovery of larger areas is usually less accurate since they may contain multiple objects and different textures. Targeting the concealment on smaller areas involves more homogeneous data which yields lower reconstruction errors.

One of the filling order approaches, specifically designed for EC tasks, is RSF [2]. We briefly summarize it here. Given a pixel p, its reliability ρ is set according to

$$\begin{array}{ll} \rho(p) = 1 & \text{if } p \text{ is correctly received} \\ \rho(p) = 0 & \text{if } p \text{ is lost and unknown} \\ 0 < \rho(p) < 1 & \text{if } p \text{ is already concealed} \end{array} \quad (1)$$

The filling order is determined by a priority parameter so that the block with the highest priority is recovered first. Let us denote $\mathcal{Y}_i$ the set of pixels belonging to a spatial context $\mathbf{y}_i$. The priority parameter π of a block $\mathbf{x}_i$ is computed as

$$\pi_i = \sum_{p \in \mathcal{Y}_i} \rho(p). \quad (2)$$

According to Eq.(2), the priority depends on the amount of pixels around $\mathbf{x}_i$ and on their reliabilities. Finally, the block $\mathbf{x}_u$ to be reconstructed at the current step is determined as

$$u = \underset{i}{\operatorname{argmax}}(\pi_i). \quad (3)$$

After $\mathbf{x}_u$ is estimated, the reliabilities of its pixels are set to the average reliability of the pixels within the corresponding spatial context $\mathbf{y}_u$ decreased by a constant α, i.e.,

$$\rho(q) \;=\; \frac{\pi_u}{|\mathcal{Y}_u|}\alpha = \frac{\alpha}{|\mathcal{Y}_u|} \sum_{p \in \mathcal{Y}_u} \rho(p); \quad q \in \mathcal{X}_u \quad (4)$$

where $\mathcal{X}_u$ is the set of pixels belonging to the block $\mathbf{x}_u$ and $0 \leq \alpha \leq 1$. We will employ $\alpha = 0.9$ (as in [2]). The concealed block $\mathbf{x}_u$ is then moved from the region of lost pixels $\mathcal{L}$ to the support area $\mathcal{S}$.

From Eq.(4) it is clear that the reliability is a decreasing parameter as the lost region is recursively filled. Moreover, low priority blocks produce less reliable pixels. This filling order not only distinguishes between correctly received pixels and the concealed ones but it also assigns high reliability to the pixels that have been concealed using reliable pixels.

However, none of the aforementioned filling order approaches, including RSF, takes into account the reconstruction quality. This behaviour favours error propagation. In order to solve this issue, regions that yield high quality reconstructions should be given higher priorities. This way we reduce the propagation of the reconstruction error, produced by low quality concealment, throughout the lost area. We will adopt the same framework as for RSF and we modify it so the reconstruction errors are considered.

3. THE PROPOSED METHOD

As already mentioned, many state-of-the-art techniques conceal the lost area by minimizing a reconstruction error criterion over $\mathbf{y}_i$ (e.g. sum of absolute differences, squared error, etc.). This error, normalized by the number of pixels within $\mathbf{y}_i$, may be different for every EC algorithm and we will denote it as $\varepsilon_y(\mathbf{y}_i)$. It follows that the dynamic range and the mean value of the reconstruction error may significantly vary among different EC techniques.

In this paper, we will follow the assumption of the underlying EC algorithms that minimizing $\varepsilon_y(\mathbf{y}_i)$ yields also the minimum error $\varepsilon_x(\mathbf{x}_i)$ over $\mathbf{x}_i$. In order to take into account the reconstruction quality, we modify the update of the reliability parameter, defined in Eq.(4), as follows

$$\rho(q) \;=\; \left(\frac{\pi_u}{|\mathcal{Y}_u|}\alpha\right) f(\varepsilon_y(\mathbf{y}_u)); \quad q \in \mathcal{X}_u \quad (5)$$

where $f(\varepsilon_y)$ is a penalty function that additionally decreases the reliability according to the reconstruction error with $\varepsilon_y \in \mathcal{R}^+$. Let us now analyze how this function should behave:

1. Perfect reconstructions ($\varepsilon_y = 0$) should not be penalized, i.e. $f(0) = 1$.

2. It should be monotonously decreasing. The larger the reconstruction error, the higher the penalization.

3. In order to prevent any block to be hard or even impossible to be selected, the penalization should yield non-zero reliabilities even for large reconstruction errors. In other words, the lower boundary of the function should be larger than zero.

A suitable candidate, fulfilling all the aforementioned conditions, is the following function

$$f(\varepsilon_y) = \frac{1}{1 + \exp\left(-\frac{\delta}{\varepsilon_y}\right)} \tag{6}$$

where δ is a scale parameter that controls the smoothness of the penalization. Smaller values of δ yield a sharper decrease while larger values produce a more forgiving function (see Fig. 2). As already remarked, the mean value and the dynamic range of the reconstruction error may be different for different EC algorithms. In order to make the penalization algorithm independent of the EC algorithm, a suitable selection of δ, which scales the reconstruction error, is required.

Let us denote $\bar{\varepsilon}_y$ the average value of the reconstruction error for a given EC technique. Then, we impose that reconstructions that produce lower error than the average one should not be penalized. Since the penalization function (Eq.(6)) is continuous and monotonously decreasing, it is equal to 1 (no penalization) if and only if the reconstruction error is zero. Thus, in order to relax this condition, we apply a tolerance factor τ of 0.1% ($\tau = 0.001$). This implies that reconstruction errors below $\bar{\varepsilon}_y$ will be penalized by no more than 0.1%, i.e. $f(\varepsilon_y) \geq 0.999, \forall \varepsilon_y \leq \bar{\varepsilon}_y$. Introducing the tolerance factor and rearranging Eq.(6) we obtain

$$\delta = -\bar{\varepsilon}_y \log\left(\frac{1}{1-\tau} - 1\right). \tag{7}$$

The average reconstruction error $\bar{\varepsilon}_y$ for different EC techniques are calculated using the images of *Lena* (512×512), *Soccer* (512×480), *Peppers* (512×384), *Baboon* (512×512) and *Cameraman* (256×256). Since these images are employed here for validation purposes, they will not be utilized to measure the performance later in Section 4.

Note that according to Eqs.(6) and (7) the penalization depends on the ratio between the observed error and the average. Thus, for instance, reconstructions that yield errors twice as large as the average are penalized by the same amount, regardless of the EC algorithm. It follows that the proposed recursive filling technique adapts itself to the algorithm involved.

4. SIMULATIONS AND RESULTS

The performance of our proposal is tested on the images of *Foreman* (352×288), *Boat* (512×512), *Goldhill* (720×576)

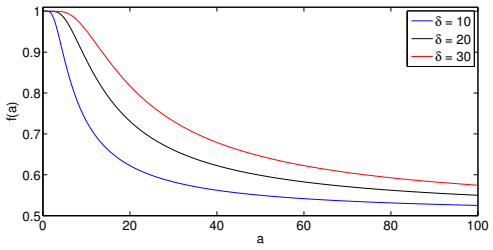


Fig. 2. Penalty function $f(a)$ for different values of δ.

Barbara (512×512) and *Tire* (232×205). In addition, the set of 24 images (768×512) by Kodak [11] is also used. We have tested different filling order (FO) techniques: concentric layer FO (CLF), gradient guided FO (GGF) [7], parallel FO (PFO) [5], RSF [2] and the proposed error-based recursive filling ordering (ε-RF). We have applied these techniques to different EC algorithms, namely overlapping boundary matching algorithm (OBMA) [8] (with the corresponding reconstruction error described in Eq.(3) of the reference), multi-hypothesis EC (MHEC) [9] (Eq.(1) in [9]), sparse linear prediction (SLP) [2] (Eq.(10) in [2]) and frequency selective extrapolation (FSE) [3] (Eq.(8) in [3]). We have also included an oracle trial (ORA) using the optimal value of δ which maximizes the PSNR for every image (found by exhaustive search). Dispersed error pattern and consecutive losses have been applied, as shown in Fig. 3(b) and Fig. 4(b), respectively.

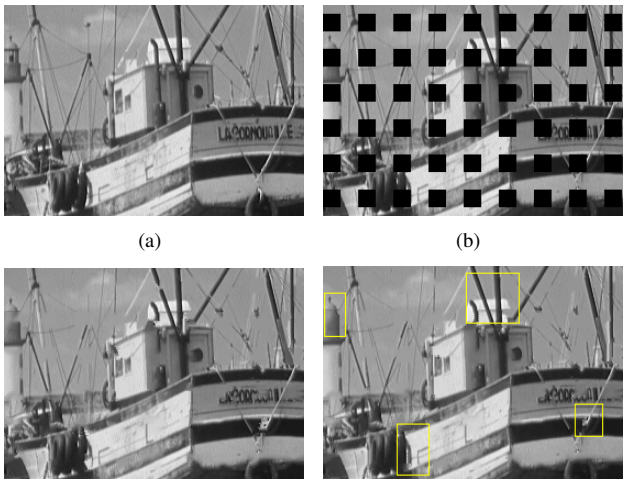
Table 1 shows the results in terms of PSNR (in dB) for the dispersed error pattern. In Table 2, we apply consecutive losses and we compare our proposal with RSF, which performs the best among the previous FO techniques tested. By applying our proposal, the best reconstruction quality is achieved. In fact, the proposed technique outperforms RSF by more than 1dB in some cases and with negligible additional computational load, as shown later. Moreover, the proposed value for δ (Eq.(7)) differs on average only marginally from the performance achieved by the optimal value of δ.

Figures 3 and 4 show the subjective comparison for the images of *Boat* and *Foreman* using SLP and MHEC, respectively. The reduction of error propagation, reflected on PSNR, yields as well reconstructions with fewer and less noticeable artifacts.

Regarding the computational complexity, the additional burden, with respect to RSF, is due to the evaluation of the penalty term (Eq.(6)) and the update of the pixels reliabilities. Simulations reveal that the computational cost is increased only from 0.5% to 1.5% of the total burden, depending on the considered EC algorithm.

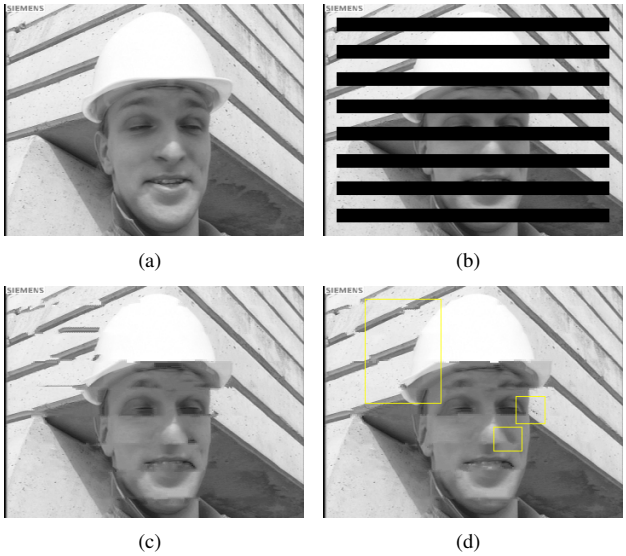
5. CONCLUSIONS

In this paper, we have proposed a recursive ordering approach that takes into account the reconstruction quality. We have



(a) (b) (c) (d)

Fig. 3. Subjective comparison for a fraction of *Boat* (reconstructed by SLP). (a) Original image. (b) Received data. (c) Reconstructed using RSF. (d) Reconstructed using ε-RF. The most outstanding differences are marked with yellow boxes.



(a) (b) (c) (d)

Fig. 4. Subjective comparison for the image of *Foreman* (reconstructed by MHEC). (a) Original image. (b) Received data. (c) Reconstructed using RSF. (d) Reconstructed using ε-RF. The most outstanding differences are marked with yellow boxes.

	CLF	GGF	PFO	RSF	ε-RF	ORA	
OBMA $\bar{\varepsilon}_y = 6.95$	33.00	33.22	34.74	34.85	**35.14**	35.48	*Foreman*
	27.51	27.19	27.91	28.13	**28.69**	28.85	*Boat*
	29.96	29.76	30.61	30.60	**30.85**	30.96	*Goldhill*
	30.01	29.10	30.17	30.32	**30.66**	30.79	*Barbara*
	27.13	27.04	27.44	27.21	**28.48**	28.72	*Tire*
	29.52	29.26	30.17	30.22	**30.83**	30.96	Average
	28.87	28.52	29.07	29.18	**29.35**	29.37	Kodak
MHEC $\bar{\varepsilon}_y = 5.62$	34.02	33.69	35.08	**35.63**	**35.63**	36.40	*Foreman*
	28.01	27.73	28.51	28.67	**29.27**	29.33	*Boat*
	31.01	30.54	31.29	31.26	**31.60**	31.67	*Goldhill*
	30.66	29.23	30.83	30.93	**31.14**	31.36	*Barbara*
	27.51	26.56	27.54	27.90	**28.33**	28.64	*Tire*
	30.24	29.55	30.65	30.88	**31.28**	31.48	Average
	29.68	29.23	29.86	29.93	**30.09**	30.11	Kodak
SLP $\bar{\varepsilon}_y = 4.65$	34.27	33.17	35.05	35.38	**35.75**	36.11	*Foreman*
	28.30	27.83	28.65	28.89	**29.49**	29.55	*Boat*
	31.45	30.89	31.69	31.88	**31.97**	31.98	*Goldhill*
	30.62	29.65	30.85	30.96	**31.22**	31.29	*Barbara*
	27.81	27.91	28.26	28.45	**28.87**	28.87	*Tire*
	30.49	29.89	30.90	31.11	**31.42**	31.56	Average
	29.89	29.43	30.02	30.08	**30.24**	30.24	Kodak
FSE $\bar{\varepsilon}_y = 0.59$	33.08	32.15	33.05	33.15	**33.17**	33.42	*Foreman*
	28.68	28.45	28.71	28.73	**28.83**	28.89	*Boat*
	31.47	31.15	31.36	31.38	**31.50**	31.50	*Goldhill*
	30.85	30.53	30.88	31.01	**31.09**	31.09	*Barbara*
	28.06	28.64	28.69	28.70	**28.91**	28.93	*Tire*
	30.43	30.18	30.53	30.59	**30.76**	30.77	Average
	29.98	29.78	30.00	30.05	**30.15**	30.17	Kodak

Table 1. PSNR values (in dB, whole images) for test images reconstructed by several algorithms using dispersed error pattern (the average error $\bar{\varepsilon}_y$ is also indicated). The average PSNR for the images listed in the table as well as the average for the Kodak set are also included. Different filling orders are applied: concentric layers FO (CLF), gradient guided FO (GGF), parallel FO (PFO), reliability-based FO (RSF), the proposed error-based FO (ε-RF) and ε-RF using the optimal value of δ (ORA). The best performances (excluding the oracle trial) are in bold face.

	Foreman	*Boat*	*Goldhill*	*Barbara*	*Tire*	Kodak	
OBMA	26.72	22.14	25.11	24.56	20.66	23.93	RSF
	27.94	**22.54**	**25.43**	**25.14**	**21.53**	**24.08**	ε-RF
MHEC	26.54	22.57	25.56	24.91	20.61	24.38	RSF
	28.25	**22.84**	**25.90**	**25.55**	**21.67**	**24.57**	ε-RF
SLP	27.68	**22.90**	26.20	25.01	20.52	24.66	RSF
	28.08	22.88	**26.41**	**25.80**	**21.39**	**24.86**	ε-RF
FSE	26.20	23.24	25.80	25.15	21.94	**24.50**	RSF
	26.61	**23.39**	**25.84**	**25.61**	**22.48**	**24.50**	ε-RF

Table 2. PSNR values for test images reconstructed by several algorithms using consecutive error pattern. The average PSNR for the Kodak set is also included. The filling orders RSF and the proposed ε-RF are employed. The best performances are in bold face.

introduced a penalty function that maps the reconstruction error into a penalization factor that controls pixel reliabilities. Blocks that are surrounded by high quality reconstructions, i.e., by more reliable pixels, are prioritized. The proposed technique is applicable to any EC algorithm that recovers the lost pixels by reconstructing a wider area and then cutting out the region of interest. Improvements of up to 1dB (in PSNR) are achieved with respect to RSF with only marginal increase in computational complexity. Ongoing work is focused on a more accurate filling order that exploits local data statistics.

6. REFERENCES

[1] G. Zhai, X. Yang, W. Lin, W. Zhang, and Y. Xu, "Bayesian error concealment with DCT pyramid," in *Proceedings of ICASSP*, 2010, pp. 1366–1369.

[2] J. Koloda, J. Østergaard, S. H. Jensen, V. Sánchez, and A. M. Peinado, "Sequential error concealment for video/images by sparse linear prediction," *IEEE Transactions on Multimedia*, pp. 957–969, June 2013.

[3] J. Seiler and A. Kaup, "Complex-valued frequency selective extrapolation for fast image and video signal extrapolation," *IEEE Signal Processing Letters*, vol. 17, pp. 949–952, 2010.

[4] G. Zhai, J. Cai, W. Lin, X. Yang, and W. Zhang, "Image error-concealment via block-based bilateral filtering," *IEEE International Conference on Multimedia and Expo*, pp. 621–624, June 2008.

[5] J. Seiler and A. Kaup, "Optimized and parallelized processing order for improved frequency selective extrapolation," in *Proceedings of EUSIPCO*, September 2011, pp. 269–273.

[6] A. Criminisi, P. Pérez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," *IEEE Transactions on Image Processing*, vol. 13, pp. 1200–1212, September 2004.

[7] A.S. Hareesh and V. Chandrasekaran, "A fast and simple gradient function guided filling order prioritization for exemplar-based color image inpainting," in *Proceedings of ICIP*, September 2010, pp. 409–412.

[8] T. Thaipanich, P.H. Wu, and C.C. Jay Kuo, "Video error concealment with outer and inner boundary matching algorithms," in *Proceedings of SPIE*, 2007.

[9] K. Song, T. Chung, C.-S. Kim, Y.-O Park, Y. Kim, Y. Joo, and Y. Oh, "Efficient multi-hypothesis error concealment technique for H.264," in *Proceedings of ISCAS*, May 2007, pp. 973–976.

[10] J. Koloda, A.M. Peinado, and V. Sanchez, "On the application of multivariate kernel density estimation to image error concealment," in *Proceedings of ICASSP*, May 2013, pp. 1330–1334.

[11] "Kodak test images," http://r0k.us/graphics/kodak/.

Chapter 3

Conclusions and Future Work

3.1 Conclusions

This thesis has been focused on designing reconstruction algorithms that outperform other state-of-the-art EC techniques for image/video communication. In order to do so, we have adopted different approaches. In this section, we briefly summarize the four research topics tackled in this thesis.

- First, we have have dealt with interpolation based EC. The main problem here is the lack of robustness and the inability to recover complex edges. We have tackled these issues by introducing a novel scanning procedure which, along with the Hough transform, provides a robust and accurate descriptor of the relevant edges around the missing area. Moreover, since this area can be affected by multiple edges, several directional interpolations are employed in order to obtain the final reconstruction. These interpolations are combined using a set of weights unique for each missing pixel. We have introduced a visual clearness parameter associated to every edge which is used to estimate these weights. This proposal successfully recovers complex edges and significantly outperforms other state-of-the-art interpolation based EC techniques.

- The second main area, addressed in this thesis, is to take advantage of local data statistics in order to recover the missing samples. We have proposed a vector linear prediction scheme which, under a sparsity constraint, dynamically adapts itself to the amount of available and useful data. This scheme has been tested for image, video and also speech transmission applications. The method can be further simplified by adopting an exponential approximation. By doing so, comparable or better performance is obtained with much less computational burden. This approach can be generalized by adopting a multivariate kernel-based MMSE framework. The key issue consists in a novel kernel bandwidth estimation, especially oriented to multimedia signal reconstruction. These proposals provide high quality reconstructions of complex structures and fine textures, both on objective and subjective levels.

- Third, we have studied EC in transformed domain. Natural images are low-pass signals and we have proposed to use this information in order to improve the reconstruction quality. This is achieved by introducing a residual frequency filtering into the existing frequency selective extrapolation algorithm. For every iteration, this filtering affects the selection of the best fitting basis function and the update of the corresponding Fourier coefficient. This approach

almost completely suppresses the overfitting, present in the original FSE algorithm, and provides an improvement of up to 1dB (in terms of PSNR) with respect to FSE.

- Finally, we have proposed a novel filling ordering, especially conceived for EC tasks. Regions with lower reconstruction errors are prioritized in order to achieve better overall reconstruction quality. Reconstruction error is estimated from the available samples and a penalty function is introduced in order to reduce error propagation. Simulations reveal that improvements of up to 1dB (in terms of PSNR) are achieved with marginal additional computational cost. This approach is also applicable to a large variety of EC techniques.

Moreover, in order to encourage reproducible research, we have made the implementations of our techniques and various EC algorithms available online at [59]. By publishing the source codes we hope to contribute to the advances in image reconstruction since obtaining a thorough benchmark will be easier.

Conclusiones

Esta tesis se ha centrado en el diseño de técnicas de reconstrucción cuyo rendimiento supere a los algoritmos EC del estado de arte. Durante la fase de diseño se han adoptado diferentes aproximaciones. En esta sección se ofrece un breve resumen de las cuatro aproximaciones tratadas en esta tesis:

- En primer lugar hemos trabajado con técnicas EC basadas en interpolación. Su mayor problema es la falta de robustez y la incapacidad de reconstruir bordes más comlejos. Hemos tratado de resolver este problema introduciendo un procedimiento novedoso de barrido que, aplicando la transformada de Hough, proporciona un descriptor robusto y preciso de las fronteras relevantes alrededor de la región perdida. Además, como esta región puede verse afectada por múltiples fronteras, empleamos varias interpolaciones con el fin de obtener la reconstrucción final. Estas interpolaciones se combinan utilizando un conjunto de pesos que son únicos para cada pixel. Hemos introducido el parámetro de claridad visual que se asocia a cada frontera y que se utiliza para estimar los pesos. Esta técnica es capaz de recuperar fronteras complejas y proporciona un rendimiento considerablemente superior al resto de técnicas EC del estado de arte basadas en interpolación.

- La segunda área, considerada en esta tesis, es reconstruir las muestras perdidas aprovechando la estadística local. Hemos propuesto un esquema de predicción lineal vectorial que, bajo las condiciones de *sparsity*, se adapta dinámicamente a la cantidad de muestras disponibles. El esquema se ha aplicado a la transmisión de imágenes, vídeo y también voz. Este método se puede simplificar aplicando una aproximación exponencial. Así se consigue un rendimiento similar o incluso superior al esquema original pero con una carga computacional significativamente menor. Este esquema se puede generalizar adoptando un marco MMSE multivariado basado en kernels. El aspecto clave consiste en una novedosa estimación del ancho de banda, orientada especialmente a la reconstrucción de señales multimedia. Las técnicas propuestas ofrecen altas calidades de reconstrucción de estructuras complejas y texturas finas, tanto en el plano objetivo como sujetivo.

- En tercer lugar hemos estudiado los algoritmos EC en el dominio transformado. Las imágenes naturales suelen ser señales paso-baja así que hemos propuesto utilizar esta información

con el fin de mejorar la calidad de reconstrucción. Hemos modificado el algoritmo FSE (*frequency selective extrapolation*) introduciendo un filtrado de frecuencias residuales. En cada iteración, este filtrado afecta el proceso de selección de la mejor función base y la actualización del correspondiente coeficiente de Fourier. Esta aproximación elimina casi por completo el *overfitting*, presente en el algoritmo FSE original, y produce una mejora de hasta 1dB (en términos de PSNR) con respecto a FSE.

- Por último, hemos propuesto un algoritmo de orden de relleno especialmente diseñado para EC. Se tiende a priorizar regiones con menor error de reconstrucción para conseguir una calidad de reconstrucción global más alta. El error de reconstrucción se estima a partir de las muestras disponibles y se introduce una función de penalización para reducir la propagación de errores. Experimentalmente se ha conseguido una mejora de hasta 1dB (en términos de PSNR) con un coste computacional adicional mínimo. Cabe mencionar que esta propuesta es aplicable a una amplia gama de algoritmos EC.

Además, para propiciar la investigación reproducible, hemos publicado las implementaciones de nuestras técnicas y de otros algoritmos del estado de arte. Los códigos fuente se pueden descargar en [59]. Esperamos que la publicación de los códigos contribuya a los avances en la investigación de reconstrucción de imágenes dado que obtener una comparación exhaustiva con el resto de técnicas será mucho más asequible.

3.2 Future work

Having developed the techniques presented in this thesis, various research topics have arisen to be further explored.

Kernel-based MMSE can be employed to improve the filling order. It can provide an estimation of the reconstruction error that will determine the filling priorities.

A thorough study on the low-pass filter, used in our improved FSE algorithm, is desirable. It could provide further improvements and this approach could be later generalized to any signal source. It is also worth exploring an adaptive filtering that would control the evolution of the FSE algorithm according to the spectral features of the residual energy.

EC techniques can be extended to other related applications, such as inpainting, object removal, superresolution or texture diffusion. The objective of such applications is a visually plausible result and not the minimization of an error criterion between the original and reconstructed signal. In fact, there is no well-defined unique solution [60]. It is worth exploring the possibility of modifying the proposed algorithms so they can be employed in the aforementioned applications.

It has been shown that the techniques proposed in this thesis can successfully recover corrupt images and video streams, leading to high quality reconstructions. That means that the received streams still contain a significant amount of redundancy. This issue can be further studied in order to propose **new compression algorithms** for multimedia signals.

Bibliography

[1] Ericsson Mobility Report:, "On the pulse of the networked society," *http://www.ericsson.com/ericsson-mobility-report*, June 2013.

[2] T. Tröger and A. Kaup, "Inter-sequence error concealment techniques for multi-broadcast TV reception," *IEEE Transactions on Broadcasting*, vol. 57, pp. 777–793, December 2011.

[3] T. Wiegand, G.J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, pp. 560–576, July 2003.

[4] G.J. Sullivan, J. Ohm, W.J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, pp. 1649–1668, December 2012.

[5] ITU-T, "ITU-T Recommendation H.264," International Telecommunication Union, 2010.

[6] I.E.G. Richardson, "The H.264 advanced video compression standard," Wiley, 2010.

[7] Z. Wang, E.P. Simoncelli, and A.C. Bovik, "Multi-scale structural similarity for image quality assessment," *IEEE Signals, Systems and Computers*, vol. 2, pp. 1398–1402, November 2003.

[8] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assesment: From error visibility to structural visibility," *IEEE Transactions on Image Processing*, vol. 13, pp. 600–612, April 2004.

[9] S. Chikkerur, V Sundaram, M. Reisslein, and L.J. Karam, "Objective video quality assessment methods: A classification, review, and performance comparison," *IEEE Transactions on Broadcasting*, vol. 57, pp. 165–182, June 2011.

[10] P. Salama, N.B. Shroff, E.J. Coyle, and E.J. Delp, "Error concealment techniques for encoded video streams," in *Proceedings of ICIP*, 1995, pp. 9–12.

[11] V. Varsa and M.M. Hannuksela, "Non-normative error concealment algorithms," *ITU-T SG16, VCEG-N62*, vol. 50, September 2001.

[12] E. Ong, W. Lin, Z. Lu, S. Yao, and M. Etoh, "Visual distortion assessment with emphasis on spatially transitional regions," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 559–566, April 2004.

[13] W.Y. Kung, C.S. Kim, and C.C.J. Kuo, "Spatial and temporal error concealment techniques for video transmission over noisy channels," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, pp. 789–802, July 2006.

[14] Z. Rongfu, Z. Yuanhua, and H. Xiaodong, "Content-adaptive spatial error concealment for video communication," *IEEE Transactions on Consumer Electronics*, vol. 50, pp. 335–341, February 2004.

[15] D. Agrafiotis, D.R. Bull, and C.N. Canagarajah, "Enhanced error concealment with mode selection," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 960–973, August 2006.

[16] M. Chen, Y. Zheng, and M. Wu, "Classification-based spatial error concealment for visual communications," *EURASIP Journal on Advances in Signal Processing*, pp. 1–17, March 2006.

[17] W. Kim, J. Koo, and J. Jeong, "Fine directional interpolation for spatial error concealment," *IEEE Transactions on Consumer Electronics*, pp. 1050–1056, August 2006.

[18] H. Asheri, H.R. Rabiee, N. Pourdamghani, and M. Ghanbari, "Multi-directional spatial error cocnealment using adaptive edge thresholding," *IEEE Transactions on Consumer Electronics*, vol. 58, pp. 880–885, 2012.

[19] D.L. Robie and R.M. Mersereau, "The use of Hough transforms in spatial error concealment," in *Proceedings of ICASSP*, 2000, vol. 4, pp. 2131–2134.

[20] H. Gharavi and S. Gao, "Spatial interpolation algorithm for error concealment," in *Proceedings of ICASSP*, April 2008, pp. 1153–1156.

[21] X. Li and M.T. Orchard, "Novel sequential error-concealment techniques using orientation adaptive interpolation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, pp. 857–864, October 2002.

[22] Y. Zhao, H. Chen, X. Chi, and J.S. Jin, "Spatial error concealment using directional extrapolation," in *Proceedings of DICTA*, 2005, pp. 278–283.

[23] S. Shirani, F. Kossentini, and R. Ward, "An adaptive Markov random field based error concealment method for video communication in error prone environment," in *Proceedings of ICIP*, 1999, vol. 6, pp. 3117–3120.

[24] Y. Zhang, X. Xiang, D. Zhao, S. Ma, and W. Gao, "Packet video error concealment with auto regressive model," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 12–27, January 2012.

[25] G. Zhai, J. Cai, W. Lin, X. Yang, and W. Zhang, "Image error-concealment via block-based bilateral filtering," *IEEE International Conference on Multimedia and Expo*, pp. 621–624, June 2008.

[26] Z. Wang, Y. Yu, and D. Zhang, "Best neighborhood matching: An information loss restoration technique for block-based image coding systems," *IEEE Transactions on Image Processing*, vol. 7, pp. 1056–1061, July 1998.

[27] K. Song, T. Chung, C.-S. Kim, Y.-O Park, Y. Kim, Y. Joo, and Y. Oh, "Efficient multi-hypothesis error concealment technique for H.264," in *Proceedings of ISCAS*, May 2007, pp. 973–976.

[28] D. Nguyen, M. Dao, and T. Tran, "Video error concealment using sparse recovery and local dictionaries," in *Proceedings of ICASSP*, May 2011, pp. 1125–1128.

[29] A.M. Peinado, V. Sánchez, and A. Gómez, "Error concealment based on MMSE estimation for multimedia wireless and IP applications," in *Proceedings of PIMRC (invited paper)*, September 2008, pp. 1–5.

[30] C.S. Kim, J.W. Kim, I. Katsavounidis, and C.C.J. Kuo, "Robust MMSE video decoding: Theory and practical implementations," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, pp. 39–51, January 2005.

[31] D. Persson, T. Eriksson, and P. Hedelin, "Packet video error concealment with gaussian mixture models," *IEEE Transactions on Image Processing*, vol. 17, pp. 145–154, 2008.

[32] D. Persson and T. Eriksson, "Mixture model- and least squares-based packet video error concealment," *IEEE Transactions on Image Processing*, vol. 18, pp. 1048–1054, 2009.

[33] H. Sun and W. Kwok, "Concealment of damaged block transform coded images using projections onto convex sets," *IEEE Transactions on Image Processing*, vol. 4, no. 4, April 1995.

[34] J.W. Park, J.W. Kim, and S.U. Lee, "DCT coefficients recovery-based error concealment technique and its application to the MPEG-2 bit stream error," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 7, pp. 845–854, December 1997.

[35] Z. Alkachouh and M.G. Bellanger, "Fast DCT-based spatial domain interpolation of blocks in images," *IEEE Transactions on Image Processing*, pp. 729–732, April 2000.

[36] G. Zhai, X. Yang, W. Lin, and W. Zhang, "Bayesian error concealment with DCT pyramid for images," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, pp. 1224–1232, September 2010.

[37] A. Kaup, K. Meisinger, and T. Aach, "Frequency selective signal extrapolation with applications to error concealment in image communication," *AEUE - International Journal of Electronics and Communications*, vol. 59, pp. 147–156, 2005.

[38] J. Seiler and A. Kaup, "Complex-valued frequency selective extrapolation for fast image and video signal extrapolation," *IEEE Signal Processing Letters*, vol. 17, pp. 949–952, November 2010.

[39] J. Seiler, K. Mesinger, and A Kaup, "Orthogonality deficiency compensation for improved frequency selective image extrapolation," in *Proceedings of PCS*, November 2007.

[40] J. Seiler and A Kaup, "Fast orthogonality deficiency compensation for improved frequency selective image extrapolation," in *Proceedings of ICASSP*, March 2008, pp. 781–784.

[41] J. Seiler and A. Kaup, "Optimized and parallelized processing order for improved frequency selective extrapolation," in *Proceedings of EUSIPCO*, September 2011, pp. 269–273.

[42] A. Criminisi, P. Pérez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," *IEEE Transactions on Image Processing*, vol. 13, pp. 1200–1212, September 2004.

[43] A.S. Hareesh and V. Chandrasekaran, "A fast and simple gradient function guided filling order prioritization for exemplar-based color image inpainting," in *Proceedings of ICIP*, September 2010, pp. 409–412.

[44] X. Qian, G. Liu, and H. Wang, "Recovering connected error region based on adaptive error concealment order determination," *IEEE Transactions on Multimedia*, pp. 683–695, June 2009.

[45] W.Y. Kung, C.S. Kim, and C.C.J. Kuo, "A spatial-domain error concealment method with edge recovery and selective directional interpolation," in *Proceedings of ICASSP*, April 2003, pp. 700–703.

[46] S. Shirani, F. Kossentini, and R. Ward, "A concealment method for video communications in an error-prone environment," *IEEE Journal on Selected Areas in Communications*, vol. 18, pp. 1122–1128, June 2000.

[47] S. Roth and M.J. Black, "Fields of experts," *International Journal of Computer Vision*, vol. 82, pp. 205–229, April 2009.

[48] K. Guo, X. Yang, R. Zhang, S. Yu, and H. Zha, "Interpolating fine textures with fields of experts prior," in *Proceedings of ICIP*, November 2009, pp. 353–356.

[49] Y. Zhang, X. Xiang, S. Ma, D. Zhao, and W. Gao, "Auto regressive model and weighted least squares based packet video error concealment," in *Proceedings of DCC*, March 2010, pp. 455–464.

[50] X. Xiang, Y. Zhang, D. Zhao, M. Sa, and W. Gao, "A high efficient error concealment scheme based on auto-regressive model for video coding," in *Proceedings of PCS*, May 2009, pp. 1–4.

[51] G. Yu, G. Sapiro, and S. Mallat, "Solving inverse problems with piecewise linear estimators: From Gaussian mixture models to structured sparsity," *IEEE Transactions on Image Processing*, vol. 21, pp. 2481–2499, May 2012.

[52] A. M. Gomez, A.M. Peinado, V. Sanchez, and A.J. Rubio, "Combining media specific FEC and error concealment for robust distributed speech recognition over loss-prone packet channels," *IEEE Transactions on Multimedia*, vol. 8, pp. 1228–1238, November 2006.

[53] J. Romberg, "Imaging via compressive sensing," *IEEE Signal Processing Magazine*, vol. 25, no. 2, March 2008.

[54] L. Vandenberghe and S. Boyd, "Semidefinite programming," *Society for Industrial and Applied Mathematics*, p. 49–95, 1996.

[55] S.V. Vaseghi, "Advanced digital signal processing and noise reduction," Wiley, 2006.

[56] D.W. Scott, "Multivariate density estimation: Theory, practice, and visualization," Wiley, 1992.

[57] M. Kristan, A. Leonardis, and D. Skočaj, "Multivariate online kernel density estimation with gaussian kernels," *Pattern Recognition*, vol. 44, pp. 2630–2642, 2011.

[58] J. O'Neal and T. Natarajan, "Coding isotropic images," *IEEE Transactions on Information Theory*, pp. 697–707, November 1977.

[59] [Online], "Available: `http://dtstc.ugr.es/~jkoloda/download.html`," .

[60] C. Guillemot and O. Le Meur, "Image inpainting : Overview and recent advances," *IEEE Signal Processing Magazine*, vol. 31, pp. 127–144, January 2014.