MARÍA ÁNGELES MENDOZA PÉREZ

RECONOCIMIENTO DE ACCIONES HUMANAS BASADO EN MODELOS PROBABILÍSTICOS DE ESPACIO DE ESTADOS

RECONOCIMIENTO DE ACCIONES HUMANAS BASADO EN MODELOS PROBABILÍSTICOS DE ESPACIO DE ESTADOS

MARÍA ÁNGELES MENDOZA PÉREZ

Universidad de Granada

Ciencias de la Computación e Inteligencia Artificial

2009

.

Editor: Editorial de la Universidad de Granada
Autor: Mª Ángeles Mendoza Pérez
D.L.: GR 2385-2010
ISBN: 978-84-693-1320-6

DECLARACIÓN

D. Nicolás Pérez de la Blanca Capilla
Profesor Catedrático del Departamento de Ciencias de la Computación de la Universidad de Granada

CERTIFICA:

Que la presente memoria, titulada *Reconocimiento de Acciones Humanas Basado en Modelos Probabilísticos de Espacio de Estados*, ha sido realizada por Dña. María Ángeles Mendoza Pérez bajo mi dirección en el Departamento de Ciencias de la Computación e Inteligencia Artificial de la Universidad de Granada. Esta memoria constituye la Tesis que Dña. María Ángeles Mendoza Pérez presenta para optar al grado de Doctor por la Universidad de Granada.

Granada, 2009

DIRECTOR: Nicolás Pérez de la Blanca Capilla

DOCTORANDO: María Ángeles Mendoza Pérez

Caminante no hay camino se hace camino al andar.
— Antonio Machado

a mis padres

RESUMEN

El campo del reconocimiento de acciones humanas ha recibido y recibe en la actualidad una especial atención debido a sus numerosas áreas de aplicación. De entre todas las técnicas aplicables a este campo los modelos de espacios de estados que usan herramientas estadísticas se han revelado como un efectivo método capaz de afrontar la incertidumbre asociada a las acciones humanas. Un amplio rango de modelos gráficos probabilísticos han sido propuestos en la literatura, pero no todos han sido evaluados con profundidad en el campo del reconocimiento de acciones, como por ejemplo los modelos ocultos de Markov factoriales que se han enfocado principalmente en reconocer personas caminando, o en el caso de los campos aleatorios condicionales con estados ocultos en el reconocimiento de gestos.

En esta tesis, contextualizamos los principales tipos de modelos gráficos probabilísticos, dirigidos y no dirigidos, en el ámbito del reconocimiento de acciones humanas, examinando sus ventajas y desventajas frente al problema tratado y realizando un completo estudio experimental. Mientras otros trabajos encontrados en la literatura generalmente se limitan a comparar dos o tres de estos modelos, en esta tesis atacamos esta comparación bajo una metodología común y una misma base de datos considerada estándar en este campo que nos permite inferir conclusiones generales.

Así mismo, adaptamos a nuestro problema otros modelos gráficos propuestos en otros campos pero que por sus propiedades creemos que pueden ser especialmente adecuados en la tarea que nos ocupa. Este es el caso de los campos aleatorios de Markov generativos, especificados para el reconocimiento de objetos en imágenes estáticas y que nosotros extendemos a tratar señales temporales, su definición nos va a permitir estudiar densas relaciones entre los estados de las variables en distintos instantes de tiempo. El producto de modelos ocultos de Markov, usados

en el procesamiento del lenguaje escrito, se revela en este trabajo como una interesante alternativa a los modelos probabilísticos actuales en el reconocimiento de acciones, con mayor capacidad representativa que los tradicionales modelos ocultos de Markov, son capaces de modelar acciones generadas por movimientos concurrentes que evolucionan con distinta dinámica, así como la estructura de un movimiento en múltiples escalas de tiempo de un modo eficiente frente a otros modelos de Markov multi-cadena como los factoriales o los acoplados gracias a su topología. Nosotros reformulamos este modelo para incorporar distribuciones estadísticas con varias componentes gaussianas, ya que usualmente las distintas realizaciones de una misma acción humana presentan demasiada variabilidad para ser modeladas por una distribución simple.

En este trabajo también abordamos el principal escollo en los modelos con conexiones no dirigidas, la constante de normalización global, generalmente de coste computacional prohibitivo, nosotros proponemos un esquema de muestreo por importancia basado en enfriamiento para estimar esta constante de forma plausible en el caso del producto de modelos ocultos de Markov.

PUBLICACIONES

Algunas ideas y figuras de esta tesis han aparecido previamente en las siguientes publicaciones:

- María Ángeles Mendoza, Nicolás Pérez de la Blanca y Manuel J. Marín-Jiménez, *Fitting Product of HMM to Human Motions*, Proceedings of the 13th International Conference on Computer Analysis of Images and Patterns, LNCS 5702, pp. 824–831, 2009.
- María Ángeles Mendoza, Nicolás Pérez de la Blanca y Manuel J. Marín-Jiménez, *PoHMM-based Human Action Recognition*, International Workshop on Image Analysis for Multimedia Interactive Services, editado por IEEE Computer Society, pp. 85–88, 2009.
- Manuel J. Marín-Jiménez, Nicolás Pérez de la Blanca, María Ángeles Mendoza, Manuel Lucena y José M. Fuertes, *Learning Action Descriptor for Recognition*, International Workshop on Image Analysis for Multimedia Interactive Services, editado por IEEE Computer Society, pp.5–8, 2009.
- María Ángeles Mendoza y Nicolás Pérez de la Blanca, *Applying Space State Models in Human Action Recognition: A Comparative Study*, Proceedings of the 5th international conference on Articulated Motion and Deformable Objects, LNCS 5098, pp. 53–62, 2008.
- María Ángeles Mendoza y Nicolás Pérez de la Blanca, *HMM-Based Action Recognition Using Contour Histograms*, Proceedings of the 3rd Iberian conference on Pattern Recognition and Image Analysis, Part I, LNCS 4477, pp. 394–401, 2007.

AGRADECIMIENTOS

En primer lugar quiero dar las gracias a Nicolás Pérez la Blanca, cuya visión y búsqueda de superación fueron los verdaderos artífices de esta tesis.

Quisiera también agradecer a mis compañeros de grupo, Manuel Jesús, Manolo, José Manuel y Nacho, que me enseñaron el arte del buen comer.

Al resto de mis compañeros que durante todo este tiempo se preocuparon por mí y me dieron ánimo, Ignacio, Jesús, Nacho, Nico, Silvia, Waldo y un largo etcétera.

A Javi, mi compañero de despacho, y a Guillermo, mi compañero en la vida, por sufrirme.

Y en especial, quiero dar las gracias a mi hermana Consuelo, que durante este último año ha sido mi chica para todo, por su inestimable ayuda.

CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

NOTACIÓN

A lo largo de esta tesis usaremos las siglas de las palabras correspondientes en el lenguaje inglés, ya que al lector versado en el tema esta terminología le resultará más familiar. A continuación listamos el significado de las siglas que más aparecen a lo largo de este trabajo.

- Campos aleatorios condicionales - CRF (*Conditional Random Fields*).
- Campos aleatorios de Markov - MRF (*Markov Random Fields*).
- Modelo gráfico - GM (*Graphical Model*).
- Modelo gráfico dirigido - DGM (*Directed Graphical Model*).
- Modelo gráfico no dirigido - UGM (*Undirected Graphical Model*).
- Modelo oculto de Markov - HMM (*Hidden Markov Model*).
- Modelo oculto de Markov acoplado - CHMM (*Coupled Hidden Markov Model*).
- Modelo oculto de Markov factorial - FHMM (*Factorial Hidden Markov Model*).
- Modelo oculto de Markov paralelo - PaHMM (*Parallel Hidden Markov Model*).
- Producto de modelos ocultos de Markov - PoHMM (*Product of Hidden Markov Model*).
- Reconocimiento de acciones humanas - HAR (*Human action recognition*).
- Red bayesiana - BN (*Bayesian Net*).

- Red Bayesiana dinámica - DBN (*Dynamic Bayesian Net*).

1

INTRODUCCIÓN GENERAL

1.1 MOTIVACIÓN

Un sistema capaz de reconocer automáticamente acciones humanas capturadas mediante una videocámara integra un amplio abanico de aplicaciones como: a) *Vídeo-vigilancia*, existen muchos escenarios en los que un sistema que llame la atención en tiempo real sobre acciones sospechosas es altamente beneficioso (por ejemplo, bancos, aeropuertos, fronteras, etc.); b) *Sistemas interactivos* para fines sociales (por ejemplo descripción oral de una secuencia de vídeo a personas invidentes), o para la industria del entretenimiento (como la nueva generación de consolas en las que el juego interactúa con el usuario atendiendo a sus movimientos); c) *Anotación e indexación de vídeos* en bibliotecas digitales ya que el enorme crecimiento que ha experimentado en los últimos años los datos multimedia (véase youtube) hace tedioso y en ocasiones impracticable el etiquetado manual, además de estar sujeto a la subjetividad del etiquetador.

El reconocimiento de acciones humanas (HAR – *Human Action Recognition*) es sin embargo un problema complejo debido a los numerosos factores implicados, como la gran diversidad existente entre las personas tanto en su apariencia (constitución física, ropas,. . .) como en el estilo de ejecución de la acción; el escenario donde se realiza, a menudo se trata de entornos concurridos, afectados por sombras, cambios de iluminación o oclusiones; y otros factores como el ángulo de vista y la distancia del sujeto respecto a la cámara. Las acciones humanas llevan asociadas una componente espacial y una temporal, ambas altamente aleatorias, por lo que la realización de una misma acción nunca es idéntica.

Los modelos gráficos probabilísticos tratan de forma natural esta incertidumbre, manejando con éxito señales que varían en el tiempo y eventos con gran variabilidad espacial

y temporal, por lo que unido a su capacidad de representar visualmente las dependencias entre las variables implicadas y la existencia de eficientes algoritmos de inferencia sobre ellos, los convierte en una potente técnica para el modelado y el reconocimiento de acciones humanas.

Distintas estructuras de grafo (las variables y sus relaciones) especifican distintos tipos de modelos. En el caso en que esté formado por variables discretas que contengan información sobre la evolución de los estados por los que atraviesa el fenómeno modelado definen los modelos de espacio de estados probabilísticos, los cuales son capaces de explotar el ordenamiento temporal de la acción humana, capturando tanto características estructurales como dinámicas de forma compacta. Así por ejemplo, el HMM es un caso particular sencillo de estos modelos.

1.2 ANTECEDENTES

El reconocimiento de acciones humanas es una área de investigación muy activa y en constante expansión debido a la gran cantidad de potenciales aplicaciones. Existen por tanto numerosas técnicas que se aplican esta tarea, bien basadas en la definición de un modelo del cuerpo humano (2D o 3D) o en las características extraídas directamente de las muestras de vídeo; bien usando información sobre la forma del sujeto o sobre su patrón temporal. En [1, 2, 3, 4, 5] podemos encontrar una amplia recopilación de los trabajos más destacados sobre el análisis del movimiento humano y el reconocimiento de acciones. Los métodos propuestos se agrupan principalmente de dos categorías:

a. Aquellos que aproximan la cuestión como un tradicional problema de emparejamiento de patrones, en los que se extrae un patrón de la secuencia de imágenes y se compara con un prototipo almacenado. La principal dificultad asociada a esta categoría estriba en la representación de la información visual siendo la clasificación usualmente simple y directa. Algunos de los clasificadores más usados son el vecino más cercano [6, 7, 8], las

máquinas de vectores soporte [9, 10] o las redes neuronales artificiales [11, 12].

b. Los modelos de espacios de estados, que definen un espacio de estados en el que cada punto de este espacio representa un evento (pose, movimiento, ...) de la acción, proporcionando un modo compacto de describirla. Estos modelos pueden ser deterministas, como las maquinas de estado finitas [13, 14, 15] (FSM -*Finite State Machines*), que conciben las acciones como procesos totalmente observables en el que el conjunto de estados es conocido; o probabilísticos, en los que a cada estado se le asocia una probabilidad, siendo la probabilidad conjunta del camino a través de los estados que conforma el modelo de cada acción el criterio usado en la clasificación.

Recientemente, se ha definido una nueva categoría [5]: Los modelos de eventos semánticos [16, 17, 18], en el que las acciones son definidas en términos cualitativos mediante reglas semánticas o lógicas entre los eventos que las componen. Nosotros no nos ocuparemos en este trabajo de este tipo de acciones, sino de aquellas cuyos vectores de características son extraídos directamente de las imágenes.

Las técnicas englobadas en la categoría (a) suelen tener bajo coste computacional durante la clasificación pero la extracción de características puede ser muy compleja [19, 20], además son más susceptibles al ruido y a otros factores como el punto de vista de la cámara, o la distinta duración de una misma acción. Los modelos de espacios de estados, en cambio, manejan de forma natural esta variabilidad temporal en la ejecución de las acciones ya que el mismo estado puede visitarse repetidamente.

En general los modelos de espacio de estados se adecuan convenientemente al problema de HAR, ya que una acción puede interpretase intuitivamente como una sucesión ordenada de eventos en el espacio y el tiempo. De ellos, los probabilísticos además tratan estadísticamente amplias variaciones en el dominio del espacio y del tiempo y en concreto la gran variabilidad en el movimiento y la apariencia entre las personas, por lo que son intensivamente explota-

dos en este campo. No obstante, presentan el inconveniente de que la búsqueda de la secuencia de estados óptima que identifique a la acción requiere generalmente de algoritmos iterativos que pueden ser computacionalmente costosos.

De estos modelos los que gozan de mayor popularidad son los modelos ocultos de Markov (HMM) y sus variantes, de probado éxito en otros problemas como el reconocimiento de gestos [21, 6, 21], y de los que inclusive existen trabajos de desarrollo de hardware para acelerar la decodificación en los HMM [22]. Los HMM se emplean para reconocer desde acciones sencillas (como andar, correr, etc.) hasta acciones de semántica más compleja [23, 24], y son alimentados con características de muy diverso tipo: de bajo [25], medio [26] o alto nivel [27], basadas en la forma [28, 29] o en el movimiento [30] del sujeto que realiza la acción o en una combinación de ambas [31, 32].

HMM con topologías no lineales han sido propuestos para solucionar algunas de las limitaciones de los tradicionales HMM, como su capacidad representativa. En [33] HMM factoriales son utilizados para modelar el andar humano, en ellos la secuencia de observación depende del estado actual de distintas cadenas de Markov. En [34, 35] se emplean HMM paralelos, donde cada cadena genera una secuencia de observación distinta y la probabilidad resultante es la multiplicación de las probabilidades de cada HMM obtenidas independientemente. Brand *et al.* [36] usan los HMM acoplados para clasificar tres típicos movimientos de Tai Chi, en este modelo las cadenas de estados responsables de cada secuencia de observación están conectadas entre sí, lo que permite codificar la coordinación en el movimiento de distintas partes del cuerpo humano (por ejemplo, los gestos de ambas manos son independientes pero actúan coordinadamente para formar un movimiento global de Tai Chi). Encontramos también otros trabajos basados en modelos de arquitectura más compleja, como los HMM en capa [37] (*layered HMM*) y los HMM jerárquicos [38], generalmente estos modelos suelen utilizarse en acciones formadas por la combinación de varias actividades (por ejemplo las tareas diarias en una oficina). Una recopilación

sobre los trabajos más recientes en HMM y sus principales extensiones se recoge en [39].

Todos los modelos anteriormente citados son casos particulares de redes bayesianas dinámicas (DBN). Existen trabajos que directamente utilizan las DBN [40, 41, 42, 43] e incluso BN [44, 45], que permite introducir de forma sencilla el conocimiento a priori del problema en el modelo y que hacen uso de técnicas genéricas de inferencia.

Los campos aleatorios del Markov (MRF), que a diferencias de las BN sus conexiones no son dirigidas (la influencia entre variables es simétrica), tienen una larga historia en visión por ordenador [46], pero sólo en los últimos años un caso particular de los MRF, los campos aleatorios condicionales (CRF), han sido explorados en el área de HAR. Los CRF fueron originariamente definidos por Lafferty *et al.* [47] para el modelado del lenguaje natural en textos, estos modelan directamente la secuencia de etiquetas que definen las clases condicionada sobre las observaciones, no contienen variables ocultas y todas las variables son conocidas durante el entrenamiento. En la bibliografía encontramos trabajos que refieren iguales o mejores resultados en nuestro problema de HAR [48] cuando se usa la probabilidad condicional en lugar de la probabilidad conjunta para la clasificación, lo que señala a estos clasificadores discriminativos como una buena alternativa frente a los clasificadores generativos basados en DBN [49].

El número de trabajos en el campo de HAR basados en CRF es aún modesto con respecto a los HMM pero está creciendo rápidamente. En [50] los CRF son alimentados con histogramas de alta dimensión formados por las posiciones relativas entre puntos y pares de puntos extraídos de la silueta humana en diferentes escalas espaciales. En [51] se emplean histogramas 2D constituidos únicamente de las posiciones entre puntos de la silueta en una sola escala mostrando que el uso de características más sencillas también da lugar a buenos resultados. Los eventos con una maquina expendedora son reconocidos usando la posición, la velocidad y la pose del sujeto en [52]. Los CRF ocultos (HCRF) y los CRF de dinámica latente (LDCRF) son extensiones a los CRF que incorporan una capa de variable

ocultas de forma que sea posible modelar la estructura subyacente en los gestos humanos, como en [53] con un HCRF o en [54] con un LCRF. En [55] la capa de observación del LDCRF es remplazada por un estimador de pose que extrae características más compactas desde las observaciones de alta dimensión.

En los trabajos citados, se han utilizado CRF sencillos con una topología similar a los HMM, CRF con estructuras más complejas han sido también aplicados a este campo [56, 57, 58, 59]. En [56] se incluyen distintos nodos para distintos observables para el seguimiento de poses y el reconocimiento de acciones simultáneo. En [57] una cascada de CRF aprende el patrón del movimiento de las distintas partes en que se ha descompuesto jerárquicamente el cuerpo humano. En [58] el reconocimiento de acciones se basa en los CRF factoriales (FCRF) definidos por Sutton [60]. En [59] se comparan los CRF jerárquicos y los FCRF en acciones donde se manipulan objetos. No obstante, debido a las dependencias que se establecen entre las variables en estos modelos, el coste de la inferencia es considerablemente mayor que en los CRF de estructura más simple, por lo que al igual que en el caso de las DBN son aplicados a acciones de alto nivel con complejas relaciones entre los eventos que las componen [61, 62].

En general, el coste computacional de los CRF lineales es equivalente al de los HMM en la fase de reconocimiento pero requiere significativamente más tiempo en la fase de entrenamiento, mayor cuanto mayor sea la dimensión del vector de características. En [63] sugieren optimizar el entrenamiento minimizando el error cuadrático de un subconjunto de muestras de entrenamiento en lugar de maximizar la verosimilitud logarítmica. En [64] proponen un nuevo algoritmo para mejorar la inferencia en los LDCRF combinando una estrategia de búsqueda eficiente con programación dinámica.

1.3 OBJETIVOS

El objetivo general de esta tesis es el análisis del comportamiento de los modelos de estado probabilísticos en la resolución del problema del reconocimiento automático de acciones humanas desde secuencias de vídeo monoculares.

Para ello, en primer lugar examinamos los principales modelos gráficos probabilísticos aplicados al reconocimiento de acciones humanas (HAR). Analizando sus ventajas y debilidades en la realización de esta tarea y llevando a cabo un exhaustivo estudio comparativo de todos ellos en el mismo marco experimental. Algunos de estos modelos no han sido explorados con profundidad en este campo, como es el caso de los *modelos ocultos de Markov factoriales* [65] o el de los *campos aleatorios condicionales* con variables ocultas [66, 54], por otro lado los trabajos encontrados en la literatura sólo comparan dos o tres de estos modelos, por lo que una comparación que abarque un amplio rango de estos modelos evaluados en iguales condiciones y bajo una base de datos común es necesario para extraer conclusiones globales.

Y en segundo lugar, exploramos otros modelos gráficos que aunque propuestos en otros campos, como el reconocimiento de objetos en el caso de *campos aleatorios de Markov generativos* [67] o el procesamiento del lenguaje escrito en el caso del *producto de modelos ocultos de Markov* [68], creemos que son apropiados para el problema de HAR debido a sus características por lo que los adaptamos a esta tarea.

Estos modelos son evaluados usando características visuales de bajo nivel, las cuales son generales a distintos problemas de visión por ordenador y usualmente fácil y rápidamente extraíbles de las imágenes con técnicas sencillas (como primer paso a sistemas de reconocimiento que operen a tiempo real), y que puedan obtenerse en condiciones realistas de trabajo (sombras, cambios de iluminación, del punto de vista, del zoom de la cámara, etc.).

1.4 ORGANIZACIÓN DE LA TESIS

Esta tesis se organiza de la siguiente forma: El capítulo 2 introduce los fundamentos de los modelos gráficos probabilísticos sobre los que se apoyan los modelos que aplicamos al problema de HAR, así como los algoritmos de inferencia que utilizamos en su resolución.

En el capítulo 3 se describe los principales modelos de espacios de estados utilizados en la tarea de HAR, los HMM clásicos y sus extensiones multi-cadena (factoriales, acoplados y paralelos), los cuales son casos particulares de DBN. En este capítulo también describimos el modelo mixto denominado *el producto de modelos ocultos de Markov*, cuya topología permite modelar distintos aspectos de los datos de forma eficiente. Complementariamente, atacamos la estimación de su constante de normalización global en un modo computacionalmente viable basándonos en el muestreo de una serie de distribuciones muy similares (*annealed importance sampling*) partiendo desde una distribución con constante de normalización conocida. Por último, todos los modelos son analizados en el contexto del problema de HAR, evaluando sus aportaciones específicas a esta tarea.

En el capítulo 4 se aborda la tarea de HAR usando los campos aleatorios de Markov (MRF). En concreto, mediante dos casos particulares: un modelo discriminativo, el campo aleatorio de Markov (CRF) y aquellas extensiones que incorporan variables ocultas modelando la estructura interna de la acción (HCRF y LDCRF); y un modelo generativo que denominamos *MRF generativo*, en el que las variables de estado se conectan entre sí a través de conexiones no dirigidas y mediante conexiones dirigidas con las variables de observación (modelo mixto).

La base de datos con las secuencias de acciones evaluadas y la extracción de características son detalladas en el capítulo 5. Nosotros consideramos una acción desde dos puntos de vista, como una sucesión de poses estáticas en el que usamos características basadas en el contorno del sujeto que realiza la acción, o como una composición de pequeños movimientos en el que usamos características extraídas del flujo óptico entre dos fotogramas de la secuencia.

En el capítulo 6 se describe el marco experimental sobre el que examinamos los modelos expuestos y discutimos los resultados obtenidos: comparación de los modelos generativos frente a los discriminativos, de los modelos lineales frente a los modelos multi-cadena y de los modelos mixtos que hemos propuesto para su aplicación en el campo de HAR, el MRF generativo y el producto de HMM.

En el capítulo 7 se exponen las conclusiones extraídas de los resultados obtenidos en esta tesis y se esbozan las principales líneas de trabajo futuro.

Por último, en el apéndice A se listan las derivadas con respecto a los parámetros del producto de HMM con distribución de salida formada por la combinación lineal de gaussianas.

Parte I

Modelos

2

MODELOS GRÁFICOS PROBABILÍSTICOS

En este capítulo describimos brevemente los principales fundamentos de los modelos gráficos probabilísticos, de los cuales los modelos que aplicaremos al problema de HAR son casos particulares, así como aquellos algoritmos asociados a los GM que empleamos en la resolución de esta tarea.

2.1 LOS MODELOS GRÁFICOS PROBABILÍSTICOS

Los modelos gráficos probabilísticos [69] (GM – *graphical model*), han sido empleados en una gran variedad de áreas, como diagnóstico clínico, codificación de señales, modelado del lenguaje escrito, reconocimiento de voz, etc. En la actualidad, su aplicación al problema del reconocimiento de acciones humanas ha experimentado un notable auge debido fundamentalmente a que manejan la aleatoriedad inherente en las acciones humanas de forma natural y pueden expresar intuitivamente los fenómenos físicos subyacentes así como sus relaciones mediante el lenguaje gráfico visual. Por otro lado, estos modelos proporcionan una metodología general de inferencia que explotan eficientemente la estructura del grafo para hallar probabilidades marginales y condicionales.

Combinación de teoría de probabilidad y teoría de grafos, los GM representan gráficamente la independencia condicional entre variables aleatorias y definen una factorización de la distribución probabilidad $P(X)$. Un grafo se define por un conjunto de nodos V que representan las variables aleatorias, las cuales pueden ser discretas (que toman un conjunto finito de valores) o continuas (descritas por una distribución paramétrica), y un conjunto de aristas E que establecen relaciones entre estos nodos codificando las propiedades de independencia condicional sobre las variables aleatorias. Se dice que un nodo (variable) del grafo es observado cuando

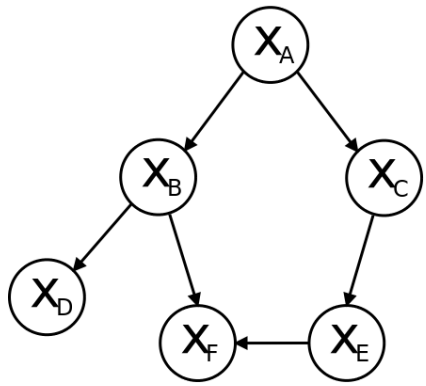

Figura 1: **Ejemplo de un modelo gráfico dirigido.** Los nodos representan variables aleatorias y los arcos indican el orden en los nodos codificando las independencias condicionales.

conocemos los valores que toma esa variable. Un nodo es no observado cuando no conocemos su valor (desconocido) o los datos no son asequibles directamente (oculto).

De aquí en adelante denotaremos con X las variables observadas, con Y las variables que queremos predecir, y con S las variables ocultas, bien porque sean demasiado costosas de obtener o porque se traten de variables hipotéticas, y de las que sólo conocemos su distribución condicional dadas las variables observadas.

2.2 MODELOS GRÁFICOS DIRIGIDOS

En los modelos gráficos dirigidos [69] (DGM – *Directed Graphical Model*) las conexiones llevan asociadas una dirección (convencionalmente se les denominan como arcos) tal que X_i precede a X_j, o equivalentemente X_i es el padre de X_j.

La redes bayesianas [70] (BN – *Bayesian Network*) son un subconjunto de los DGM, en los que los grafos no contienen ciclos (acíclico), es decir, para cada nodo del grafo no hay un camino dirigido que empiece y acabe en dicho nodo. La figura 1 presenta un ejemplo de BN. Las BN pueden tener una interpretación causal, si X_A es el padre de X_B

puede implicar una relación causa-efecto entre X_A y X_B respectivamente.

Las propiedades de independencia condicional en este tipo de grafos son definidas mediante el criterio de d-separación [71]. Sean X_{C_i}, X_{C_j} y X_{C_k} tres conjuntos disjuntos de variables (nodos) en un grafo dirigido, X_{C_i} y X_{C_j} son condicionalmente independientes dado X_{C_k} si todos los caminos entre cualquier nodo de X_{C_i} y cualquier nodo de X_{C_j} están bloqueados por X_{C_k}. El camino está bloqueado si tiene un nodo X_v tal que:

1. Los arcos a lo largo del camino no convergen en X_v y X_v pertenece a X_{C_k}.
2. Los arcos a lo largo del camino convergen en X_v y ni X_v ni sus descendientes pertenecen a X_{C_k}.

La función de probabilidad puede entonces factorizarse según la propiedad local de Markov dirigida [69], derivada del criterio de d-separación, por el cual una variable X_i es condicionalmente independiente de sus no descendientes dado el valor de sus padres:

$$P(X_1, X_2, ..., X_N) = \prod_{i=1}^{N} P(X_i|X_{1:i-1}) = \prod_{i=1}^{N} P(X_i|pa(X_i)) \quad (2.1)$$

donde $pa(X_i)$ representa los padres de X_i. Por ejemplo, sea el grafo de la figura 1, la función de probabilidad conjunta queda como:

$$P(X_A,X_B,X_C,X_D,X_E,X_F) = P(X_A)P(X_B|X_A)P(X_C|X_A)P(X_D|X_B)P(X_E|X_C)P(X_F|X_B,X_E)$$

Por lo que dada la ley distributiva, la marginalización de la probabilidad conjunta con respecto a la variable X_A se reduce a:

$$\begin{aligned} P(X_A) &= \sum_{B,C,D,E,F} P(X_A, X_B, X_C, X_D, X_E, X_D) \\ &= P(X_A) \sum_{B} P(X_B|X_A) \sum_{C} P(X_C|X_A) \\ &\quad \sum_{D} P(X_D|X_B) \sum_{E} P(X_E|X_C) \sum_{F} P(X_F|X_B, X_E) \end{aligned}$$

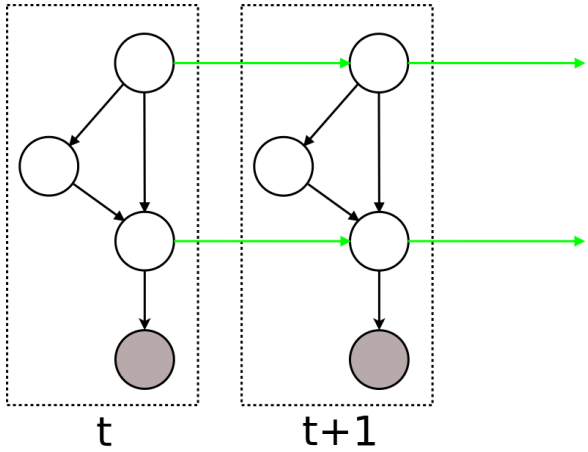

Figura 2: **Ejemplo de una DBN.** La estructura de la red se repite a lo largo del tiempo. Los arcos en color representan la dirección temporal.

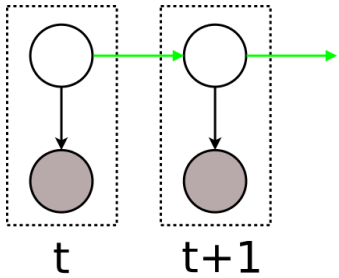

Figura 3: **Ejemplo de un HMM.** Representa el caso más simple de DBN.

Otra propiedad que se obtiene del criterio de d-deparación es que cada nodo del grafo depende condicionalmente de sus padres, sus hijos y de los padres de sus hijos, a este conjunto de variables se le denomina manto de Markov del nodo.

Redes bayesianas dinámicas

Si la estructura de las BN se repite a lo largo del tiempo con nodos conectados en la dirección temporal se les denomina redes bayesianas dinámicas (DBN – *Dynamic Bayesian Network*). La figura 2 muestra un ejemplo de DBN, los arcos en color representan la dirección temporal. Las DBN son particularmente adecuadas para modelar señales tempo-

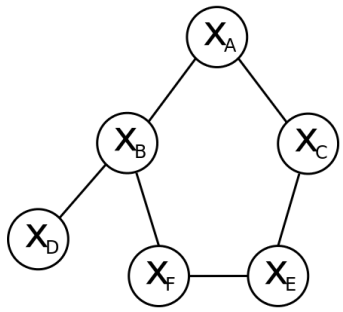

Figura 4: **Ejemplo de un modelo gráfico no dirigido.** Los nodos representan variables aleatorias y las aristas relaciones de tipo correlacional.

rales [72], como es el caso de la evolución de las poses o la sucesión de movimientos en acciones humanas. El caso más simple de DBN correspondería a un HMM (ver figura 3).

2.3 MODELOS GRÁFICOS NO DIRIGIDOS

En los modelos gráficos no dirigidos [69] (UGM – *Undirected Graphical Model*), también conocidos como campos aleatorios de Markov (MRF – *Markov Random Field*), no se especifica un orden entre sus variables aleatorias (nodos), se caracterizan porque los valores que toman las variables conectadas entre sí están correlacionados [73] (influencia entre nodos simétrica). Gráficamente estas relaciones se representan mediante aristas, las cuales no llevan asociadas una dirección.

Las propiedades de independencia condicional en los UGM son más sencillas que en los DGM [74]. Sean X_{C_i}, X_{C_j} y X_{C_k} tres conjuntos disjuntos de variables (nodos) en un grafo no dirigido, X_{C_i} y X_{C_j} son condicionalmente independientes dado X_{C_k} si todos los caminos entre cualquier nodo de X_{C_i} y cualquier nodo de X_{C_j} contienen al menos un nodo de X_{C_k}. El manto de Markov para un nodo del UGM está formado por los vecinos de dicho nodo.

Mientras que en los DGM los arcos indican una distribución condicional local entre el nodo padre y el nodo hijo, en los UGM las relaciones son modeladas mediante funciones

potenciales tal que $\Psi_C : \mathcal{V}^n \rightarrow \mathcal{R}^+$ definidas sobre subgrafos completamente conectados (cliques). La probabilidad conjunta puede entonces factorizarse como el producto de estas funciones locales, de forma que la distribución sobre un gran número de variables aleatorias queda representada por el producto de funciones que depende sólo de un pequeño numero de variables:

$$P(X_1, X_2, ..., X_N) = \frac{\prod_{C \in \mathcal{C}} \Psi_C(X_C)}{Z} \tag{2.2}$$

$$Z = \sum_X \prod_{C \in \mathcal{C}} \Psi_C(X_C) \tag{2.3}$$

con $\mathcal{C}$ el conjunto de cliques del grafo. Ψ_C es la función potencial del clique C especificando como influye el conjunto de variables de este clique, X_C, en la probabilidad resultante dependiendo de los valores que toman. Z es una constante de normalización denominada función de partición, que se calcula sumando sobre todas las configuraciones posibles de valores de las variables aleatorias, por lo que generalmente su cálculo exacto es computacionalmente costoso.

Por ejemplo, dado el grafo no dirigido de la figura 4 la distribución de probabilidad queda como:

$$P(X_A, X_B, X_C, X_D, X_E, X_F) = \frac{1}{Z}\Psi(X_A, X_B)\Psi(X_A, X_C)\Psi(X_B, X_D) \Psi(X_B, X_F)\Psi(X_C, X_E)\Psi(X_E, X_F) \tag{2.4}$$

2.4 MODELOS GRÁFICOS MIXTOS

En los modelos gráficos mixtos [69] se combinan aristas y arcos relacionando las variables aleatorias. La figura 5 ilustra un ejemplo de este modelo. Teóricamente estos modelos podrían capturar más relaciones estadísticas que aquellos modelos con conexiones de un sólo tipo [75]. Estos grafos pueden descomponerse en grafos dirigidos y no

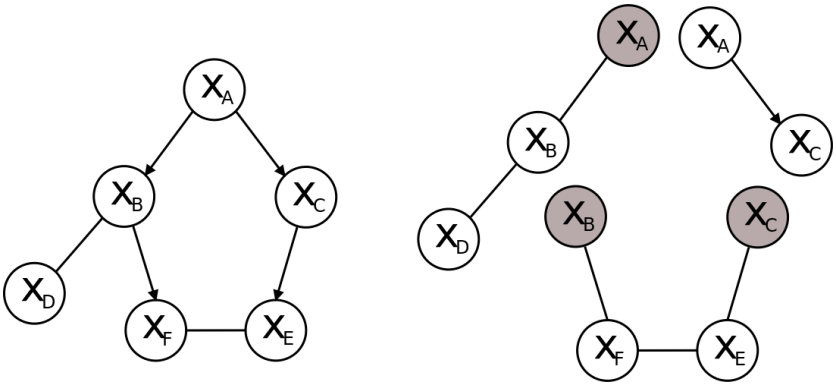

Figura 5: **Modelo gráfico mixto.** Derecha, ejemplo de un GM mixto. Izquierda, descomposición del grafo en componentes dirigidos y no dirigidos.

dirigidos [76] (figura 5), en cada subgrafo se aplicaría las propiedades de los grafos dirigidos o no dirigidos respectivamente.

$$P(X_1, X_2, ..., X_N) = \frac{\prod_{C \in \mathcal{C}} \Psi_C(X_C | pa(X_C)/X_C)}{Z} \tag{2.5}$$

Las funciones potenciales del clique C están condicionadas en los padres de las variables X_C del clique, $pa(X_C) = \cup_{c \in C} pa(X_c)$, excepto aquellas que pertenecen al propio clique. Dado el grafo de la figura 5 su función de probabilidad conjunta toma la forma:

$$P(X_A, X_B, X_C, X_D, X_E, X_F) = \frac{1}{Z} p(X_A) p(X_C|X_A) \Psi(X_B, X_D|X_A) \Psi(X_E, X_F|X_B, X_C) \tag{2.6}$$

2.5 INFERENCIA PROBABILÍSTICA

La inferencia consiste en el cálculo de las probabilidades marginales y condicionales dadas algunas observaciones mediante un conjunto de algoritmos asociados al grafo, las cuales son esenciales para efectuar predicciones basadas en el modelo y en el aprendizaje de sus parámetros.

Dichos algoritmos emplean una estrategia de programación dinámica, como es el caso del algoritmo de Avance-

Retroceso [77] y el algoritmo de Viterbi [78], los dos algoritmos tradicionalmente empleados con los HMM.

2.5.1 *Algoritmo Avance-Retroceso*

El algoritmo avance-retroceso (FB - *Forward-Backward*) define dos probabilidades, una probabilidad denominada hacia adelante $\alpha_t(i) = P(X_{1:t}, S_t = i|\Theta)$, esto es, dado el modelo Θ la probabilidad conjunta de observar una secuencia de observaciones hasta el instante t, $X_{1:t}$, y que el estado de la variable de estado en ese instante sea i, $S_t = i$; y la probabilidad hacia atrás $\beta_t(i) = P(X_{t+1:T}|S_t = i, \Theta)$, que es la probabilidad de observar la secuencia de observaciones $X_{t+1:T}$ desde el instante $t+1$ hasta T dado el estado $S_t = i$ en el instante t y el modelo Θ. Ambas variables pueden calcularse recursivamente:

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{K} \alpha_t(i) P(S_{t+1} = j|S_t = i))\right] P(X_{t+1}|S_{t+1} = j) \tag{2.7}$$

con $\alpha_1(i) = P(S_1 = i)P(X_1|S_1 = i)$ y K el número total de estados que pueden tomar las variables S .

$$\beta_t(i) = \sum_{j=1}^{K} P(S_{t+1} = i|S_t = j)\beta_{t+1}(j)P(X_{t+1}|S_{t+1} = j) \tag{2.8}$$

con $\beta_T(i) = 1$.

La probabilidad de observar la secuencia $\{X_{1:T}\}$ dados los parámetros del modelo Θ es:

$$P(X_{1:T}|\Theta) = \sum_{i=1}^{K} \alpha_T(i) \tag{2.9}$$

Y la secuencia más probable de estados ocultos de $\{S_1, S_2, \ldots, S_T\}$ dada por esa secuencia de observaciones viene dada por:

$$P(S_t = i|X_{1:T}, \Theta) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{S_t} \alpha_t(i)\beta_t(i)} \tag{2.10}$$

La complejidad computacional de este algoritmo es K^2T, donde K es el número de estados posibles y T la longitud de la secuencia de observaciones.

2.5.2 *Algoritmo de Viterbi*

En la sección anterior se describió como obtener la secuencia de estados más probable de las variables $\{S_1, \ldots, S_t, \ldots, S_T\}$ dado $X_{1:T}$ y Θ mediante el algoritmo FB, sin embargo, en la práctica se emplea el algoritmo de Viterbi [78] para este propósito debido a que este algoritmo no necesita calcular la probabilidad total del modelo (ecuación 2.10). El algoritmo de Viterbi [78] sigue una estrategia similar al algoritmo FB pero sustituye el operador suma por el operador máximo. Se define la probabilidad:

$$\delta_t(i) = \max_{S_{1:t-1}} P(S_{1:t-1}, S_t = i, X_{1:t}|\Theta) \tag{2.11}$$

esto es, dado el modelo Θ la probabilidad del mejor camino hasta t en el espacio de estados con $S_t = i$ y las primeras t observaciones $X_{1:t}$, por recursión:

$$\delta_{t+1}(j) = \max_i \Big[\delta_{t-1}(i)P(S_{t+1} = j|S_t = i)\Big] P(X_{t+1}|S_{t+1} = j) \tag{2.12}$$

con $t = \{1, \ldots, T-1\}$ y donde

$$\begin{aligned} \delta_1(1) &= 1 \\ \delta_1(i) &= P(S_1 = i)P(X_1|S_1 = i) \end{aligned} \tag{2.13}$$

El estado más probable en el instante T viene dado por:

$$S_T^* = argmax_i[\delta_T(i)] \tag{2.14}$$

La secuencia óptima de estados puede calcularse entonces hacia atrás:

$$\begin{aligned} S_t^* &= \psi_{t+1}(S_{t+1}^*) \\ \psi_t(j) &= argmax_i[\delta_{t-1}(i)P(S_t = j|S_{t-1} = i) \end{aligned} \tag{2.15}$$

con $t = T-1, T-2, \ldots, 1$.

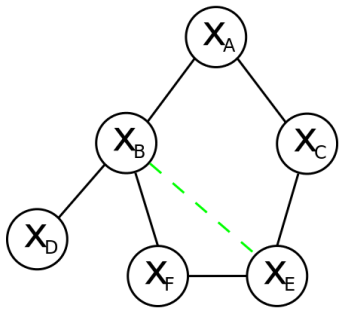

Figura 6: **Moralización del grafo dirigido.** Grafo moral del grafo dirigido de la figura 1. Se añade una conexión entre los dos padres del nodo X_F.

2.5.3 *Algoritmo de Árboles de Expansión*

Los algoritmos de inferencia anteriormente descritos sólo pueden aplicarse a DBN sencillas (como es el caso del HMM), para calcular probabilidades en grafos más complejos es necesario usar un algoritmo de propósito general, como es el algoritmo de Árboles de Expansión [79] (JT - *Junction Tree*), capaz de explotar grafos con ricas estructuras con o sin ciclos. Smyth *et al.* [74] demuestra que el algoritmo FB es un caso particular de este algoritmo.

El algoritmo consta de dos procesos: 1) El proceso de construcción, en el que se construye un árbol (sin ciclos) desde el grafo original. 2) El proceso de propagación, el cual se basa en la propagación de información local (mensajes) a través de todo el grafo tal que un acuerdo global sea alcanzado de forma eficiente.

En el primer proceso se aplica un conjunto de transformaciones que convierte el grafo original en un grafo de orden superior con estructura de árbol formado con los nodos del grafo inicial. Los pasos son los siguientes:

1. **Moralización.** Este paso sólo se aplica si el grafo original es un grafo dirigido. Antes de transformar los arcos en aristas (conexiones sin una dirección definida), se añade una arista entre padres con hijos comunes (figura 6), de esta forma el UGM resultante no desobedece las suposiciones de independencia condicional

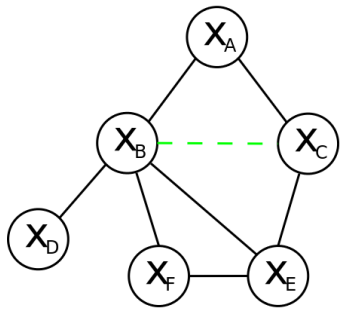

Figura 7: **Triangulación del grafo.** Ejemplo de una de las triangulaciones posibles para el grafo de la figura 6.

del DGM original, de otro modo la inferencia podría no ser correcta [80].

2. **Triangulación.** Este proceso añade aristas al UGM hasta que todos los ciclos en el grafo de longitud 4 o mayor contengan un par de nodos no consecutivos conectados por una arista [71]. El conjunto de distribuciones representadas por el nuevo grafo conteniendo a la familia de distribuciones del grafo original es ahora mayor (más aristas implica menos suposiciones de independencia). Este proceso no tiene solución única, distintas triangulaciones dan lugar a grafos compuestos por distintos cliques [81]. La figura 7 muestra el caso para una de las triangulaciones posibles del grafo de la figura 6. La eficiencia en la propagación de información en el grafo depende del tamaño (número de variables aleatorias) de los cliques, obtener una triangulación que minimice este tamaño es un problema NP-completo [82].

3. **Agrupación en cliques**. Finalmente los cliques máximos son agrupados formando un estructura de árbol (árbol de expansión), tal que verifica la propiedad de intersección consecutiva: la intersección de nodos entre dos cliques cualquiera en el árbol es contenida en todos los cliques que hay en el camino único entre estos cliques (figura 8). El peso de una arista entre dos cliques se establece como el número de variables en la

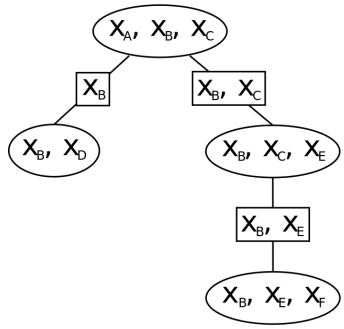

Figura 8: **Árbol de expansión.** Construcción del árbol de expansión a partir del grafo de la figura 7.

intersección de los cliques. El nuevo grafo suele representarse introduciendo nodos de separación entre dos cliques, estos nodos contienen las variables comunes S a ambos cliques y llevan asociado un potencial de separación $\phi(S)$, convencionalmente son representados con rectángulos.

En el proceso de propagación del algoritmo JT, un procedimiento de paso de mensajes es llevado a cabo sobre este árbol para asegurar que los cliques vecinos sean consistentes localmente, es decir, tengan distribuciones marginales iguales en aquellas variables que tienen en común [70]:

$$\sum_{C_i/S_{ij}} \psi_{C_i}(X_{C_i}) = \phi(S_{ij}) = \sum_{C_j/S_{ij}} \psi_{C_j}(X_{C_j}) \tag{2.16}$$

donde C_i es el clique i, S_{ij} el conjunto de variables contenidas en los cliques C_i y C_j, $S_{ij} = S_i \cap S_j$.

Dadas las características del nuevo grafo, el procedimiento converge en dos pasos:

Paso 1. El clique C_j absorbe información de C_i. Cuando el clique C_i ha recibido mensajes de todos sus vecinos excepto de C_j envía un mensaje a éste.

$$\phi^*_{ij}(S_{ij}) = \sum_{C_i/S_{ij}} \psi_{C_i}(C_i), \qquad \psi^*_{C_j}(C_j) = \frac{\phi^*_{ij}(S_{ij})}{\phi_{ij}(S_{ij})} \psi_{C_j}(C_j) \tag{2.17}$$

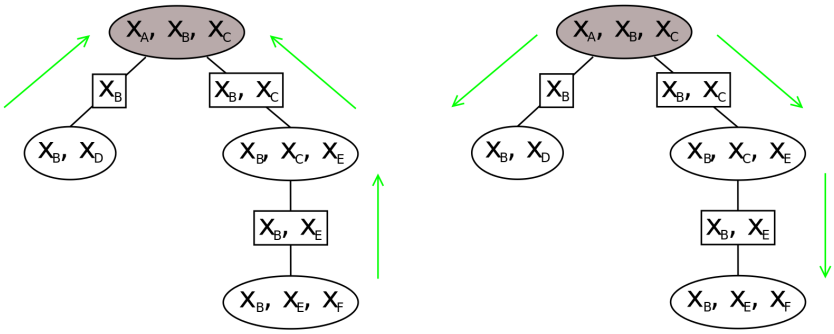

Figura 9: **Colección y distribución de la evidencia en un árbol de expansión.** El nodo sombreado corresponde al nodo raíz escogido. Derecha: la raíz recoge la evidencia de todos los nodos del grafo. Izquierda: la raíz distribuye la evidencia a todos los nodos.

Paso 2. El clique C_j absorbe información de C_i. Una vez que el clique C_j recibe información de todos sus vecinos distintos de C_i.

$$\phi^{**}_{ij}(S_{ij}) = \sum_{C_j/S_{ij}} \psi^{*}_{C_j}(C_j), \qquad \psi^{**}_{C_i}(C_i) = \frac{\phi^{**}_{ij}(S_{ij})}{\phi^{*}_{ij}(S_{ij})}\psi_{C_i}(C_i) \tag{2.18}$$

Después de convertir el grafo original en un grafo JT, se inicializa los potenciales de clique y separadores y se elige un nodo raíz. La propagación de la evidencia se lleva a cabo mediante el algoritmo HUGIN [83]. El nodo raíz pide información al resto de nodos del grafo (mensajes hacia la raíz), una vez recibida la información la distribuye (mensajes desde la raíz), lo que se denomina colección y distribución de la evidencia (ver figura 9).

La consistencia local garantiza consistencia global [80], pero sólo operaciones locales son requeridas en este procedimiento, por lo que la inferencia puede ser muy rápida. En el caso de grafos cuyos nodos corresponde a variables discretas la complejidad del algoritmo de propagación es la suma sobre todos los cliques del producto de la cardinalidad de cada variable en el clique, $O\big(\sum_{C_i \in \mathcal{C}} \prod_{c_j \in C_i} |X_{c_j}|\big)$, donde $\mathcal{C}$ es el conjunto de cliques en el árbol, C_i es un

clique de $\mathcal{C}$, c_j corresponde a la variable X_{c_j} en el clique C_i y $|X_{c_j}|$ representa su cardinalidad. Las probabilidades marginales son halladas sumando sobre las variables X_C de cada clique, la computación es exponencial al tamaño del clique mayor, por lo que en el paso de triangulación conviene formar cliques con pequeño número de variables.

3

HMM Y EXTENSIONES MULTI-CADENA

En este capítulo describimos los modelos gráficos más importantes aplicados a la tarea de HAR, como son el HMM y sus extensiones multi-cadena (FHMM, CHMM y PaHMM), todos ellos casos particulares de las DBN. También proponemos un modelo mixto multi-cadena, el PoHMM, cuyas propiedades pensamos lo convierte en una interesante alternativa a dichos modelos en la resolución de nuestro problema de HAR.

3.1 MODELOS OCULTOS DE MARKOV

La idea básica de los modelos ocultos de Markov (HMM – *Hidden Markov Model*) es la existencia de un proceso que atraviesa un serie de estados $S_{1:T} = \{S_1, \ldots, S_T\}$, la disposición de estos estados no pueden conocerse directamente (ocultos) sino la secuencia de observaciones $X_{1:T} = \{X_1, \ldots, X_T\}$ generada por ellos En esta sección planteamos los HMM desde un punto de vista bayesiano en términos de variables de estado y cardinalidad de dichas variables. En cada instante de tiempo t existe un nodo representando una variable de estado S y un nodo de observación X, padre–hijo respectivamente, cuya estructura se repite a lo largo del tiempo a través de S. La figura 10 muestra el esquema gráfico de un HMM. Los nodos en blanco representan las variables de estado ocultas S en cada instante t y los nodos sombreados la variables de observación X, los arcos entre nodos representan relaciones causales. S es una variable multinomial que toma K valores discretos, $k \in \{1, \ldots, K\}$, la cardinalidad de S corresponde al número de estados del HMM, K.

Dado que cada nodo es independiente de sus no descendientes dados sus padres, se extraen las dos siguientes relaciones de independencia condicional:

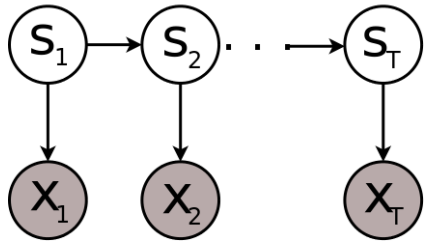

Figura 10: **HMM.** X representan las variables de observación (nodos sombreados), S las variables ocultas de estado, los arcos representan relaciones de tipo causal.

1. $S_t \perp \{S_1, X_1, \ldots, S_{t-2}, X_{t-2}, X_{t-1}\} | S_{t-1}$
2. $X_t \perp \{S_1, X_1, \ldots, S_{t-1}, X_{t-1}\} | S_t$

con $2 \leqslant t \leqslant T$ y que se traduce en las siguientes dos suposiciones sobre la distribución de probabilidad conjunta $P(X, S)$: 1) la probabilidad de estar en un estado en el instante t depende sólo del estado en el instante anterior $t-1$ (propiedad de Markov); 2) la probabilidad de la observación depende sólo del estado actual, es decir, las observaciones son independientes entre sí conocido el estado. La distribución de probabilidad se factoriza entonces como:

$$P(X_{1:T}, S_{1:T}) = P(S_1) \prod_{t=2}^{T} P(S_t|S_{t-1}) \prod_{t=1}^{T} P(X_t|S_t) \qquad (3.1)$$

Donde $P(S_1)$ representa las probabilidades de estado inicial, habitualmente denotadas con el vector $\pi = \{\pi_k\}$ de K elementos tal que $\pi_k = P(S_1 = k)$ es la probabilidad de que el modelo se encuentre en el estado k en el instante inicial $t = 1$; $P(S_t|S_{t-1})$ son las probabilidades de transición entre estados, definida por la matriz $A = \{a_{ij}\}$ de tamaño $K \times K$, a_{ij} es la probabilidad de que el modelo evolucione desde el estado i en el instante $t-1$ al estado j en el instante t; $P(X_t|S_t)$ representa las probabilidades de observación o de salida, usualmente caracterizadas por la matriz B de tamaño $K \times T$, $b_k(X_t = i)$ es la probabilidad de que un estado k en el instante t genere la observación

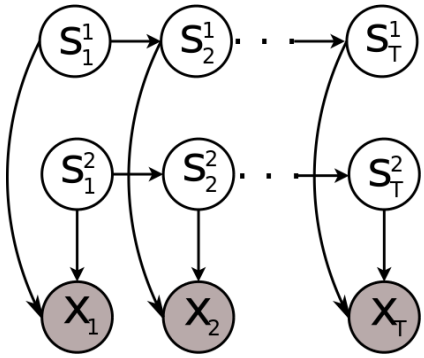

Figura 11: **FHMM.** Esquema de un FHMM con 2 cadenas de Markov, $S_t^{(m)}$ es la variable de estado de la cadena de Markov m en el instante t, y X_t la observación generada por estas cadenas en ese instante.

$X_t = i$. En nuestro caso, emplearemos características continuas obtenidas sobre el flujo óptico o el contorno extraídos de los fotogramas de la secuencia de la acción y que suponemos generados por un función de densidad gaussiana, $b_k(x) = N(x; \mu_k, \Sigma_k)$, la probabilidad de salida en cada estado es entonces caracterizada por su media μ_k y covarianza Σ_k. En este sentido los HMM pueden interpretarse como la modelización de la secuencia temporal de variables de un modelo de mixtura gaussiana. Los parámetros del modelo, $\theta = \{\pi, A, \{\mu_1, \ldots, \mu_K\}, \{\Sigma_1, \ldots, \Sigma_K\}\}$, pueden hallarse con complejidad $O(TK^2)$ mediante el algoritmo de Avance-Retroceso [77].

3.2 MODELOS OCULTOS DE MARKOV FACTORIALES

Los modelos ocultos de Markov factoriales (FHMM – *Factorial Hidden Markov Model*) fueron desarrollados por Ghahramani y Jordan [65] como extensión de los HMM tradicionales. En un HMM básico toda la información sobre la historia de la señal está contenida en una sola variable discreta (el estado oculto actual), lo que limita la capacidad representativa del HMM.

EL FHMM está compuesto por varias cadenas de Markov (sucesión de variables aleatorias que cumplen la propiedad

de Markov) cuya salida se combina para generar una única secuencia de observación $X_{1:T}$. En la figura 11 se observa que el valor de la observación X_t depende de la combinación de los valores discretos que toman sus padres. El estado S_t en el instante t es entonces representado como un conjunto de variables de estado:

$$S_t = S_t^{(1)}, \ldots, S_t^{(m)}, \ldots, S_t^{(M)} \tag{3.2}$$

El espacio de estado estará formado por el producto vectorial de estas variables.

Se ha supuesto (ver figura 11) que cada cadena de Markov evoluciona independientemente del resto de cadenas de acuerdo a su propia dinámica, entonces la probabilidad de transición se factoriza como:

$$P(S_t|S_{t-1}) = \prod_{m=1}^{M} P(S_t^{(m)}|S_{t-1}^{(m)}) \tag{3.3}$$

En total tenemos M matrices de transición entre estados, una para cada cadena m, con dimensión $K^{(m)} \times K^{(m)}$ respectivamente. Un HMM con una capacidad representativa equivalente tendría una matriz de transición de tamaño $K^M \times K^M$.

La observación en instante t depende del estado actual en cada una de las cadenas ocultas S_t, $P(X_t|S_t)$. Ésta se modela por una gaussiana con covarianza común y media especificada por una combinación lineal de las variables de estado $\mu^{(m|S_t)}$:

$$P(X_t|S_t) = \frac{e^{-\frac{1}{2}\left(X_t - \sum_{m=1}^{M} \mu^{(m|S_t)}\right)^T \Sigma^{-1} \left(X_t - \sum_{m=1}^{M} \mu^{(m|S_t)}\right)}}{\sqrt{(2\pi)^d \Sigma}} \tag{3.4}$$

donde $\mu^{(m|S_t)} = W^{(m)} S_t^{(m)}$ es la media en la cadena m dado el estado conjunto S_t (ecuación 3.2). Cada variable de estado $S_t^{(m)}$ se representa por un vector de dimensión $K^{(m)} \times 1$ con 1 en la posición del estado correspondiente y cero en el resto. $W^{(m)}$ es una matriz $d \times K^{(m)}$ cuyas columnas son las contribuciones a la media de cada estado en la cadena

m. Los parámetros de los FHMM son entonces, las probabilidades de estado inicial $\pi^{(m)}$ y las probabilidades de transición entre estados $A^{(m)}$ en cada cadena, y las medias μ y covarianzas Σ para la combinación de los estados de todas las cadenas:

$$\theta = \left\{ \left\{ \pi^{(m)}, A^{(m)} \right\}_{m=1}^{M}, \left\{ \mu_1, \ldots, \mu_{K^{(1)} \times \ldots \times K^{(M)}} \right\}, \left\{ \Sigma_1, \ldots, \Sigma_{K^{(1)} \times \ldots \times K^{(M)}} \right\} \right\}$$

El espacio de estados del FHMM está formado por el producto cartesiano del número de estados asociado a cada cadena (cardinalidad de las variables), lo que permite representar con pocos parámetros grandes espacios de estados. Por otro lado, cada cadena puede aprender la evolución de distintos procesos independientes. Sin embargo, la inferencia es más costosa que en los HMM, el árbol de expansión derivado de este modelo está formado por cliques de tamaño $M + 1$ (ver sección 2.5.3), por lo que el coste de la inferencia es $O(TMK^{M+1})$, frente a los T cliques de tamaño 2 del HMM (TK^2), suponiendo que todas las variables discretas posean la misma cardinalidad, $K^{(m)} = K \quad \forall m$.

3.3 MODELOS OCULTOS DE MARKOV PARALELOS

Los modelos ocultos de Markov paralelos (PaHMM – *Parallel Hidden Markov Model*) ofrecen la posibilidad de combinar varias características sin necesidad de concatenarlas en un único vector de características extendido. Distintas observaciones son modeladas con una cadena de Markov distinta, la probabilidad del PaHMM es el producto de la probabilidades de cada HMM (ecuación 3.1):

$$P(X_{1:T}, S_{1.T}) = \prod_{m=1} P(X_{1:T}^{(m)}, S_{1:T}^{(m)}) \tag{3.5}$$

Al igual que los FHMM, los PaHMM pueden modelar independientemente las dinámicas de distintas partes del cuerpo humano, dividiendo el vector de características en sub-vectores correspondientes a las distintas partes y modelando cada sub-vector con una cadena distinta, por lo que

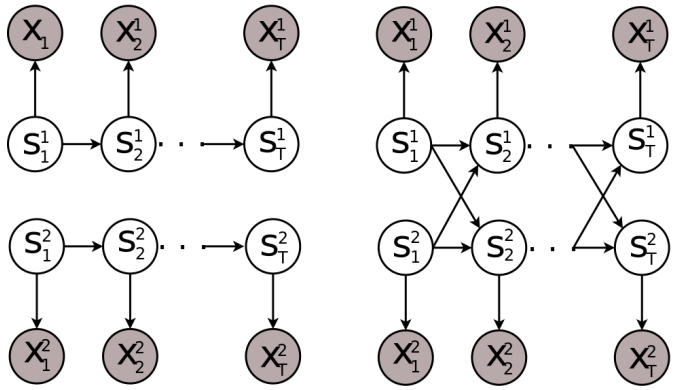

Figura 12: **PaHMM y CHMM.** Esquema de un PaHMM y un CHMM con 2 cadenas de Markov, izquierda y derecha respectivamente. En los PaHMM cada vector de observación es generado por una cadena de Markov independiente. En los CHMM, las cadenas que generan los distintos vectores de observación están conectadas entre sí.

ahora el número de parámetros del modelo es sólo M veces los parámetros de un HMM clásico:

$$\theta = \left\{\pi^{(m)}, A^{(m)}, \{\mu_1, \ldots, \mu_{K^{(m)}}\}, \{\Sigma_1, \ldots, \Sigma_{K^{(m)}}\}\right\}_{m=1}^{M}$$

por tanto el número de muestras de entrenamiento necesarias y el coste computacional ($O(TMK^2)$) es mucho menor que para el FHMM. Además, en este modelo las observaciones no están contaminadas por la existencia de ruido en partes del cuerpo no relevantes para la acción concreta, por ejemplo, por sombras en el suelo en la acción de aplaudir.

3.4 MODELOS OCULTOS DE MARKOV ACOPLADOS

Los modelos ocultos de Markov acoplados (CHMM - *Coupled Hidden Markov Model*) fueron propuestos para modelar las relaciones entre distintas señales [84]. En los CHMM, distintas secuencias de observación son generadas por distintas cadenas de Markov, las cuales están conectadas entre sí.

Algoritmo de entrenamiento en DGM

1 $\Theta = \{\theta_k\}$ iniciales
2 Se construye el grafo JT desde el grafo original
3 **while** $\mathcal{L}(\Theta)$ no converja
4 Ejecutar EM
5 ***Paso 1.***
6 $\Theta' = \Theta$
7 **for** $\forall$ muestra $X(i)$
8 introducir evidencia $X(i)$ en el JT
9 calcular las prob. marginales $P(S_t|X(i),\Theta')$, $P(S_t, S_{t-1}|X(i),\Theta')$ y $\mathcal{L}_i$ (*algoritmo HUGIN sobre JT*)
10 $\mathcal{L} = \mathcal{L} + \mathcal{L}_i$
11 **endfor**
12 Hallar el valor esperado $E[\log P(S, X|\theta_k)|X, \Theta']$
13 ***Paso 2.***
14 $\theta_k \longleftarrow argmax_{\theta_k} E[\log P(S, X|\theta_k)|X, \Theta']$
15 **endwhile**

Figura 13: Algoritmo de entrenamiento en DGM.

La probabilidad de cada variable de estado está condicionada por el valor de todas las variables de estado en el instante $t-1$ (ver figura 12):

$$A^{(m)} = P(S_t^{(m)}|S_{t-1}^{(1)}, S_{t-1}^{(2)}, \ldots, S_{t-1}^{(M)}) \tag{3.6}$$

donde M es el número total de cadenas, $S_t^{(m)}$ es el estado oculto de la cadena m en el tiempo t (determinado por el estado en todas las variables multinomiales en el instante anterior), siendo $A^{(m)}$ la matriz de transición para la cadena m, con tamaño $K^{(m)} \times (K^{(1)} \times \ldots \times K^{(M)})$. Para topologías de CHMM con más de dos cadenas la inferencia es computacionalmente prohibitiva.

3.5 APRENDIZAJE EN DGM: ALGORITMO EM

El aprendizaje de los parámetros que constituyen el modelo cuando existen variables desconocidas u ocultas puede realizarse de manera directa con el algoritmo EM [85] (*Expectation – Maximization*) en el caso de los DGM. El algoritmo Baum-Welch [86] empleado en los HMM es un caso particular de esta técnica más general. El algoritmo EM consiste en un método iterativo para estimar la máxima verosimilitud (ML – *Maximum Likelihood*) de las observaciones que hayan sido generadas por un modelo dado Θ, para ello alterna entre los siguientes dos pasos hasta que el valor de los parámetros del modelo converge:

PASO E. Se calcula el valor esperado condicional de la verosimilitud logarítmica conjunta, $\log P(S, X|\Theta)$, para las secuencias de entrenamiento X dado los parámetros del modelo actual $\Theta^{(i)}$. Dado que no se conoce el valor de las variables ocultas S se consideran sus valores esperados bajo la distribución posterior de las variables ocultas $P(S|X, \Theta^{(i)})$:

$$\Theta^{(i)} \longrightarrow Q(\Theta^{(i+1)}|\Theta^{(i)}) = E[\log P(S, X|\Theta^{(i+1)})|X, \Theta^{(i)}] \tag{3.7}$$

PASO M. Se busca $\Theta^{(i+1)}$ que maximiza la función Q.

$$\arg \max_{\Theta^{(i+1)}} Q(\Theta^{(i+1)}|\Theta^{(i)}) \longrightarrow \Theta^{(i+1)} \tag{3.8}$$

El algoritmo EM converge a un óptimo local, por lo que el resultado depende de la parámetros θ iniciales.

La probabilidad posterior $P(S|X, \Theta^{(i)})$ puede calcularse mediante el algoritmo JT (sección 2.5.3). No obstante, para algunos modelos gráficos el paso E puede ser computacionalmente intratable debido a las dependencias entre variables, en este caso se aproxima la probabilidad posterior por una distribución más simple, el paso E calcula entonces un límite inferior y en el paso M se maximiza ese límite.

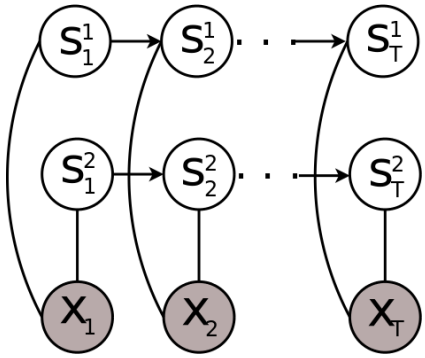

Figura 14: **PoHMM.** Diagrama de un PoHMM formado por 2 cadenas de Markov. Los nodos de color blanco representan las variables de estado ocultas S y los nodos sombreados representan las variables de observación X.

3.6 EL PRODUCTO DE MODELOS OCULTOS DE MARKOV

Como alternativa a los modelos bayesianos anteriormente descritos nosotros proponemos el producto de modelos ocultos de Markov (PoHMM – *Product of Hidden Markov Models*), el cual es un modelo gráfico mixto introducido por Brown y Hinton [68] para el modelado del lenguaje natural escrito. Nosotros creemos que debido sus características su aplicación a la tarea de HAR puede ser muy beneficiosa. Hemos reformulado el modelo enunciado en [68] para incluir como distribución de observación una combinación lineal de gaussianas y así manejar mejor la fuerte variabilidad inherente a nuestro problema.

3.6.1 *Fundamentos del PoHMM*

El PoHMM toma su origen en el producto de expertos definido por Hinton [87], en este modelo la distribución de probabilidad del modelo se debe a las distribuciones individuales de los expertos que contribuyen independientemente a la distribución total, se les denomina expertos

porque cada uno de ellos describe un aspecto concreto de los datos.

El PoHMM especifica la distribución como un producto de expertos donde los expertos son HMM. En definitiva, distintas cadenas de Markov (de primer orden) de variables de estado ocultas describen una secuencia de observación válida para cada una de las cadenas, los nodos en cada cadena de Markov se conecta mediante arcos pero las conexiones entre estos y los nodos de observación son aristas. La figura 14 muestra un diagrama de un PoHMM formado por dos cadenas. La verosimilitud del PoHMM se define entonces multiplicando las distribuciones individuales de los HMM y normalizando:

$$P(X_{1:T}|\Theta) = \frac{\prod_{m=1}^{M} P^{(m)}(X_{1:T}|\theta^{(m)})}{Z(\Theta)} \tag{3.9}$$

con

$$Z(\Theta) = \sum_{X \in X} \prod_{m=1}^{M} P^{(m)}(X_{1:T}|\theta^{(m)}) \tag{3.10}$$

donde M es el número de cadenas de Markov en el PoHMM, $X_{1:T} = \{X_1, \ldots, X_T\} \in \mathcal{X}$ es la secuencia de observación y Θ es el conjunto de parámetros del PoHMM formado por los parámetros de los HMM individuales θ en el producto, $\Theta = \left\{\theta\right\}_{m=1}^{M}$. La función de partición Z se calcula sumando sobre todas las posibles secuencias de observación $\mathcal{X}$.

Sea $s_{1:T}^{(m)} = \{s_1^{(m)}, \ldots, s_T^{(m)}\}$ la secuencia de estados más probable para la cadena m, $P^{(m)}(X_{1:T}|\theta^{(m)})$ se aproxima por:

$$P^{(m)}(X_{1:T}|\theta^{(m)}) = \max_{s^{(m)}} \left\{ P(S_1^{(m)}) \prod_{t=2}^{T} P(S_t^{(m)}|S_{t-1}^{(m)}) \prod_{t=1}^{T} P(X_t|S_t^{(m)}) \right\} \tag{3.11}$$

donde $P(S_1^{(m)}) = \pi^{(m)}$ es el vector de probabilidad de estado inicial de la cadena m, $P(S_t^{(m)}|S_{t-1}^{(m)}) = A^{(m)}$ la matriz de probabilidad de transición entre estados, $P(X_t|S_t^{(m)})$ es la

probabilidad de observación (ver sección 3.1). Según las propiedades de independencia condicional sobre aristas, dada la observación en el instante X_t las variables de estado en ese instante $\{S_t^{(1)}, \ldots, S_t^{(M)}\}$ son independientes entre sí, por lo que es posible efectuar la inferencia eficientemente ejecutando el algoritmo FB (sección 2.5.1) en cada cadena independientemente. El número de estados totales crece exponencialmente con el número de cadenas de Markov en el PoHMM pero la complejidad computacional de la inferencia sólo crece linealmente ($O(TMK^2)$).

3.6.2 *PoHMM con distribución de observación combinación lineal de gaussianas*

Nosotros generalizamos el PoHMM formulado en [88] tal que incorporen probabilidades de observación de cada estado, $P(X_t|S_t^{(m)})$, definidas como una combinación lineal de gaussianas, con el objetivo de que estas ajusten mejor la variabilidad existente en las acciones realizadas por personas:

$$P(X_t|S_t^{(m)}) = \sum_{g=1}^{G} C_{S_t^{(m)}}^{(g)} \mathcal{N}\left(mu_{S_t^{(m)}}^{(g)}, \Sigma_{S_t^{(m)}}^{(g)}\right) \tag{3.12}$$

donde G es el número de gaussianas en la mezcla, C_k^g es el peso de la g^{th} gaussiana del estado k y μ_k^g y Σ_k^g su media y covarianza respectivamente. Los parámetros totales del modelo son:

$$\Theta = \{\theta\}_{m=1}^{M} = \left\{ \pi^{(m)}, A^{(m)}, \left[C_1^{(m)}, \ldots, C_{K^{(m)}}^{(m)}, \mu_1^{(m)}, \ldots, \mu_{K^{(m)}}^{(m)}, \Sigma_1^{(m)}, \ldots, \Sigma_{K^{(m)}}^{(m)} \right]_{g=1}^{G} \right\}_{m=1}^{M}$$

con $K^{(m)}$ el número de estados para la cadena m.

3.6.3 *Aprendizaje en PoHMM: la divergencia contrastiva*

El aprendizaje de los parámetros del PoHMM es complejo a causa de la constante de normalización global Z en el denominador de la ecuación 3.9. La presencia de esta suma sobre todas las posibles secuencias hace computacionalmente prohibitivo el cálculo del gradiente exacto de la verosimilitud logarítmica con respecto a los parámetros.

$$\begin{aligned}
&\frac{\partial}{\partial\theta^{(m)}}\log P(X|\Theta)\\
&=\frac{\partial}{\partial\theta^{(m)}}\log P^{(m)}(X|\theta^{(m)})-\frac{\partial}{\partial\theta^{(m)}}\log Z\\
&=\frac{\partial}{\partial\theta^{(m)}}\log P^{(m)}(X|\theta^{(m)})-\sum_{X\in\mathcal{X}}P(X|\Theta)\frac{\partial}{\partial\theta^{(m)}}\log P^{(m)}(X|\theta^{(m)})\\
&=\frac{\partial}{\partial\theta^{(m)}}\log P^{(m)}(X|\theta^{(m)})-\left\langle\frac{\partial}{\partial\theta^{(m)}}\log P^{(m)}(X|\theta^{(m)})\right\rangle_{P(X|\Theta)}
\end{aligned} \tag{3.13}$$

No obstante, maximizar la verosimilitud logarítmica es equivalente a minimizar la divergencia de Kullback-Leibler:

$$KL(P||P^{\infty})=\sum_{i}^{N}P(X_i)\log P(X_i)-\sum_{i}^{N}P(X_i)\log P^{\infty}(X_i) \tag{3.14}$$

donde N es número de muestras X_i de entrenamiento, P^0 representa la distribución de los datos observados y P^{∞} la distribución del modelo (distribución de equilibrio). En lugar de minimizar esta función, Brown y Hinton [68] proponen minimizar la divergencia contrastiva [89] (CD-*Contrastive Divergence*) para cancelar el último término de la ecuación 3.13:

$$CD=KL\big(P^0(X)||P^{\infty}(X|\Theta)\big)-KL\big(P^1(X|\Theta)||P^{\infty}(X|\Theta)\big) \tag{3.15}$$

con P^1 la distribución obtenida en una iteración hacia P^{∞}, la cual se reconstruye desde los datos empíricos aplicando el muestreo de Gibbs (muestreo desde $P(S|X,\theta)$ y $P(X|S,\theta)$

Algoritmo de entrenamiento en PoHMM

$\Theta = \{\theta_k^{(m)}\}_{m=1}^{M} \longleftarrow \forall m$ (algoritmo figura 13)
while $\mathcal{L}(\Theta)$ no converja
  $\Theta' = \Theta$
  for $\forall$ muestra $X(i)$
    introducir evidencia $X(i)$
    calcular las prob. marginales $P(S_t|X(i), \theta'^{(m)})$, $P(S_t, S_{t-1}|X(i), \theta'^{(m)})$ y $\mathcal{L}_i$ (*algoritmo FB*)
    $\mathcal{L} = \mathcal{L} + \mathcal{L}_i$
    //se calcula el primer término de la CD
    $\frac{\partial \log P(X(i)|\theta_k'^{(m)})}{\partial \theta_k'^{(m)}}$ (ecuación 3.20)
    //se ejecuta un paso del muestreo de Gibbs ($P(X|S)$ y $P(S|X)$)
    $\{S_{1:T}\}_{m=1}^{M} \longleftarrow P(S_t|X(i), \theta'^{(m)})$ (*algoritmo de Viterbi*)
    $X^1(i) \longleftarrow P(X|S, \Theta')$, con $P(X|S, \Theta')$ calculada desde las ecuaciones 3.21 y 3.22
    //se calcula el segundo término de la CD
    $\frac{\partial \log P(X^1(i)|\theta_k'^{(m)})}{\partial \theta_k'^{(m)}}$ (ecuación 3.20)
    //se calcula CD
    $\Delta\theta_k^{(m)}(i) = CD$ (ecuación 3.19)
    $\Delta\theta_k^{(m)} = \Delta\theta_k^{(m)} + \Delta\theta_k^{(m)}(i)$
  endfor
  $\Delta\theta_k^{(m)} = \Delta\theta_k^{(m)}/N$ (N es el número de muestras)
  $\theta_k^{(m)} = \theta_k'^{(m)} - \Delta\theta_k^{(m)}$
endwhile

Figura 15: Algoritmo de entrenamiento en PoHMM.

alternativamente) una sola vez. Dado que P^1 está más cerca que P^0 de la distribución del equilibrio P^∞, $KL(P^0||P^\infty)$ es mayor que $KL(P^1||P^\infty)$ por lo que la CD es siempre positiva.

Derivando la ecuación 3.14 respecto a los parámetros del PoHMM, $\Theta = \{\theta_m\}$, obtenemos:

$$\frac{\partial KL\big(P^0(X)||P^\infty(X|\Theta)\big)}{\partial \theta^{(m)}} = -\left\langle \frac{\partial}{\partial \theta^{(m)}} \log P^\infty(X|\Theta) \right\rangle_{P^0(X)} \tag{3.16}$$

$$\begin{aligned}\frac{\partial KL\big(P^1(X|\Theta)||P^\infty(X|\Theta)\big)}{\partial \theta^{(m)}} = &-\left\langle \frac{\partial}{\partial \theta^{(m)}} \log P^\infty(X|\Theta) \right\rangle_{P^1(X|\Theta)} \\ &+ \frac{\partial P^1(X|\Theta)}{\partial \theta^{(m)}} \frac{\partial KL\big(P^1(X|\Theta)||P(X|\Theta)^\infty\big)}{\partial P^1(X|\Theta)}\end{aligned} \tag{3.17}$$

El último término de la ecuación 3.17 se debe a que $P^1(X|\Theta)$ depende de los parámetros $\theta^{(m)}$, en la práctica su valor es muy pequeño por lo que este término puede despreciarse [87]. Entonces:

$$\begin{aligned}&-\frac{\partial}{\partial \theta^{(m)}} KL\big(P^0(X)||P^\infty(X|\Theta) - KL\big(P^1(X|\Theta)||P^\infty(X|\Theta)\big) \propto \\ &\left\langle \frac{\partial}{\partial \theta^{(m)}} \log P^\infty(X|\Theta) \right\rangle_{P^0(X)} - \left\langle \frac{\partial}{\partial \theta^{(m)}} \log P^\infty(X|\Theta) \right\rangle_{P^1(X|\Theta)}\end{aligned} \tag{3.18}$$

Promediando la ecuación 3.13 sobre las distribuciones $P^0(X)$ y $P^1(X|\Theta)$ respectivamente y sustituyendo en 3.18 obtenemos la siguiente fórmula para la actualización de los parámetros del PoHMM:

$$\Delta\theta^{(m)} \propto \frac{\partial}{\partial \theta^{(m)}} \log P\big(X^0_{1:T}|\theta^{(m)}\big) - \frac{\partial}{\partial \theta^{(m)}} \log P\big(X^1_{1:T}|\theta^{(m)}\big) \tag{3.19}$$

El gradiente $\frac{\partial \log P}{\partial \theta^{(m)}}$ se calcula usando la distribución posterior sobre las variables ocultas $P(S_{1:T}|X_{1:T}, \theta^{(m)})$:

$$\frac{\partial \log P(X_{1:T}|\theta^{(m)})}{\partial \theta^{(m)}} = \left\langle \frac{\partial \log P(X_{1:T}, S_{1:T}|\theta^{(m)})}{\partial \theta^{(m)}} \right\rangle_{P(S_{1:T}, X_{1:T}|\theta^{(m)})}$$

(3.20)

donde $P(S_{1:T}|X_{1,T}, \theta^{(m)})$ son calculadas con el algoritmo FB (sección 2.5.1).

En definitiva, nosotros partimos de una estimación inicial de los parámetros $\{\theta^{(m)}\}$ entrenando cada HMM independientemente y a continuación los actualizamos en el contexto de los PoHMM usando la ecuación 3.19. X_0 es una secuencia de observación tomada de los datos experimentales y X_1 es una secuencia 'reconstruida': dada la distribución posterior de variables ocultas $P(S_{1:T}|X_{1:T}, \theta^{(m)})$ para cada HMM hallamos la secuencia de estados ocultos más probable para cada una de las cadenas de Markov $\{s_{1:T}^{(1)}, \ldots, s_{1:T}^{(M)}\}$ mediante el algoritmo de Viterbi (sección 2.5.2); X_1 es entonces muestreada desde la distribución $P(X_{1:T}|S_{1:T})$, que en nuestro caso se trata de una combinación de gaussianas condicionada sobre $\{s_{1:T}^{(1)}, \ldots, s_{1:T}^{(M)}\}$ con medias y covarianzas:

$$\Sigma^{(g)}_{X_t|\{s_t^{(1)},\ldots,s_t^{(M)}\}} = \left[\sum_{m=1}^{M} \left(\Sigma^{(m)\ (g)}_{s_t^{(m)}}\right)^{-1}\right]^{-1} \tag{3.21}$$

$$\mu^{(g)}_{X_t|\{s_t^{(1)},\ldots,s_t^{(M)}\}} = \Sigma^{(g)}_{X_t|\{s_t^{(1)},\ldots,s_t^{(M)}\}} \left[\sum_{m=1}^{M} \left(\Sigma^{(m)\ (g)}_{s_t^{(m)}}\right)^{-1} \mu^{(m)\ (g)}_{s_t^{(m)}}\right] \tag{3.22}$$

3.7 FUNCIÓN DE PARTICIÓN PARA POHMM

Las aristas definen una función potencial sobre las variables (nodos) conectadas. Estas funciones potenciales no están normalizadas por lo que es necesario dividirlas por una constante de normalización de forma que la probabilidad total del modelo se encuentre entre 0 y 1. Cuando el modelo es simple, como en el caso de los grafos lineales de primer orden, esta constante de normalización o función de partición Z puede ser calculada con coste de computación razonable usando programación dinámica, pero conforme

el grafo está más densamente conectado o la longitud en la secuencia es mayor (más combinaciones posibles de los valores que toman las variables aleatorias) su cálculo se vuelve intratable. La estimación de Z es entonces un problema difícil sólo resuelto para grafos sencillos mediante técnicas complejas [90]. Generalmente, se recurre a una estimación de la razón de Z, Z_2/Z_1, empleando la regresión sobre los datos pertenecientes a dos clases distintas [88] para comparar los dos modelos correspondientes a estas clases. En esta sección describimos este método y presentamos una nueva propuesta para obtener una estimación no relativa de Z en el caso de los PoHMM mediante un muestreo de importancia por enfriamiento (AIS - *Annealed Importance Sampling*) [91].

3.7.1 *Estimación relativa de Z mediante regresión logística*

Para clasificar una secuencia bajo dos PoHMM correspondientes a dos clases distintas de datos (ecuación 3.23) es necesario conocer la función de partición Z sumando sobre todas las posibles secuencias de observación (ecuación 3.10), lo que es computacionalmente prohibitivo.

$$\log P(X_{1:T}|\Theta) = \sum_{m=1}^{M} \log P^{(m)}(X_{1:T}|\theta^{(m)}) - \log Z(\Theta) \quad (3.23)$$

El primer término a la derecha es la verosimilitud logarítmica no normalizada, $\log P^*(X|\Theta) = \sum_{m=1}^{M} \log P^{(m)}(X_{1:T}|\theta^{(m)})$, la cual se calcula sumando la probabilidad de cada HMM en el producto (calculadas independientemente dados los parámetros del modelo $\Theta = \{\theta^{(m)}\}$). Una vez obtenido el $\log P^*(X|\Theta)$ de los dos PoHMM, la diferencia entre sus $\log Z$ ($\Delta \log Z$) puede estimarse considerando ésta como una desviación de la diferencia entre las verosimilitudes logarítmicas de ambos PoHMM:

$$\Delta \log P(X|\Theta) = \Delta \log P^*(X|\Theta) - \Delta \log Z(\Theta) \quad (3.24)$$

$\Delta \log Z(\Theta)$ se calcula mediante una simple regresión logística [89] sobre los datos de entrenamiento:

$$p(y = 1|x) = \frac{1}{1 + e^{-(ax+b)}} \quad (3.25)$$

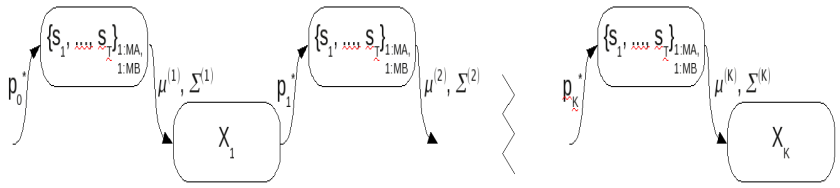

Figura 16: **Reconstrucción de X_k mediante muestreo de Gibbs.** Se muestrea los estados ocultos S usando $p(S|X)$ y a continuación X por $p(X|S)$ para cada distribución p^*_k.

con $x = \log P^*(X|\Theta)$ y $b = \Delta \log Z$.

3.7.2 *Estimación relativa y absoluta de Z mediante AIS*

Para calcular la verosimilitud de un modelo (ecuación 3.23) de forma absoluta es necesario una estimación no relativa de Z, lo que adquiere más importancia en tareas de clasificación multi-clase, ya que se puede evaluar los distintos modelos en un sólo paso, lo que permite manejar bases de datos con mayor número de acciones.

En [92], Salakhutdinov y Murray proponen un método basado en AIS [91] para hallar una estimación de la función de partición Z en maquinas restringidas de Boltzmann (RBM - *Restricted Boltzmann Machine*). Nosotros nos basamos en un esquema similar para estimar la función de partición en los PoHMM.

Dados dos PoHMM, siendo P^*_A y P^*_B las probabilidades no-normalizadas, y Z_A y Z_B sus respectivas funciones de partición. Definimos una secuencia de distribuciones intermedias entre P^*_A y P^*_B que difieren ligeramente entre sí, $\{p^*_0, \ldots, p^*_K\}$, las cuales satisfacen $p^*_k \neq 0$ si $p^*_{k+1} \neq 0$:

$$p^*_k(X) \propto P^*_A(X)^{(1-\beta_k)} P^*_B(X)^{\beta_k} = \left(\prod_{m_A=1}^{M_A} P_A^{*(m_A)} \right)^{(1-\beta_k)} \left(\prod_{m_B=1}^{M_B} P_B^{*(m_B)} \right)^{\beta_k} \tag{3.26}$$

siendo $P_I^{*(m_I)}$ la probabilidad debida a la cadena m_I del PoHMM I. Es decir, partimos de la distribución de un PoHMM y evolucionamos hacia la distribución del otro PoHMM ($p_0^* = P_A^*$ and $p_K^* = P_B^*$). La distribución de β que define la secuencia de distribuciones p_k^* ($\beta_0 = 0 < \beta_1 < \ldots < \beta_{K-1} < \beta_K = 1$) se fija empíricamente ya que su número y espaciado depende de cada problema en particular, en [91] afirman que el esquema óptimo es un espaciado uniforme en $\log \beta_j$.

En cada paso k reconstruimos una muestra $X_k = \{X_{1_k}, \ldots, X_{T_k}\}$ (con T_k la longitud temporal de cada muestra) desde la distribución anterior p_{k-1}^*, para lo cual muestreamos alternativamente $P(S|X)$ y $P(X|S)$ como se indica en la figura 16. Para generar X_k, en primer lugar tomamos X_{k-1} y calculamos usando p_{k-1}^* mediante el algoritmo de Viterbi la secuencia de estados ocultos más probable S para cada instante de tiempo t en todas las cadenas de los dos PoHMM: $S = \{s_t^{(1)}, \ldots, s_t^{(M_A)}, s_t^{(1)} \ldots, s_t^{(M_B)}\}$; posteriormente, la muestra aleatoria X_k se extrae de la mezcla de gaussianas condicionada en esta particular configuración de estados ocultos. Cada PoHMM contribuye en una determinada proporción (que se va desplazando en cada paso k) a la varianza, media y peso de cada gaussiana:

$$\Sigma_{\{s_t^{(1)},\ldots,s_t^{(M_A)},s_t^{(1)},\ldots,s_t^{(M_B)}\}} = \left[(1-\beta_k)\sum_{m_A=1}^{M_A}\left(\Sigma_{s_t^{(m_A)}}^{(m_A)}\right)^{-1} + \beta_k \sum_{m_B=1}^{M_B}\left(\Sigma_{s_t^{(m_B)}}^{(m_B)}\right)^{-1}\right]^{-1} \tag{3.27}$$

$$\mu_{\{s_t^{(1)},\ldots,s_t^{(M_A)},s_t^{(1)},\ldots,s_t^{(M_B)}\}} = \Sigma_{\{s_t^{(1)},\ldots,s_t^{(M_A)},s_t^{(1)},\ldots,s_t^{(M_B)}\}} \times \left[(1-\beta_k)\sum_{m_A=1}^{M_A}\left(\Sigma_{s_t^{(m_A)}}^{(m_A)}\right)^{-1}\mu_{s_t^{(m_A)}}^{(m_A)} + \beta_k \sum_{m_B=1}^{M_B}\left(\Sigma_{s_t^{(m_B)}}^{(m_B)}\right)^{-1}\mu_{s_t^{(m_B)}}^{(m_B)}\right] \tag{3.28}$$

$$C_{\{s_t^{(1)},\ldots,s_t^{(M_A)},s_t^{(1)},\ldots,s_t^{(M_B)}\}} = (1-\beta_k)\sum_{m_A=1}^{M_A} C_{s_t^{(m_A)}}^{(m_A)} + \beta_k \sum_{m_B=1}^{M_B} C_{s_t^{(m_B)}}^{(m_B)} \tag{3.29}$$

El peso de importancia de la muestra i^{th} es definido como:

$$w^{(i)} = \frac{p_1^*(X_1^{(i)})}{p_0^*(X_1^{(i)})} \quad \frac{p_2^*(X_2^{(i)})}{p_1^*(X_2^{(i)})} \cdots\cdots \frac{p_{K-1}^*(X_{K-1}^{(i)})}{p_{K-2}^*(X_{K-1}^{(i)})} \quad \frac{p_K^*(X_K^{(i)})}{p_{K-1}^*(X_K^{(i)})} \tag{3.30}$$

El promedio de los pesos para todas las muestras de entrenamiento nos da la razón entre Z_B y Z_A [91].

$$\frac{Z_B}{Z_A} \approx \frac{1}{N}\sum_{i=1}^{N} w^{(i)} \tag{3.31}$$

Si tomamos como P_A un simple modelo con Z_A conocida podemos estimar directamente Z_B de la ecuación (3.31). Nosotros vamos a tomar como P_A un PoHMM con una sola cadena, es decir, un HMM clásico, con P_A igual a la ecuación (3.1) y $Z_A = 1$, la estimación absoluta de Z_B es dada entonces por $Z_B \approx \frac{1}{N}\sum_{i=1}^{N} w^{(i)}$.

La estimación de Z mediante este método tiene un mayor coste computacional que la estimación de la razón de Z por el método de regresión logística descrito en la sección 3.7.1, no obstante, permite una mayor eficiencia en la etapa de reconocimiento, ya que para una base de datos con M acciones para reconocer a qué clase pertenece una muestra hay que realizar una sola comparación sobre las distribuciones de cada modelo frente a las $M-1$ comparaciones necesarias con razones de Z. Además, cuando el número de acciones M en la base de datos es grande, el coste en la estimación absoluta de Z se ve recompensado debido a que para el método de la regresión es necesario calcular $M!/2!(M-2)!$ razones de Z frente a las M estimaciones de Z con AIS.

Algoritmo de estimación de Z basado en AIS

// Se calcula la probabilidades no normalizadas para ambos PoHMM (P^*_A y P^*_B numerador de la ecuación 3.9)
for $\forall$ muestra $X(i)$:
  for $\forall$ muestra β_k:
    introducir evidencia $X(i)$ en P^*_A y P^*_B
    calcular las prob. marginales $P(S_t|X(i), \theta^{(m)})$, $P(S_t, S_{t-1}|X(i), \theta^{(m)})$ para ambos (*algoritmo FB*)

    //se calcula X_{k+1} desde p^*_k (ecuación 3.26)
    calculamos $\{S^{(m_A)}_{1:T}\} \longleftarrow P^*_A(S|X, \theta^{(m_A)})$ y $\{S^{(m_B)}_{1:T}\} \longleftarrow P^*_B(S|X, \theta^{(m_B)})$ (*algoritmo de Viterbi*)
    $X_{k+1}(i) \longleftarrow P(X|S, \Theta)$ (ecuaciones 3.27, 3.28 y 3.29)

    // calcular $\mathcal{L}$ de $X_{k+1}(i)$ para la distribución $p^*_k = (1-\beta_k)P^*_A + \beta_k P^*_B$
    introducir evidencia $X_{k+1}(i)$ en P^*_A y P^*_B
    se calcula $\mathcal{L}_{\mathcal{A}i}$ y $\mathcal{L}_{\mathcal{B}i}$ (*algoritmo FB*)
    $\mathcal{L}_i = (1-\beta_k)\mathcal{L}_{\mathcal{A}i} + \beta_k \mathcal{L}_{\mathcal{B}i}$
    $\log w(i) = \log w(i) - \mathcal{L}_i$

    // calcular $\mathcal{L}$ de $X_{k+1}(i)$ para la distribución $p^*_{k+1} = (1-\beta_{k+1})P^*_A + \beta_{k+1}P^*_B$
    introducir evidencia $X_{k+1}(i)$ en P^*_A y P^*_B
    se calcula $\mathcal{L}_{\mathcal{A}i}$ y $\mathcal{L}_{\mathcal{B}i}$ (*algoritmo FB*)
    $\mathcal{L}_i = (1-\beta_{k+1})\mathcal{L}_{\mathcal{A}i} + \beta_{k+1}\mathcal{L}_{\mathcal{B}i}$
    $\log w(i) = \log w(i) + \mathcal{L}_i$
  endfor
endfor
$\frac{Z_B}{Z_A} \approx \sum_{i=1}^{N} w(i)$
if P^*_A es un HMM $(Z_A = 1) \Rightarrow Z_B \approx \sum_{i=1}^{N} w(i)$

Figura 17: Algoritmo de entrenamiento en PoHMM.

3.8 DISCUSIÓN

Los HMM llevan siendo extensamente explotados en el campo de HAR desde hace más de una década, debido a

que proporcionan un herramienta sencilla y eficiente para modelar las acciones humanas gracias a su habilidad en manejar información espacio-temporal con alto grado de aleatoriedad. El aspecto dinámico de la acción es modelado estadísticamente por la matriz de transición entre estados, A, que absorbe la variabilidad temporal en la ejecución de una acción concreta, y la matriz de probabilidad de salida, B, modela la variabilidad en el espacio de características.

No obstante, toda la información sobre la evolución del proceso está contenida en una única variable multinomial discreta. Así que, un HMM con pocos estados capturará ineficientemente la estructura de acciones formadas por muchos movimientos o poses distintas, sin embargo, aumentar el número de estados K supone una mayor necesidad de datos de entrenamiento para que el HMM funcione correctamente y un mayor coste computacional.

Los FHMM, al usar una representación distribuida de estado, pueden manejar más información que los HMM clásicos ($M \log K$ vs. $\log K$). Por otro lado, posibilitan que los procesos de dinámica independiente sean modelados con una cadena distinta separadamente. Por ejemplo, el movimiento de la parte inferior del cuerpo humano (piernas, caderas) por una parte y el de la superior (tronco, brazos, cabeza) por otra. Lo que permite modelar acciones de movimientos más complejos que los HMM e incluso acciones formadas por la intervención de dos o más personas.

Debido a la suposición independencia en la evolución de cada cadena, los FHMM se comportan bien cuando la señal a modelar puede descomponerse en componentes pocos correlacionadas, en caso contrario Ghahramani y Jordan [65] afirman que los modelos factoriales no ofrecen ventajas frente a los HMM simples. Por tanto, la elección de características para alimentar a los FHMM es un asunto fundamental, sin embargo, la extracción de características independientes en una secuencia de imágenes no siempre es una tarea sencilla.

En los PaHMM, cada cadena de Markov modela una secuencia de observación distinta consideradas totalmente independientes. Por ejemplo, podemos dividir el vector de ob-

servación en secuencias correspondientes a distintas partes del cuerpo humano y conectar cada secuencia a una cadena de Markov para aprender su evolución. Aunque el coste computacional es entonces proporcional al de un HMM ($O(TMK^2)$), a menudo, esta simplificación es demasiado restrictiva en las acciones humanas. Además, para aquellas acciones donde una parte del cuerpo realiza movimientos similares, como puede ser el caso de los brazos en las acciones de correr o boxear, conduce a una mayor confusión en la clasificación.

A diferencia de los FHMM y los PaHMM, los CHMM no exigen poca o nula correlación entre las señales a modelar por cada cadena de Markov. Por lo que son capaces de modelar procesos que interactúan entre sí, es decir, cada proceso evoluciona según su propia dinámica interna pero ésta es influenciada por las del resto. Si como en el caso de los PaHMM cada secuencia corresponde a una parte distinta del cuerpo, los CHMM permiten codificar coordinación en los movimientos, es decir, cada parte del cuerpo se mueve independientemente pero se coordinan entre sí para formar un movimiento determinado. Por ejemplo, cuando las personas caminan balancean sus brazos para conservar el equilibrio, de forma que el estado de los brazos en cada instante depende del estado de los brazos y de las piernas en los instantes anteriores.

Los CHMM son también adecuados para combinar información de señales de diferente naturaleza, por ejemplo características audiovisuales, ya que permiten la progresión asíncrona de las distintas secuencias de observación a la vez que conservan su correlación en el tiempo.

Calcular exactamente la distribución posterior sobre las variables de estado en los FHMM y los CHMM es costoso, debido a que las variables observadas inducen dependencias entre las variables de estado (ver capítulo 2). Es necesario, por tanto, considerar todas las combinaciones posibles en los estados de las variables ocultas para evaluar su probabilidad posterior ($O(TMK^{M+1})$), por lo que para más de dos cadenas el coste computacional de la inferencia es prohibitivo.

En los PoHMM, sin embargo, puede incorporarse tantas cadenas como sea necesario para modelar adecuadamente nuestro problema a coste lineal ($O(TMK^2)$). A diferencia de los FHMM donde cada variable compite para explicar los datos, en los PoHMM todas las cadenas de Markov son responsables de describir los datos, por lo que en el proceso de aprendizaje cada cadena aprende un aspecto distinto de los datos. Por tanto, los PoHMM no sólo son capaces de modelar las dinámicas de distintas partes del cuerpo como lo otros modelos multi-cadena, sino que son capaces de aprender la estructura presente en los vectores de característica en múltiples escalas temporales. Muchas acciones exhiben posturas comunes así que analizar la información contenida a distinta escala temporal puede ser muy útil para discernir entre estas acciones (por ejemplo entre correr y trotar).

4

CAMPOS ALEATORIOS DEL MARKOV

En los modelos gráficos descritos en el capítulo anterior existe una relación de precedencia entre las variables de estado. En este capítulo abordamos el problema de HAR empleando dos modelos gráficos distintos en los que la relación entre las variables de estado es de tipo correlacional (simétrica): El primero de ellos es un caso particular de los MRF, los campos aleatorios condicionales, que se caracterizan directamente por su probabilidad condicional sobre los datos observados, se trata por tanto de un modelo discriminativo. Y el segundo es un modelo mixto denominado MRF generativo, que nosotros adaptamos a manejar señales temporales, y que a diferencia del MRF discriminativo pretende dar cuenta de los procesos que generan las observaciones.

4.1 CAMPOS ALEATORIOS CONDICIONALES: CRF

Los campos aleatorios condicionales [93] (CRF - *Conditional Random Fields*) son un caso particular de los UGM que representan directamente distribuciones de probabilidad condicional $P(Y|X)$ sobre las variables multinomiales Y dadas las observaciones X. A diferencia de la probabilidad conjunta $P(X, Y)$ no requieren modelar explícitamente las observaciones, $P(X)$.

La probabilidad conjunta se factoriza según la ecuación 2.5 como:

$$P(Y|X) = \frac{\prod_C \Psi_C(X_C, Y_C)}{Z(X)}, \quad Z = Z(X) = \sum_{Y \in \mathcal{Y}} \prod_C \Psi_C(X_C, Y_C) \tag{4.1}$$

donde X_C y Y_C son las variables observadas y las variables multinomiales a predecir en el clique C respectivamente. El conjunto de factores en el grafo $\{\Psi_C\}$ se describen usualmente mediante combinaciones log-lineales de las funciones

de características f que contienen la información relevante extraída de los datos:

$$\Psi_C(X_C, Y_C) = exp\Big\{ \sum_k \theta_{C_k} f_{C_k}(X_C, Y_C) \Big\} \tag{4.2}$$

Estas funciones pueden comprender valores entre menos infinito e infinito aunque generalmente toman valores binarios $\{0, 1\}$ [94]. θ son los pesos asociados a cada función de características y definen la importancia de una particular función f respecto al resto en el potencial. Estos pesos constituyen los parámetros del modelo que son aprendidos durante el entrenamiento minimizando la log-verosimilitud condicional $\log P(Y|X)$ de los datos etiquetados.

En nuestra tarea de HAR la variable Y representa la etiqueta asociada a cada acción y puede tomar los valores discretos $Y \in \{1, \ldots, Y\}$ donde Y es el número de acciones a comparar; la variable X tiene el mismo significado que en los modelos anteriores, el vector de características extraído en cada fotograma. Nosotros usamos CRF con una estructura lineal, similar a los HMM, representando las variables Y en cada instante de tiempo (fotogramas). Suponemos conexiones de primer orden entre las variables Y, es decir, cliques formados por pares de variables vecinas (Y_{t-1}, Y_t), dado que los nodos cercanos en el tiempo tendrán intuitivamente más influencia entre sí, además, incrementar la conectividad entre las variables Y incrementa la complejidad de la inferencia notablemente. No obstante, en contraste con los modelos descritos en el capítulo 3 las variables Y pueden depender de las observaciones X en cualquier instante de tiempo si estas son conocidas y fijadas. La figura 18 muestra un esquema de este modelo, la estructura del CRF se extiende en el tiempo con tantos nodos de las variables Y y X como fotogramas en cada secuencia. Ambas variables son conocidas durante el entrenamiento, pero las variables discretas Y que asocian una acción concreta a cada fotograma han de ser inferidas durante el reconocimiento. A diferencia de los modelos descritos en el capítulo anterior no existe una capa de variables ocultas que el modelo tenga que aprender. Sustituyendo la ecuación 4.2 en 4.1 para este tipo de CRF obtenemos:

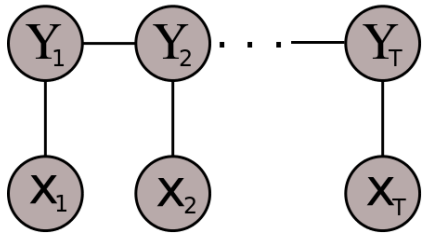

Figura 18: **CRF de estructura lineal.** Y son las variables discretas etiquetando la acción representada en los fotogramas y X las variables de observación. Los nodos sombreados indican que los valores para ambas variables son conocidos durante el entrenamiento.

$$\begin{aligned} P(Y_{1:T}|X_{1:T}) &= \frac{1}{Z(X)} exp\left\{ \sum_{t=1}^{T} \sum_{k=1} \theta_k f_k(Y_{t-1}, Y_t, X) \right\} \\ &= \frac{1}{Z(X)} exp\left\{ \sum_{t=1}^{T} \sum_{k=1} \theta_k f_k(Y_{t-1}, Y_t, X_{W_t}) \right\} \end{aligned} \tag{4.3}$$

donde por razones de eficiencia hemos restringido las observaciones consideradas para predecir el valor de Y en el instante t, Y_t, a las observaciones que ocurren en la ventana temporal centrada en t de tamaño $2W+1$ $(t-W : t+W)$ bajo la suposición que las relaciones entre observaciones vecinas son más fuertes.

La función de partición $Z(X)$ que normaliza esta probabilidad depende de las observaciones y se define como suma sobre todas las configuraciones de Y (ecuación 4.1), por tanto, el coste computacional de Z se incrementa exponencialmente con el tamaño de las acciones a considerar, Y, y la longitud T de las secuencias, ésta puede ser calculada por programación dinámica [71].

En la ecuación 4.3 hemos supuesto que los pesos θ_k no dependen del instante de tiempo (están fijos a través del tiempo) lo que permite manejar secuencias de distinta longitud. Como funciones de características usamos una función f_i asociada a la etiqueta de la acción que codifica las relaciones entre las etiquetas y los vectores de observación

(clique Y_t-X_t), y una función f_{ij} asociada a las etiquetas entre fotogramas consecutivos que captura la consistencia en la etiquetas y la transición entre acciones (clique Y_{t-1}-Y_t):

$$P(Y_{1:T}|X_{1_T}) = \frac{1}{Z(X)} exp\left\{ \sum_{t=1}^{T} \sum_{i,j} \theta_{ij} f_{ij}(Y_t, Y_{t-1}) + \sum_{t=1}^{T} \sum_{i} \theta_i f_i(Y_t, X_{W_t}) \right\} \tag{4.4}$$

Comúnmente las funciones f_{ij} y f_i son definidas como deltas de Dirac [50, 53]:

$$f_i = \delta_i(Y_t) X_{W_t} \begin{cases} X_W & \text{si } Y_t = i, \\ 0 & \text{en otro caso} \end{cases} \tag{4.5}$$

$$f_{ij} = \delta_{ij}(Y_{t-1}, Y_t) \begin{cases} 1 & \text{si } Y_{t-1} = i \text{ y } Y_{t-1} = j \\ 0 & \text{en otro caso} \end{cases} \tag{4.6}$$

La primera función tiene asociados $Y \times d \times (2W+1)$ pesos θ_{ij}, y la segunda $Y \times Y$ pesos θ_i, donde Y es el número de etiquetas (acciones a comparar), d es el tamaño del vector de características y $2W+1$ el tamaño de la ventana de observación (número de fotogramas).

4.2 CAMPOS ALEATORIOS CONDICIONALES OCULTOS

En el modelo descrito en la sección anterior los valores de todas las variables que intervienen son conocidos en la fase de entrenamiento, se dice por tanto que el modelo es completamente observable. En esta sección describimos brevemente los campos ocultos condicionales (HCRF - *Hidden Conditional Random Fields*) [95], a los que se ha añadido una capa de variables ocultas $\{S_1, \ldots, S_T\}$ condicionadas sobre las observaciones para aprender la estructura subyacente en el fenómeno modelado.

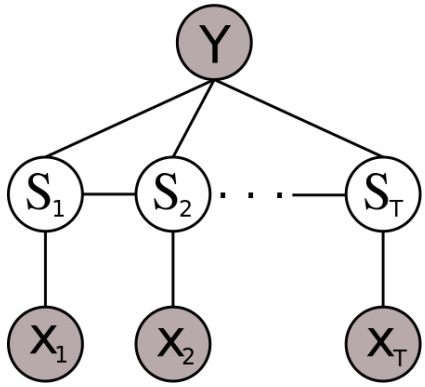

Figura 19: **HCRF.** La variable Y representa la clase (etiqueta), las variables S_t y X_t son las variables de estado y las de observación respectivamente para cada instante de tiempo. Los nodos no sombreados indican que las variables S son ocultas.

La probabilidad condicional $P(Y|X)$ se calcula marginando en S la probabilidad conjunta de las variables Y y S dadas las observaciones X:

$$P(Y|X) = \sum_{S} P(Y, S|X) \tag{4.7}$$

factorizando $P(Y, S|X)$ como producto de potenciales de clique $\Psi_C(X_C, Y_C, S_C)$ tenemos:

$$P(Y|X) = \frac{\sum_S \prod_C \Psi_C(X_C, Y_C, S_C)}{Z(X)} \tag{4.8}$$

$$Z(X) = \sum_{Y,S} \prod_{C} \Psi(X_C, Y_C, S_C) \tag{4.9}$$

Y_C y X_C tiene el mismo significado que en la sección 4.1, S_C son las variables de estado ocultas contenidas en el clique C. Ahora la ecuación 4.3 toma la forma:

$$P(Y, S_{1:T}|X_{1:T}) = \frac{1}{Z(X)} exp\Bigg\{ \sum_{t=1}^{T} \sum_{i} \theta_i f_i(S_t, X_{W_t}) + \sum_{t=1}^{T} \sum_{i,j} \theta_{ij} f_{ij}(Y, S_t) + \sum_{t=1}^{T} \sum_{i,j,k} \theta_{ijk} f_{ijk}(Y, S_{t-1}, S_t) \Bigg\} \tag{4.10}$$

Al igual que en el caso de los CRF usaremos como funciones de características f_i, f_{ij} y f_{ijk} funciones de Dirac:

$$f_i = \delta_i(S_t) X_{W_t} \begin{cases} X_W & \text{si } S_t = i, \\ 0 & \text{en otro caso} \end{cases} \tag{4.11}$$

$$f_{ij} = \delta_{ij}(Y, S_t) \begin{cases} 1 & \text{si } Y = i \text{ y } S_t = j \\ 0 & \text{en otro caso} \end{cases} \tag{4.12}$$

$$f_{ijk} = \delta_{ijk}(Y, S_{t-1}, S_t) \begin{cases} 1 & \text{si } Y = i,\ S_{t-1} = j \text{ y } S_t = k \\ 0 & \text{en otro caso} \end{cases} \tag{4.13}$$

Los parámetros del HCRF son entonces $\Theta = \{\theta_i, \theta_{ij}, \theta_{ijk}\}$ de tamaño $d(2W+1) \times K$, $Y \times K$ y $Y \times K \times K$ respectivamente, con Y el número de acciones a comparar (etiquetas), K el número de estados ocultos en cada acción, d es la dimensión del vector de características para la observación en un instante t y W la ventana temporal de observaciones (desde el fotograma $t - W$ al $t + W$). La matriz θ_i asociada a la función de características f_i pesa la correlación entre el valor de variable de estado oculta en el instante t, S_t, y las observaciones encontradas en la ventana temporal W centrada en ese instante; θ_{ij} asociado a la función de

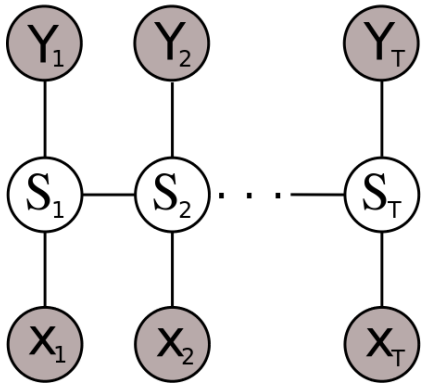

Figura 20: **LDCRF.** Y_t es la variable de etiqueta de clase, S_t es la variable de estado y X_t es la variable de observación en el instante t respectivamente. Los nodos sombreados representan variables observables y los nodos con fondo blanco variables ocultas.

características f_{ij} pesa la correlación entre un valor de la variable S en el instante t y una acción determinada (valor de Y); por último, θ_{ijk} pesa la relación entre una transición concreta entre estados ocultos, valores de S_{t-1} y S_t, y la acción dada por Y.

4.3 CRF DE DINÁMICA LATENTE

Los campos aleatorios condicionales de dinámica latente [54] (LDCRF - *Latent-Dynamic Conditional Random Fields*) fueron introducidos como una mejora a los HCRF, descritos en la sección anterior. Los LDCRF combinan los CRF y los HCRF para explotar las ventajas de ambos, mientras los CRF son capaces de aprender la dinámica extrínseca en el paso de una clase a otra, los HCRF pueden modelar la estructura intrínseca de cada clase. Este modelo consiste en una capa de variables multinomiales de etiquetas de clase $Y_{1:T} = \{Y_1, \ldots, Y_T\}$, una capa de variables ocultas discretas $S_{1:T} = \{S_1, \ldots, S_T\}$, y la capa de las observaciones $X_{1:T} = \{X_1, \ldots, X_T\}$ (figura 20).

La probabilidad condicional será:

$$\begin{aligned} P(Y_{1:T}|X_{1:T}) &= \sum_{S} P(Y_{1:T}, S_{1:T}|X_{1:T}) \\ &= \sum_{S} P(Y_{1:T}|S_{1:T}, X_{1:T}) P(S_{1:T}|X_{1:T}) \end{aligned} \tag{4.14}$$

Por cuestiones de eficiencia, se dividen los estados ocultos asociados a cada clase Y en conjuntos disjuntos $\mathcal{S}_Y$ tal que el conjunto de todos los posibles estados ocultos sea su unión $\mathcal{S} = \sum_Y \mathcal{S}_Y$ [54] . Entonces, la ecuación 4.14 quedaría como:

$$P(Y_{1:T}|X_{1:T}) = \sum_{S:\forall s_Y \in \mathcal{S}_y} P(S_{1:T}|X_{1:T}) \tag{4.15}$$

ya que $P(Y_{1:T}|S_{1:T}, X_{1:T}) = 1$ para $\forall s_Y \in \mathcal{S}_Y$ y 0 en cualquier otro caso. La ecuación 4.10 se simplifica a:

$$\begin{aligned} P(S_{1:T}|X_{1:T}) = \frac{1}{Z(X)} exp\Bigg\{ &\sum_{t=1}^{T} \sum_{i} \theta_i f_i(S_t, X_{W_t}) \\ &+ \sum_{t=1}^{T} \sum_{i,j} \theta_{ij} f_{ij}(S_{t-1}, S_t) \Bigg\} \end{aligned} \tag{4.16}$$

Ahora θ_{ij} (matriz $K \times K$ siendo K el número total de estados ocultos) aprende tanto la dinámica extrínseca como la intrínseca, si los elementos de la matriz están asociados a estados ocultos del mismo subconjunto ($s_i, s_j \in \mathcal{S}_{Y_I}$) modelan la estructura interna de esta clase, en cambio si están asociados a estados ocultos pertenecientes a distintas clases ($s_i \in \mathcal{S}_{Y_I}$, $s_j \in \mathcal{S}_{Y_J}$ y $\mathcal{S}_{Y_I} \neq \mathcal{S}_{Y_J}$) modela la dinámica de transición entre ambas clases. θ_i tiene el mismo significado que en HCRF.

Así, los LDCRF no sólo son capaces de discriminar la estructura interna de la acción a través de las variables ocultas como en el caso de los HCRF sino también de aprender del tránsito de una acción a otra, lo que puede ser muy útil en determinadas ocasiones, por ejemplo una persona que está corriendo no se para en seco cuando se aproxima a su destino sino que primero desciende el ritmo a un ligero trote y luego continúa caminando, encadenando las acciones correr, trotar, caminar. Además, como cada fotograma

Algoritmo de entrenamiento en CRF

$\Theta = \{\theta_k\}$ iniciales
Se construye el grafo JT desde el grafo original
while $\mathcal{L}(\Theta)$ no converja
    Cálculo del gradiente
    $\Theta' = \Theta$
    for $\forall$ muestra $X(i)$
        introducir evidencia $X(i)$ en el JT
        calcular las prob. marginales $P(Y_t|X(i),\Theta')$, $P(Y_t, Y_{t-1}|X(i),\Theta')$ y $Z(X(i))$ (*algoritmo HUGIN sobre JT*)
        $\mathcal{L} = \mathcal{L} + \mathcal{L}_i$
    endfor
    $\theta_k \longleftarrow \frac{\partial \mathcal{L}}{\partial \theta_k}$ (ecuaciones 4.20 y 4.21)
endwhile

Figura 21: Algoritmo de entrenamiento en CRF.

lleva asociado una etiqueta indicando la acción a que corresponde podemos entrenar directamente sin necesidad de segmentar las secuencias en acciones individuales en contraste con los HCRF, así como segmentar y etiquetar una secuencia compuesta por varias acciones simultáneamente al igual que los CRF.

4.4 APRENDIZAJE EN LOS CRF

Para estimar los parámetros Θ de los campos aleatorios condicionales se maximiza el logaritmo de la verosimilitud condicional:

$$\mathcal{L}(\Theta) = \sum_{m=1}^{N} \log P\left(Y^{(m)}_{1:T^{(m)}} | X^{(m)}_{1:T^{(m)}}, \theta\right) - \frac{1}{2\sigma^2}||\Theta||^2 \quad (4.17)$$

siendo N el número de secuencias de entrenamiento y donde los pares $\{(X^{(m)}_{1:T^{(m)}}, Y^{(m)}_{1:T^{(m)}})\}_1^N$ para cada secuencia m de longitud $T^{(m)}$ son conocidos en su totalidad en el aprendizaje. El segundo término de esta ecuación es un

término de regularización que representa el logaritmo de una probabilidad a priori gaussiana de media cero y varianza σ^2, que suaviza $\mathcal{L}(\Theta)$ cuando existe disparidad en los datos de entrenamiento [96]. En nuestro caso, $\mathcal{L}(\Theta)$ será (ver ecuación 4.3):

$$\mathcal{L}(\Theta) = \sum_{m=1}^{N} \left(\sum_{t=1}^{T} \sum_{i,j} \theta_{ij} f_{ij}(Y_t^{(m)}, Y_{t-1}^{(m)}) + \sum_{t=1}^{T} \sum_{i} \theta_i f_i(Y_t^{(m)}, X_{W_t}^{(m)}) - \log Z(X^{(m)}) \right) - \frac{\|\Theta\|^2}{2\sigma^2} \tag{4.18}$$

Derivando con respecto a θ_k:

$$\frac{\partial \mathcal{L}(\Theta)}{\partial \theta_k} = \sum_{m=1}^{N} \left(\sum_{t=1}^{T} f_k(Y_t^{(m)}, X_{W_t}^{(m)}) - \frac{1}{Z(X^{(m)})} \frac{\partial Z(X^{(m)})}{\partial \theta_k} \right) - \frac{\theta_k}{\sigma^2} \tag{4.19}$$

sustituyendo la forma de Z en el segundo término y desarrollando matemáticamente se obtiene:

$$\frac{\partial \mathcal{L}(\Theta)}{\partial \theta_k} = \sum_{m=1}^{N} \left(\sum_{t=1}^{T} f_k(Y_t^{(m)}, X_{W_t}^{(m)}) - \sum_{t=1}^{T} \left\langle f_k(Y_t^{(m)}, X_{W_t}^{(m)}) \right\rangle_{P(Y_t | X_{W_t}^{(m)}, \Theta)} \right) - \frac{\theta_k}{\sigma^2} \tag{4.20}$$

Al igual para θ_{kl}:

$$\frac{\partial \mathcal{L}(\Theta)}{\partial \theta_{kl}} = \sum_{m=1}^{N} \left(\sum_{t=1}^{T} f_{kl}(Y_t^{(m)}, Y_{t-1}^{(m)}) - \sum_{t=1}^{T} \left\langle f_{kl}(Y_t^{(m)}, Y_{t-1}^{(m)}) \right\rangle_{P(Y_t, Y_{t-1} | X_{W_t}^{(m)}, \Theta)} \right) - \frac{\theta_{kl}}{\sigma^2} \tag{4.21}$$

En definitiva, se trata de minimizar la diferencia entre los valores empíricos y el valor esperado dado por la distribución del modelo.

Las probabilidades marginales, $P(Y_t|X_{W_t},\Theta)$ y $P(Y_t,Y_{t-1}|X_{W_t},\Theta)$, necesarias para hallar los gradientes $\partial\mathcal{L}/\partial\theta_k$ y $\partial\mathcal{L}/\partial\theta_{kl}$ respectivamente, pueden calcularse mediante el algoritmo JT (sección 2.5.3).

Al ser $\mathcal{L}(\Theta)$ una función con forma $\log\left(\sum_i e^{X_i}\right)$ es convexa. La propiedad de convexidad es extremadamente útil en la estimación de parámetros cuando minimizamos esta función con respecto a dichos parámetros, puesto que asegura que el óptimo local sea también un óptimo global. El óptimo global puede ser entonces alcanzado usando algoritmos numéricos de gradiente, como el gradiente conjugado [97] cuyas direcciones de búsqueda son ortogonales entre sí.

Sin embargo, en los HCRF y los LDCRF al incorporar variables ocultas no puede garantizarse que el óptimo local alcanzado sea un óptimo global. Sea la verosimilitud logarítmica:

$$\begin{aligned}\mathcal{L}(\Theta) &= \sum_{m=1}^{N} \log P\left(Y^{(m)}_{1:T^{(m)}}|X^{(m)}_{1:T^{(m)}},\theta\right)\\ &= \sum_{m=1}^{N} \log \sum_{S} P\left(S^{(m)}_{1:T^{(m)}},Y^{(m)}_{1:T^{(m)}}|X^{(m)}_{1:T^{(m)}}\right)\end{aligned} \tag{4.22}$$

dado que los valores de Y_t son conocidos durante el entrenamiento podemos introducir sus valores reales en la probabilidad y maximizar directamente $P(S_{1:T}|Y_{1:T},X_{1:T},\Theta)$.

En el caso del LDCRF, sustituyendo la ecuación 4.16 en 4.22 y derivando respecto a los parámetros, tras varios desarrollos matemáticos [66] se obtiene:

$$\begin{aligned}\frac{\partial\mathcal{L}(\Theta)}{\partial\theta_k} = \sum_{m=1}^{N}\Bigg(&\sum_{t=1}^{T}\left\langle f_k(S^{(m)}_t,X^{(m)}_{W_t})\right\rangle_{P(S_t|Y^{(m)}_t,X^{(m)}_{W_t},\Theta)}\\ &-\sum_{t=1}^{T}\left\langle f_k(S^{(m)}_t,X^{(m)}_{W_t})\right\rangle_{P(S_t,Y_t|X^{(m)}_{W_t},\Theta)}\Bigg)\end{aligned} \tag{4.23}$$

$$\frac{\partial \mathcal{L}(\Theta)}{\partial \theta_{kl}} = \sum_{m=1}^{N} \left(\sum_{t=1}^{T} \left\langle f_{kl}(S_t^{(m)}, S_{t-1}^{(m)}) \right\rangle_{P(S_{t-1}, S_t | Y_t^{(m)}, X_{W_t}^{(m)}, \Theta)} - \sum_{t=1}^{T} \left\langle f_{kl}(S_t^{(m)}, S_{t-1}^{(m)}) \right\rangle_{P(S_{t-1}, S_t, Y_t | X_{W_t}^{(m)}, \Theta)} \right) \tag{4.24}$$

Las probabilidades marginales, $P(S_t|Y_t, X_{W_t}, \Theta)$, $P(S_t, Y_t|X_{W_t}, \Theta)$, $P(S_{t-1}, S_t|Y_t, X_{W_t}, \Theta)$ y $P(S_{t-1}, S_t, Y_t|X_{W_t}, \Theta)$, usadas para calcular los valores esperados pueden calcularse mediante el algoritmo JT ($O(TK^2)$), donde K es el número de estados del modelo y T el número de nodos de estado (fotogramas). No obstante, el algoritmo de inferencia se invoca repetidamente en cada cálculo del gradiente para cada muestra de entrenamiento en el aprendizaje. Por lo que, el coste total es $O(TK^2NG)$, con N el número de muestras de entrenamiento y G el número de cálculos del gradiente requerido para la optimización. Lo que puede ser computacionalmente muy costoso dependiendo del número de muestras de entrenamiento que necesitemos y en el caso de los HCRF Y LDCRF también del número de estados que puedan tomar las variables *S*.

La técnica que usamos para calcular el gradiente es el gradiente conjugado ya que tolera mejor las violaciones de la convexidad en los CRF con variables ocultas [47].

4.5 CAMPOS ALEATORIOS DE MARKOV GENERATIVOS

Hinton *et al.* [67] introdujo un nuevo modelo, el MRF con salida causal o MRF generativo, para aprender las relaciones entre las distintas partes de un objeto (por ejemplo, caras, dígitos, etc.) en imágenes estáticas. En contraste con los campos aleatorios condicionales (modelos discriminativos), en este modelo las variables ocultas mantienen una relación de tipo correlacional entre ellas (influencia simétrica) pero sus relaciones con las observaciones son causales (modelo generativo). Nosotros adaptamos este modelo a señales que

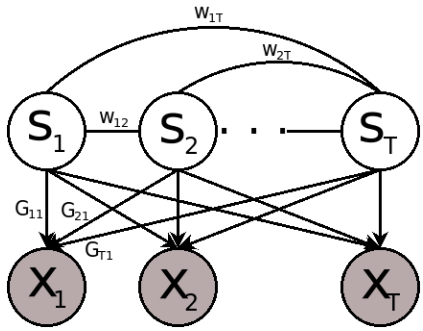

Figura 22: **MRF generativo.** X representan las variables de observación (nodos sombreados), S las variables ocultas de estado, los arcos representan relaciones causales y las aristas representan influencias simétricas entre las variables.

varían con el tiempo de manera que pueda aplicarse al problema de HAR.

4.5.1 *Fundamentos*

Los MRF generativos son especificados en función de las probabilidades $P(S)$ y $P(X|S)$ cuyas distribuciones son definidas mediante GM no dirigidos y dirigidos respectivamente [67], siendo S las variables de estado ocultas discretas y X las variables de observación continuas.

La probabilidad sobre las variables ocultas S (capa MRF) se especifica en términos de energía usando una distribución de Boltzmann:

$$P(S) = \frac{e^{-E(S)}}{Z}, \qquad Z = \sum_{s \in \mathcal{S}} e^{-E(S)} \tag{4.25}$$

Valores de energía altos se asocian a configuraciones de S que violan las suposiciones del modelo. La función de partición Z se calcula sumando sobre todas las posibles configuraciones de S, $\mathcal{S}$.

La energía de la capa MRF, $E(S)$, se modela con una función cuadrática compuesta por la energía de interacción entre nodos, $E(S_i, S_j) = s_i s_j w_{ij}$, y la energía de nodo, $E(S_i) = s_i b_i$, donde s_k es el estado del nodo S_k que toma los valores binarios 0 ó 1, $w_{ij} = w_{ji}$ es el peso de intera-

cción entre los nodos S_i y S_j ($w_{ii} = 0$) y b_i representa el desplazamiento respecto a la energía debida al nodo S_i:

$$E(S) = -\frac{1}{2}S^T W S - b^T S \tag{4.26}$$

con $b = \{b_j\}_{j=1}^J$ donde J es el número de variables (nodos) S_j y W es la matriz simétrica $W = \{w_{ij}\}_{i,j=1}^J$ de tamaño $J \times J$ que indica la correlación entre cada nodo S_j.

En Hinton *et al.* [67] suponen que las observaciones X dados los estados de las variables S son generadas conforme a una distribución gaussiana:

$$P(X|S) = \mathcal{N}(X; GS + m, \Sigma) \tag{4.27}$$

donde Σ y m representan la varianza y la media de los datos observados (de dimensión $d \times d$ y d) respectivamente, siendo d el número de componentes de los vectores de observación. La matriz G representa la contribución de cada variable oculta S_j al valor de la variable de observación X_i.

Nosotros modificamos la probabilidad $P(X \mid S)$ definida en la ecuación 4.27 de forma que el modelo pueda aplicarse a señales que evolucionan en el tiempo, como es el problema que nos ocupa. La nueva probabilidad $P(X \mid S)$ consistirá en un producto de gaussianas, cada una de las cuales modela independientemente las observaciones emitidas en un instante de tiempo. Al igual que en los modelos descritos anteriormente, las variables ocultas S_j representan el estado asociado a cada fotograma y X_i las características observables extraídas de cada uno, por tanto tenemos tantos nodos S_j y X_i como número de fotogramas T. En la figura 22 se muestra un esquema de este modelo, la probabilidad conjunta se factoriza como:

$$\begin{aligned} P(S, X|\Theta) &= P(S|\Theta)P(X|S, \Theta) \\ &= \frac{e^{\frac{1}{2}S^T W S + b}}{Z} \prod_{i=1}^{T} \frac{e^{-\frac{1}{2}(X_i - m - G_i S)^T \Sigma^{-1}(X_i - m - G_i S))}}{\sqrt{(2\pi)^d |\Sigma|}} \end{aligned} \tag{4.28}$$

la matriz G_i indica cuánto contribuye cada variable S_j al valor que toma la variable d-dimensional X_i, su tamaño

es $d \times J$, $G = \{G_i\}_{i=1}^{I}$ con I el número de variables de observación X. Nosotros consideraremos el caso más general en el que puede existir correlación entre las variables de estado en cualquier instante, es decir, entre los estados que toman cualquier par de fotogramas, y en el que el estado de cada fotograma contribuye a generar las observaciones en cualquier instante de tiempo, es decir W $(J \times J)$ y G $(d \times J \times I)$ son matrices completas.

4.5.2 *Aprendizaje*

Calcular la distribución posterior del modelo sobre las variables ocultas es intratable, en [67] se aproxima la probabilidad real $P(S|X)$ por una distribución factorial paramétrica más sencilla $Q(S|X)$, donde los parámetros del modelo son hallados maximizando un límite inferior en la verosimilitud logarítmica sobre Q. La distribución factorial se representa como:

$$Q(S|X) = \prod_j q_j^{s_j}(1 - q_j)^{(1-s_j)} \tag{4.29}$$

siendo q_j la probabilidad del estado s_j de la variable S_j en la configuración oculta $\mathcal{S}$, con $s_j \in \{0, 1\}$ (modelo con 2 estados ocultos). Las probabilidades q_j son dadas por la función sigmoide:

$$q_j = \frac{1}{1 + e^{-\sum_i R_{ji} X_i + c_j}} \tag{4.30}$$

donde R_{ij} es un vector de tamaño d que pesa la contribución de la variable d-dimensional X_i a que la variable S_j se encuentre en el estado s_j, y c_j representa el offset. Los parámetros $R = \{R_{ij}\}_{i,j=1}^{I,J}$ (matriz de tamaño $I \times d \times J$) y $c = \{c_j\}_{j=1}^{J}$ son aprendidos en la etapa de entrenamiento.

Neal y Hinton [98] demuestran que:

$$-\log P(X|\Theta) = \mathcal{F}(X|\Theta) - KL(Q(S|X,\Theta))||P(S|X,\Theta)) \tag{4.31}$$

donde $\mathcal{F}$ es la formulación variacional de la energía libre: $\mathcal{F}_Q = < E >_Q - H(Q)$. Minimizar $\mathcal{F}$ es equivalente a maximizar un límite inferior en la verosimilitud logarítmica ya que

Algoritmo de entrenamiento en un MRF generativo

1 $\Theta = \{\theta_k\}$ iniciales
2 **while** $\mathcal{L}(\Theta)$ no converja
3 $\quad\Theta' = \Theta$
4 $\quad$**for** $\forall$ muestra $X(i)$
5 $\quad\quad$se calcula $q_j^{(0)}(X_i)$ (ecuación 4.30)
6 $\quad\quad\{S_{1:T}\} \longleftarrow q_j^{(0)}(X_i)$
7 $\quad\quad$se calcula $\mathcal{L}_i(\Theta')$ tomando logaritmo en la ecuación 4.28
8 $\quad\quad\mathcal{L} = \mathcal{L} + \mathcal{L}_i$

9 $\quad\quad$//Para W y b
10 $\quad\quad$se calcula $q_j^{(k)}(X_i)$ (ecuación 4.35)

11 $\quad\quad$//Para el resto de parámetros
12 $\quad\quad$se calcula el gradiente $\frac{\partial \mathcal{F}(\times')}{\partial \theta_k}$
13 $\quad\quad\Delta\theta_k = \Delta\theta_k + \Delta\theta_k(i)$
14 $\quad$**endfor**
15 $\quad$//Se actualizan W y b
16 $\quad\Delta w_{ij} \propto \sum_{i=1}^{N} q_i^{(0)} q_j^{(0)} - q_i^{(k)} q_j^{(k)}$ y $\Delta b_i \propto \sum_{i=1}^{N} q_i^{(0)} - q_i^{(k)}$
17 $\quad$//Para los demás
18 $\quad\Delta\theta_k = \Delta\theta_k/N$ (N es el número de muestras)

19 $\quad\Theta = \Theta' - \Delta\Theta$
20 **endwhile**

Figura 23: Algoritmo de entrenamiento en un MRF generativo.

la divergencia Kullback-Leibler (KL) es siempre positiva. $\mathcal{F}$ es dada por:

$$\mathcal{F}(x) = \sum_S Q(S|X) \log Q(S|X) - \sum_S Q(S|X) \log P(S, X) \tag{4.32}$$

sustituyendo $P(S,X)$ y $Q(S|X)$ por la forma dada en las ecuaciones 4.29 y 4.28 respectivamente tenemos:

$$\begin{aligned}\mathcal{F} = &\sum_i [q_i \log q_i + (1-q_i)\log(1-q_i)] \\ &- \frac{1}{2} q^T W q - b^T q + \log Z \\ &+ \frac{1}{2}\sum_{i=1}^{T} (q^T G_i^T G_i q - 2(x_i - m)^t G_i q) + ctes\end{aligned} \tag{4.33}$$

cuya minimización nos permite obtener los parámetros del modelo $\Theta = \{(W,b),(G,m),(R,c)\}$. Los parámetros (G, m) y (R, c) pueden ser directamente ajustados calculando el gradiente exacto de la ecuación 4.33, no así W y b, ya que la función de partición Z (ecuación 4.25) depende de ambos. En este caso, se minimiza la divergencia contrastiva [89] (CD):

$$CD = \mathcal{F}_0 - \mathcal{F}_k \tag{4.34}$$

con k el número de iteraciones hacia la distribución de equilibrio. Dados los parámetros actuales de la capa MRF (W y b) y las probabilidades en la iteración $k-1$ ($\{q_j^{(k-1)}\}_{j=1}^{J}$) las probabilidades q_j en la iteración k son dadas por [67]:

$$q_j^{(k)} = \lambda q_j^{(k-1)} + \frac{1-\lambda}{1 + e^{-b_j - \sum_{m=1}^{J} q_m^{(k-1)} w_{ij}}} \tag{4.35}$$

donde λ es un coeficiente de amortiguación entre 0 y 1 usado para prevenir oscilaciones. La CD pretende forzar a que la distribución del modelo se ajuste bien a los datos experimentales siendo suficiente unas pocas iteraciones k. Minimizando la ecuación 4.34 con respecto a W y b se obtiene que dichos parámetros se actualizan como [99]:

$$\Delta w_{ij} \propto \sum_N q_i^{(0)} q_j^{(0)} - q_i^{(k)} q_j^{(k)} \qquad \Delta b_i \propto \sum_N q_i^{(0)} - q_i^{(k)} \tag{4.36}$$

siendo N el número de datos de entrenamiento. Este procedimiento se repite hasta que el valor de los parámetros del modelo Θ convergen.

Notar que en los MRF generativos el tamaño de los parámetros depende del número de nodos J de las variables ocultas S y el número de nodos I de las variables de observación X (en nuestro caso $J = I = T$), por tanto dependen de la longitud de la secuencia que registra la acción. La duración en la ejecución de una misma acción dada su naturaleza presenta una gran variabilidad, dividimos entonces cada secuencia en sub-secuencias más pequeñas de longitud fija T tomando para ello una ventana de este tamaño (correspondiente a la secuencia más corta en la base de datos) deslizándola a lo largo de la secuencia con un desplazamiento pequeño, es decir, con alto solapamiento.

4.6 DISCUSIÓN

En los modelos descritos en el capítulo 2.2 se realizan fuertes suposiciones de independencia sobre las observaciones: que son independientes entre sí y que sólo dependen del estado actual del modelo, las cuales son muy restrictivas y en ocasiones no se ajustan a la realidad. Las acciones están formadas por una combinación de pequeños movimientos, muchos de ellos comunes a diferentes acciones, así que el contexto y dependencias de largo término necesitan ser consideradas para una correcta clasificación. Además, estas suposiciones fuerzan a que las observaciones no sean compartidas entre estados, a cada observación se le asocia un estado en la etapa de entrenamiento por lo que cualquier error inicial en dicha asignación da lugar a una mala estimación de los parámetros del modelo [84].

En los CRF al modelar directamente la distribución condicional basada en las observaciones no es necesario realizar suposiciones de independencia sobre ellas. Cada nodo de estado puede compartir observaciones con otros nodos y considerar observaciones en distintos instantes de tiempo. Los CRF son por tanto capaces de incorporar restricciones contextuales, por ejemplo cuando caminamos un patrón de piernas abiertas es siempre precedido de otro de piernas semi-abiertas; además pueden alimentarse con características más ricas donde las observaciones estén relacionadas

entre sí (solapadas en el espacio o en el tiempo). Por ello, las funciones de características pueden ser diseñadas para explotar nuestro conocimiento del problema específico y extraer la mejor información posible de la secuencia de imágenes, lo que otorga a estos modelos su gran flexibilidad y potencial. La variedad de características con las que podemos alimentar estos modelos es entonces considerable, en [94] McCallun describe como inferir un subconjunto de características que incremente la verosimilitud condicional del modelo.

Por otra parte, los CRF y los LDCRF son capaces de segmentar y etiquetar simultáneamente una secuencia compuesta por distintas acciones de forma natural. Los LDCRF, al igual que los modelos de Markov, además explotan la estructura subyacente en la acción gracias a su capa de variables de estado ocultas.

Sin embargo, su principal desventaja es que el coste computacional asociado al entrenamiento es mayor que en los HMM, ya que la función de partición Z depende de las observaciones, por lo que será necesario calcularla para cada muestra. Para el caso de CRF lineales (primer orden de Markov) este coste es aceptable usando programación dinámica [71]. Lafferty [47] también señala que se podría esperar que los CRF sean estimados con menos muestras de entrenamiento que los modelos generativos, ya que las características no necesitan especificar completamente las observaciones.

Los campos condicionales aleatorios son modelos discriminativos, aunque en general no existe una clara ventaja de este tipo de modelos frente los generativos, sino que depende del problema concreto [100, 101], los modelos discriminativos están muy ligados a las muestras de entrenamiento y por tanto tienen menos capacidad de generalización.

Los MRF generativos no sufren esta limitación. Además, al estar formados por arcos y aristas permite que las técnicas de inferencia aproximada trabajen más eficientemente que si las conexiones fueran de un solo tipo [67]. Podemos entonces conectar densamente el grafo sin que el coste computacional sea prohibitivo, lo que nos permite explotar las

relaciones de consistencia entre los valores que toman las variables de estado a corto, medio y largo plazo. Muchas acciones presentan movimientos repetitivos (por ejemplo caminar, boxear, etc.) en los que los estados de los fotogramas en distintos instantes de tiempo estarán relacionados.

Este modelo también nos ayuda a entender mejor el fenómeno estudiado, los valores de G y W nos indican que conexiones tienen mayor fuerza y cuales pueden considerarse despreciables, dándonos el mapa de la influencia real entre las variables (nodos) implicadas sobre el que podemos hacer suposiciones de independencia realistas, no demasiado restrictivas.

Parte II

Experimentación

5

EXTRACCIÓN DE CARACTERÍSTICAS

5.1 BASE DE DATOS

Nuestros experimentos fueron ejecutados en la base de datos KTH [102], que es la base de secuencias de acciones pública de mayor tamaño por lo que es considerada la base estándar en el campo de HAR, y nos va a permitir cotejar nuestros resultados con los de otras técnicas aplicadas también a esta base de datos. Este corpus supone un interesante desafío al estar formadas las secuencias por imágenes en gris (no hay información de color) de baja resolución (160×120 píxeles) a 25 fps, un VCD tiene 352×288 y un DVD 720×576 píxeles; además, sufren de artificios de compresión y poco contraste, y algunas secuencias son también afectadas por sombras y tenue iluminación.

En la base KTH, 25 personas de ambos sexos ejecutan 10 acciones: *andar*, *trotar*, *correr* y *boxear* hacia la derecha y hacia la izquierda en el plano de la cámara, y *aplaudir* y *alzar los brazos* de cara a la cámara. Estas acciones se realizaron en cuatro condiciones distintas: el escenario 1 es un escenario exterior, donde se producen cambios en la iluminación (intensidad y dirección) y en ocasiones sombras en los pies de la persona dependiendo donde esté situado el sol, con un campo de hierba como fondo; el escenario 2 es el mismo entorno que el escenario 1 pero durante la grabación de las secuencias de *boxear*, *aplaudir* y *alzar los brazos* hubo un continuo zoom de la cámara (hacia adelante y hacia atrás) mientras que en el resto de acciones hay un fuerte cambio de vista de hasta 45° respecto al primer escenario, lo que provoca cambios en el tamaño relativo del sujeto a lo largo de toda la secuencia que registra la acción; el escenario 3 se sitúa en la misma ubicación que el escenario 1 y 2, pero ahora los sujetos visten ropas holgadas y portan otros objetos como mochilas, bufandas, etc.; por último, el escenario 4 se ubica en un recinto interior, con una pared

Figura 24: **Ejemplos de la base de datos KTH.** Imágenes de la base de datos KTH para las acciones *andar* (primera columna), *trotar* (segunda), *correr* (tercera), *boxear* (cuarta), *aplaudir* (quinta) y *alzar los brazos* (sexta columna) en el escenario 1 (primera fila), escenario 2 (segunda), escenario 3 (tercera) y escenario 4 (cuarta fila).

blanca de fondo donde se proyecta la sombra de la persona que realiza la acción. En la figura 24 se muestra algunas secuencias de acciones de la base de datos KTH realizadas por el mismo sujeto para los cuatro escenarios.

5.2 DETECCIÓN DEL SUJETO

La detección del sujeto que realiza la acción es en sí misma un problema no trivial que queda fuera del ámbito de este trabajo. La mayoría de las técnicas constan de dos pasos: 1) *Segmentación* de la zona de interés (ROI), siendo las técnicas más usadas: la sustracción de fondo, la estadística de píxeles, el flujo óptico y la diferenciación temporal; 2) *Clasificación*, que utiliza métodos clásicos basados en el color de la piel, la silueta del sujeto, el análisis del movimiento, etc.

Dada la naturaleza de las secuencias de la base de datos KTH en las que el único objeto en la imagen es la persona que ejecuta la acción y el fondo es casi uniforme, utilizaremos un método estadístico sencillo basado en el algoritmo

CAMSHIFT [103] (*Continuosly Adaptative Mean Shift*) sin paso de clasificación. Técnicas basadas en el movimiento fallarían en detectar las partes del cuerpo inmóviles (por ejemplo las piernas en las acciones de *boxear*, *aplaudir* o *alzar los brazos*).

5.2.1 *Descripción del método*

En primer lugar, una distribución no paramétrica del fondo (histograma) es aprendida para cada escenario de los primeros fotogramas de un conjunto de secuencias de estos escenarios. El valor de cada píxel es entonces sustituido por su probabilidad de no pertenecer al fondo.

Las zonas con mayor densidad serán las posiciones más probables del sujeto, por lo que calculamos el momento cero (M_{00}) y el primer momento para la coordenada x e y (M_{10} y M_{01} respectivamente) sobre la imagen de probabilidad:

$$\begin{aligned} M_{00} &= \sum_x \sum_y I(x,y) \\ M_{10} &= \sum_x \sum_y xI(x,y) \\ M_{01} &= \sum_x \sum_y yI(x,y) \end{aligned} \tag{5.1}$$

siendo $I(x,y)$ el valor (probabilidad) del píxel en la posición (x,y). Dada su definición M_{00} contendrá información aproximada sobre el tamaño relativo del sujeto en la imagen y M_{01} y M_{10} sobre su posición, el pico de la distribución de probabilidad se sitúa en:

$$X_c = \frac{M_{10}}{M_{00}}, \quad Y_c = \frac{M_{01}}{M_{00}} \tag{5.2}$$

Para ajustar mejor la posición del centro de masas del sujeto y su tamaño relativo, se vuelve a calcular los momentos sobre una ventana reducida de tamaño M_{00} centrada en X_c e Y_c, este proceso se repite hasta que las coordenadas del centroide no varíen. La región donde se sitúa el sujeto se extrae de una ventana centrada en este punto con tamaño:

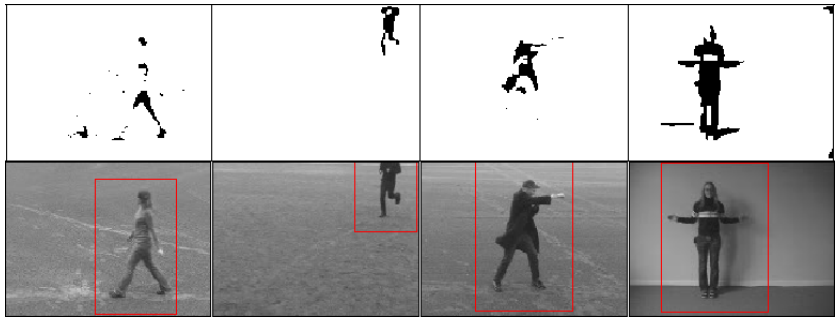

Figura 25: **Detección del sujeto.** La fila superior muestra las imágenes de probabilidad de que un píxel pertenezca al fondo (píxeles en negro corresponderían a la silueta de la persona). La fila inferior muestra el rectángulo que delimita el área conteniendo al sujeto que se usará en el reconocimiento de la acción.

$c_1 s \times c_2 s$, tal que $s = \frac{\sqrt{M_{00}}}{I_{max}}$ con I_{max} el valor máximo de píxel, y c_1 y c_2 dos constantes estimadas empíricamente que dependen de la distancia del sujeto a la cámara con la que fueron grabadas las secuencias en la base de datos. Finalmente, la región se escala a un tamaño fijo, igual para todas las secuencias.

5.2.2 *Resultados y limitaciones del método de detección*

La figura 25 presenta algunos ejemplos de los resultados de la extracción del sujeto para los cuatro escenarios con el método descrito. En la fila superior aparece la imagen de probabilidad, es decir, la probabilidad de que cada píxel pertenezca al fondo, en la figura inferior se muestra la región delimitando al sujeto de donde extraeremos las características que supondrá la entrada a los modelos evaluados.

En aquellas secuencias con presencia de sombras su distribución se integró al modelo de fondo para que no causaran un desplazamiento en la posición estimada del sujeto. Para otras bases de datos con escenarios más complejos, la distribución de aquellos elementos de la escena no pertenecientes a la persona, como por ejemplo vehículos, también puede añadirse al modelo de fondo, donde sería

además necesario una actualización periódica [104] para manejar la aparición de elementos nuevos o cambios en la iluminación a lo largo de la secuencia. La extensión a más de una persona podría atacarse inicializando varias ventanas de búsqueda (tantos como sujetos en la escena) con las posiciones y los tamaños estimados para cada sujeto e incorporando información temporal. Sin embargo, este método fallará cuando la distribución de intensidades del fondo sean similares a los de las ropas que vista la persona que queremos detectar. No obstante, existen detectores públicos que ofrecen buenos resultados en la detección de personas en entornos complejos, como el de Dalal y Triggs[1] [105] o Felzenszwalb*et al.*[2] [106], basados en histogramas 1D de gradientes de intensidad orientados. La biblioteca libre de visión artificial openCV integra el detector de caras de Viola y Jones [107] a partir de las cuales se puede extraer al individuo completo.

5.3 CARACTERÍSTICAS

La selección de las características que representen la información más relevante de la secuencia de imágenes es un importante cuestión. Nosotros usamos dos diferentes tipos de características: uno con información relativa a la forma del sujeto (basada en contornos) y el otro con información relativa a su movimiento (basada en flujo óptico), ambas son características muy comunes en visión por ordenador que pueden ser obtenidas en condiciones realistas sin muchas suposiciones a priori. Para aumentar la robustez en las características tomamos histogramas sobre ellas, ya que los histogramas han demostrado ser características robustas muy informativas [108, 109].

Planteamos entonces las acciones humanas desde dos puntos de vista, como una sucesión de posturas estáticas (características de contorno) o como la composición de pequeños movimientos (características de flujo óptico).

1 http://www.navneetdalal.com/software
2 http://people.cs.uchicago.edu/ pff/latent/

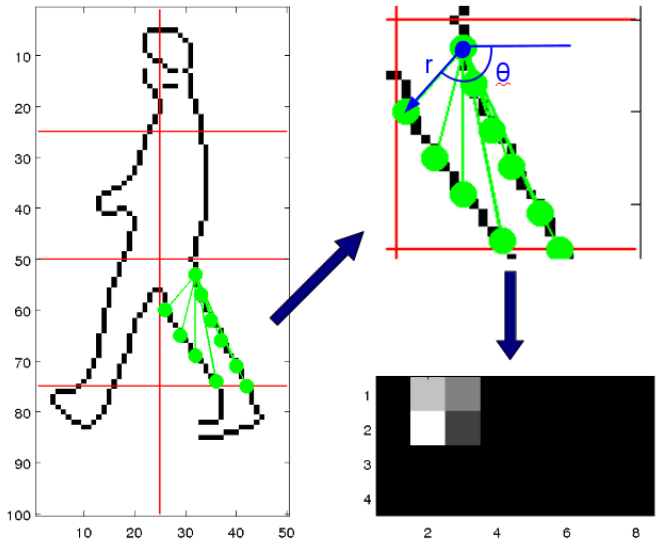

Figura 26: ***Shape-context.*** Para un punto de los 10 en que se ha muestreado el contorno.

5.3.1 *Contorno:* Shape-context

Podemos considerar cada acción humana como una sucesión temporal de determinadas poses de la persona que la ejecuta. La forma de la silueta humana ha sido ampliamente estudiada y explotada, el contorno del sujeto contiene información útil sobre ella y es una potente característica discriminante [110], asequible con baja resolución aún cuando no hay información de color o textura. Los contornos presentan como ventajas su semejanza a distintas escalas, son más robustos a pequeños cambios de iluminación y de punto de vista que por ejemplo el flujo óptico; y pueden extraerse rápidamente con técnicas sencillas como los detectores de borde. Sin embargo, son muy sensibles a escenas concurridas, a la presencia de sombras e incluso a las propias ropas del sujeto y a una mala segmentación del sujeto del fondo.

Belongie y Malik [111] introdujeron como descriptores del contorno los *shape-context* logrando buenos resultados en el reconocimiento de objetos [112]. Básicamente, son histogramas log-polares de las distancias y las direcciones de

un punto del contorno al resto de puntos del mismo contorno aleatoria y uniformemente muestreados (figura 26). Al estar basada en histogramas esta característica es razonablemente tolerante a variaciones en la pose (crucial en nuestro caso) y a los inconvenientes que sufren otras codificaciones basadas en contornos. El uso del logaritmo de la distancia otorga más peso a los puntos cercanos que a los lejanos, más probablemente afectados por la variabilidad del contorno. La invarianza a escala espacial se alcanza al normalizar todas las distancias radiales por la distancia media entre todos los pares de puntos de contorno. Nosotros referiremos los ángulos indicando la dirección respecto al eje X positivo, ya que la invarianza a rotación no es deseable en nuestro problema sino que deseamos esta información.

Los *shape-contexts* aunque recogen información de manera robusta sobre el contorno global del objeto relativa al punto de referencia no preservan la información espacial. En lugar de caracterizar el contorno global del sujeto completo nosotros caracterizamos el contorno de las distintas partes en que dividiremos el cuerpo humano. La región que contiene al sujeto es segmentada en regiones rectangulares no solapadas y se calcula los *shape-contexts* de los puntos del contorno que cae en cada región (ver figura 26). El promedio de estos *shape-contexts* de cada región son concatenados para formar nuestro vector de características. Al añadir esta información espacial se incrementa la potencia discriminante del vector de características (véase por ejemplo *andar* y *alzar los brazos*) lo que nos va a permitir distinguir el sentido de la acción (derecha o izquierda) y es también más robusto a oclusiones parciales.

La figura 27 muestra algunas imágenes originales de la base datos KTH, los contornos extraídos de ellas con el detector de bordes de Canny [113] y las características resultantes.

5.3.2 *Flujo óptico*

Las acciones humanas también pueden concebirse como una composición de pequeños movimientos. El campo de

movimiento aparente 2D también llamado flujo óptico se presenta como una característica muy intuitiva para caracterizar estos movimientos en ausencia de información 3D. El flujo óptico ya ha demostrado ser una característica muy útil en el campo del reconocimiento de acciones [114] y las técnicas que lo calculan han recibido mucha atención [115]. Los contornos pueden verse gravemente degradados en escenas con fondos no uniformes y por otro lado el flujo óptico permite resolver ambigüedades entre acciones con poses semejantes pero con diferente velocidad de realización (véase por ejemplo trotar y correr). Es sin embargo más ruidoso y costoso de calcular que los contornos.

Nosotros usaremos el flujo óptico obtenido con el algoritmo de Farnebäck [116] descrito brevemente en la siguiente sección, que proporciona una estimación de flujo óptico densa y estable con poca carga computacional [117]. Una vez hallado el flujo óptico para cada dos fotogramas consecutivos, promediamos este flujo en una ventana deslizante de tamaño 5 fotogramas para incrementar su estabilidad y calculamos histogramas 2D de la magnitud $m = \sqrt{u^2 + v^2}$ y la orientación $\theta = \arctan v/u$ del flujo (u, v). El uso de histogramas, los cuales son rápidos de extraer, pretende hacer que nuestras características sean menos sensibles a los ruidos propios del flujo óptico (punto de vista, iluminación, etc.).

Cuando se ejecuta una acción, el movimiento es principalmente gobernado por algunos de los miembros del cuerpo, por lo que al igual que en los contornos dividimos el área del sujeto en regiones no solapadas y caracterizaremos el movimiento global del sujeto caracterizando el movimiento que cae en cada región [118], lo que incrementa el potencial discriminativo del vector de características y posibilita la distinción entre una misma acción realizada en sentidos opuestos (véase sección 5.3.1). En la figura 28 se observa el flujo óptico y las características basadas en éste para algunas imágenes de la base de datos KTH.

Algoritmo de Farnëback

El método de Farnëback ajusta en cada píxel x un polinomio cuadrático para la vecindad de ese píxel:

$$f(x) = x^T A(x) x + b^T(x) x + c(x) \tag{5.3}$$

$f(x)$ representa el modelo local de la imagen expresado en un sistema de coordenadas local. La matriz simétrica $A(x)$, el vector $b(x)$ y el escalar $c(x)$ son los coeficientes de expansión polinomial locales, los cuales son estimados mediante una regresión de mínimos cuadrados pesada en los píxeles de la vecindad $\mathcal{V}$, de manera que tienen mayor peso los píxeles de la vecindad más cercanos al píxel central y éste disminuye conforme se alejan. Esto puede ser implementado eficientemente mediante un esquema jerárquico de convoluciones separables [117]. Sea d el vector de traslación del píxel x entre dos imágenes consecutivas $f_t(x)$ y $f_{t+1}(x)$:

$$f_t(x) = f_{t+1}(x - d) \tag{5.4}$$

sustituyendo esta igualdad por la ecuación 5.3 se obtiene el valor d de la restricción linear:

$$A(x) d(x) = \Delta b(x) \tag{5.5}$$

con $A = 1/2(A_t + A_{t+1})$ el promedio de ambas matrices y $\Delta b = -1/2(b_{t+1} - b_t)$. Suponiendo que el campo de desplazamiento $d(x)$ varía lentamente, se puede tomar un promedio sobre la vecindad de cada píxel para atenuar el error en la estimación de d:

$$d(x) = \left(\sum_{x \in \mathcal{V}} w A^T A \right)^{-1} \sum_{x \in \mathcal{V}} w A^T \Delta b \tag{5.6}$$

donde w es una función de peso para los píxeles en la vecindad. Esta restricción sobre cada píxel proporciona una estimación de flujo óptico densa y poco ruidosa comparada con otros algoritmos clásicos [119, 120, 121, 122]. El problema de la apertura se reduce en cada píxel a menos que toda la vecindad se vea afectada.

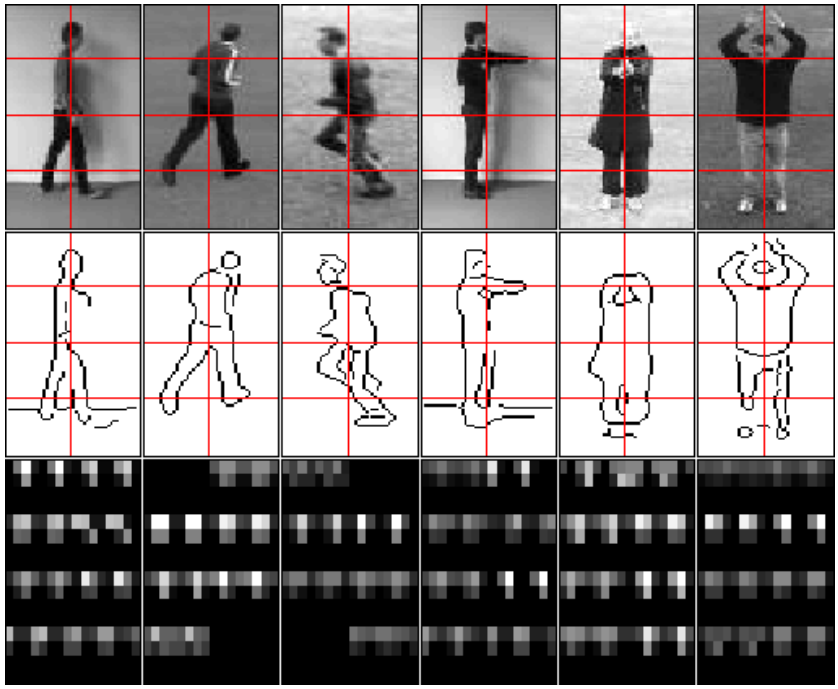

Figura 27: **Ejemplos de la codificación basada en contornos.** La primera fila muestra algunas imágenes originales de la base KTH, la segunda muestra los correspondientes contornos extraídos con Canny, y la tercera fila los *shape-context* medios en cada región en que dividimos el área conteniendo al sujeto.

5.3.3 *Vectores de características*

Los vectores de características están formados por la concatenación (de izquierda a derecha y de arriba a abajo) de las características extraídas en cada región en que dividimos el área del sujeto que realiza la acción. Nosotros consideraremos ocho primordiales regiones: dos secciones horizontales, explotando la simetría del ser humano; y cuatro secciones verticales, correspondientes aproximadamente a los segmentos básicos del cuerpo humano: cabeza-hombros, tronco-brazos, cadera-muslos y rodillas-piel. Regiones más pequeñas no capturarían los contornos o los movimientos característicos del ser humano. Un histograma 2D es calculado para cada región, con un eje para la magnitud m (relativo a la distancia en el *shape-context* o al flujo óptico)

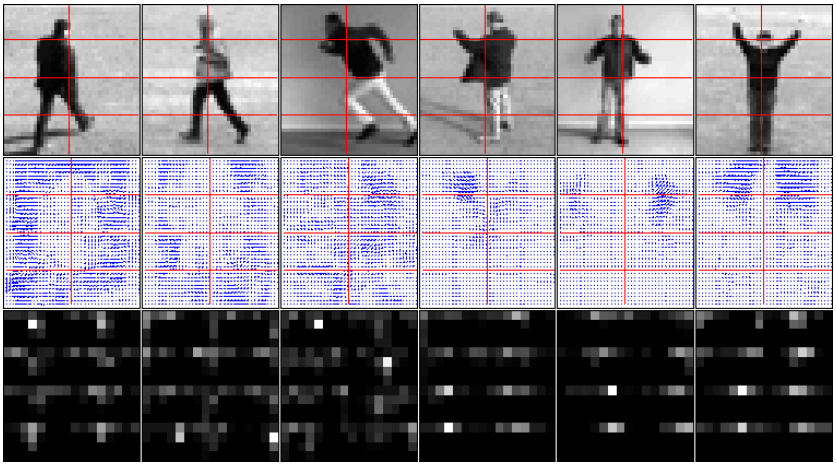

Figura 28: **Ejemplos de la codificación basada en flujo óptico.** La primera fila muestra algunas imágenes de la base KTH, la segunda muestra la estimación de flujo óptico con el algoritmo de Farnëback, y la tercera los histogramas del flujo para cada región en la que caracterizamos el movimiento.

con 4 bins y el otro eje para la orientación θ con 8 bins (de 0° a 360° en pasos de 45°). En total el vector de características tiene 256 componentes (4 bins para $m \times 8$ bins para $\theta \times 8$ regiones). Esta codificación se ilustra en las figuras 27 y 28.

Este tamaño del vector de características es bastante grande y por tanto el número de parámetros que los modelos han de aprender, lo que implica la necesidad de un mayor número de muestras de entrenamiento para que no haya sobreajuste y un mayor coste computacional. Para reducir el tamaño de este vector empleamos una de las técnicas más comunes para la compresión de la información, el análisis de componentes principales (PCA -Principal Component Analysis), en el que las variables son proyectadas en un nuevo sistema de coordenadas ortogonal, donde los ejes (autovectores) se ordenan por la información que contienen asociada a la varianza de las variables originales (autovalores). Nosotros aplicamos PCA sobre los datos de entrenamiento en las características extraídas en cada región en que dividimos el sujeto, de esta forma se mantiene la información asociada a cada parte del sujeto y se aumenta

la robustez frente a oclusiones parciales. Los datos de entrenamiento y de evaluación son proyectados en estos nuevos autovectores y tomamos sólo los autovectores con los cuatro mayores autovalores para cada región, que mantienen el 90% de la información total (para menos de cuatro la realización decae notablemente). La dimensión del vector de características se reduce de 256 a 32 componentes (8 regiones $\times 4$ autovalores).

6 ESTUDIO EXPERIMENTAL

Los modelos descritos aportan distintas soluciones al problema de HAR: introducción de información de contexto, modelado de procesos concurrentes con distinta dinámica, aprendizaje de la estructura de la acción en múltiples escalas temporales, etc. En este capítulo queremos evaluar cómo estas soluciones influyen en la tarea propuesta y cuales son las más relevantes con el objetivo de obtener el modelo más eficiente en términos de la tasa de aciertos en el reconocimiento y el coste computacional.

La bondad de estos modelos se medirá mediante la cuantificación del número de secuencias correctamente clasificadas por cada modelo. Denominamos como *clase* al grupo de secuencias que representa la misma acción. Los modelos son entrenados como clasificadores binarios. En el reconocimiento se sigue un esquema una-contra-todos (*one-vs-all*), el modelo cuyos parámetros obtienen la más alta probabilidad de observación para una secuencia asigna la etiqueta de acción a dicha secuencia.

Los resultados son ratificados usando un esquema de validación cruzada con 3 particiones: los datos son divididos aleatoriamente en tres partes de aproximadamente el mismo tamaño, una partición diferente cada vez se usa para el reconocimiento y las restantes para el entrenamiento, el resultado final es la media de los tres tests. Los tests son independientes de sujeto, es decir, los sujetos de la base KTH (sección 5.1) que participan en el entrenamiento no participan en el reconocimiento.

6.1 SOFTWARE UTILIZADO

Para los HMM y sus extensiones multi-cadena, FHMM, CHMM y HMM paralelos, se utilizó el paquete de código abierto *Bayes Net* [123] creado por Kevin Murphy en Matlab con algunas funciones en C. Con los CRF y los LDCRF

usamos la biblioteca *HCRF library* [124], desarrollada por el grupo *Vision Interface Group* del MIT en C++ con una interfaz mex para Matlab. Los modelos mixtos, los MRF generativos y los PoHMM fueron directamente implementados por nosotros en Matlab R2008a.

6.2 MODELOS GENERATIVOS VS. DISCRIMINATIVOS

En esta sección comparamos el HMM clásico, el cual es un modelo generativo con un modelo discriminativo, un CRF de estructura similar (ver figuras 10 y 18). En primer lugar, examinamos como se comportan ambos modelos con buenos datos, esto es, con un subconjunto de secuencias de la base de datos KTH de las que se extrae características no ruidosas y sin cambios de vista ni zoom, estas secuencias además han sido segmentadas manualmente en ciclos de acción comenzando desde la misma posición inicial (secuencias alineadas). Posteriormente, evaluamos los modelos con la totalidad de las secuencias de KTH, incluido el escenario 2 (ver sección 5.1), y sin segmentarlas en ciclos para observar como se comportan ambos modelos en condiciones más ruidosas y realistas.

6.2.1 *Marco experimental*

Para establecer el número teórico de estados que debe tener un HMM para modelar adecuadamente un problema concreto hay que recurrir al análisis de cuánto varía la información contenida en la secuencia de observación [125, 80]. No obstante, la mayoría de trabajos fijan empíricamente el número de estados [21, 126, 32]. Nosotros variamos el número de estados del HMM desde 2, el mínimo, hasta 10, que es la longitud de la secuencia más corta y que determina el máximo número posible de estados. La distribución de observación de cada estado se define con una función de densidad gaussiana, suficiente para modelar estadísticamente con éxito los vectores de características y que es la función de densidad clásica en estos modelos debido a sus útiles propiedades [72].

En los CRF cada fotograma es etiquetado con la clase a la que más probablemente pertenece, la etiqueta más votada en una secuencia corresponde a la clase de dicha secuencia. Los CRF son evaluados con contexto de observación de distinto tamaño, $W = 0$ (sin contexto), $W = 1$ o $W = 2$ ($2W+1$ fotogramas centrados en la observación actual). Para ventanas de observaciones mayores el coste computacional en la fase de entrenamiento es muy elevado. Respecto al término de regularización (σ en la ecuación 4.17), éste se fijó empíricamente a partir de unos pequeños experimentos preliminares tanteando de 0 a 10000 en escala logarítmica de base 10 ([0 0.001 0.01 0.1 1 10 100 1000 10000]), obteniéndose los mejores resultados con 10 en el caso de las características basadas en contornos, y con 100 en las basadas en flujo óptico, como se esperaba al ser el flujo óptico una característica más ruidosa necesita un mayor suavizado.

Los HMM y los CRF son evaluados con vectores de características de tamaño $d = 32$. Para calcular los *shape-context*, el contorno se muestreó uniformemente con 10 puntos en cada región en que se ha segmentado el sujeto. Respecto al flujo óptico, éste no se trata de flujo real sino que es una estimación sobre la región conteniendo al sujeto reducida a 40×40 píxeles, donde se llego a un compromiso entre fiabilidad y eficiencia.

Los parámetros que tienen que aprender cada modelo son:

- HMM:
  $\{\pi(1 \times K), A(K \times K), \mu_{1:K}(d \times K), \Sigma_{1:K}(d \times d \times K\}$, 1123 parámetros cuando el número de estados es $K = 2$ y 5699 para $K = 10$. En el primer experimento, como tenemos menos muestras de entrenamiento usamos matrices de covarianza diagonales por lo que el número de parámetros se reduce a 131 ($K = 2$) y 739 ($K = 10$).

- CRF:
  $\left\{\theta_i(Y \times Y), \theta_{ij}(Y \times d \times (2W+1))\right\}$ con $Y = 2$ (pertenece o no a la clase), en total 68 y 324 parámetros para $W = 0$ y $W = 2$ respectivamente.

Shape-context				Flujo óptico			
(K)	**HMM**	(*W*)	**CRF**	(K)	**HMM**	(*W*)	**CRF**
2	90.8 ± 0.9	0	96.1 ± 0.7	2	90.4 ± 0.7	0	91.7 ± 1.4
3	89.3 ± 0.5	1	97.2 ± 0.3	3	88.1 ± 0.7	1	91.6 ± 2.2
4	87.5 ± 1.0	2	97.3 ± 0.5	4	89.0 ± 0.6	2	92.9 ± 1.4
5	84.3 ± 1.4			5	85.5 ± 1.1		
6	82.2 ± 0.2			6	82.8 ± 0.8		
7	81.5 ± 0.8			7	82.1 ± 1.0		
8	79.1 ± 0.9			8	81.8 ± 1.1		
9	77.3 ± 0.4			9	80.9 ± 0.4		
10	76.5 ± 1.2			10	79.8 ± 0.9		

Tabla 1: **HMM vs. CRF (ciclos).** Resultados promedios de la precisión y su desviación estándar para un subconjunto de la base KTH. Columna de la izquierda: características basadas en *shape-context*; derecha: basadas en histogramas magnitud-orientación del flujo óptico. K es el número de estados del HMM y *W* la ventana de contexto para el CRF.

6.2.2 *Experimento 1: HMM vs. CRF con muestras no ruidosas*

En este primer experimento comparamos el HMM y el CRF usando ambos tipos de características (basadas en contornos y en flujo óptico) para un subconjunto de secuencias de KTH en las que se extraen características no ruidosas, y en las que cada vídeo fue segmentado en ciclos completos inicializados en la misma pose para cada acción. Nosotros definimos un ciclo como: un puñetazo con ambos puños en la acción *boxear*; un aplauso completo en *aplaudir*; levantar los brazos por encima de la cabeza y retornarlos a su posición de reposo en la acción *alzar los brazos*; como en la mayoría de los trabajos de detección de peatones [127, 128] para *andar* y *trotar* se toma un paso con ambos pies, lo que aumenta la robustez; sin embargo, para la acción *correr* consideramos como ciclo un sólo paso con alguna de las dos piernas, ya que en las secuencias grabadas no se completan los dos pasos. En total, el corpus tiene 737 muestras, aproxi-

madamente 492 para entrenamiento y 245 para evaluación: 42 ciclos de *andar hacia la izquierda*, 44 *andar hacia la derecha*, 33 de *trotar*, 51 *correr* y 102 *boxear*, en cada sentido, 160 ciclos de *aplaudir* y 119 de *alzar los brazos*.

Los resultados obtenidos con el CRF superan a los del HMM (tabla 1), siendo la diferencia más acusada en el caso de los *shape-context*. Examinando las matrices de confusión (tablas 2 y 3) se observa que los resultados del HMM caen debido a la acción *correr* que confunde principalmente con *trotar*. La tabla 1 ilustra que el número óptimo de estados ocultos del HMM para modelar los ciclos de acción considerados es 2. Para el CRF se observa que aumentar el contexto mejora ligeramente el reconocimiento, por lo que para el resto de experimentos utilizaremos ventanas de contexto $W = 2$. Las tablas 2 y 3 muestran las matrices de confusión obtenidas con las características basadas en contorno y en flujo óptico respectivamente para un HMM con 2 estados y un CRF con contexto $W = 2$.

6.2.3 *Experimento 2: HMM vs. CRF usando la totalidad de la base KTH*

En este experimento ambos modelos son evaluados utilizando la base de datos KTH completa, inclusive el escenario 2 en el que hay fuertes cambios de vista respecto a los otros escenarios en las acciones en las que el sujeto se desplaza (*andar*, *trotar* y *correr*) y un continuo zoom de la cámara en el resto (*boxear*, *aplaudir* y *alzar los brazos*). Muchas de las secuencias de KTH se ven afectadas por sombras y/o tienen poco contraste dando lugar a características ruidosas. Además, en este experimento no segmentamos las acciones en ciclos, cada secuencia puede comenzar en un punto arbitrario de la acción (secuencias no alineadas) y puede contener un número distinto de ciclos, lo que es más realista desde el punto de vista de un sistema de adquisición de imágenes. Aunque existen técnicas de segmentación automática que se basan en la periodicidad de la señal [129], esto implica un preprocesamiento de las secuencias lo que se traduce en un mayor coste computacional.

HMM	(I) andar	(D) andar	(I) trotar	(D) trotar	(I) correr	(D) correr	(I) boxear	(D) boxear	aplaudir	brazos	%
(I) andar	**40**	0	1	0	0	0	0	1	0	0	**95.2**
(D) andar	0	**44**	0	0	0	0	0	0	0	0	**100.0**
(I) trotar	2	0	**30**	0	0	0	1	0	0	0	**90.9**
(D) trotar	0	5	0	**25**	0	1	0	1	1	0	**75.8**
(I) correr	1	1	19	0	**19**	0	1	10	0	0	**37.3**
(D) correr	0	4	2	12	0	**19**	2	10	2	0	**37.3**
(I) boxear	0	0	0	0	0	0	**102**	0	0	0	**100.0**
(D) boxear	0	0	0	0	0	0	0	**102**	0	0	**100.0**
aplaudir	0	0	0	0	0	0	0	0	**158**	2	**98.8**
brazos	0	0	0	0	0	0	0	0	0	**119**	**100.0**
CRF											
(I) andar	**39**	0	3	0	0	0	0	0	0	0	**92.9**
(D) andar	0	**42**	0	1	1	0	0	0	0	0	**95.5**
(I) trotar	0	0	**31**	0	1	0	0	0	0	1	**93.9**
(D) trotar	0	0	0	**31**	0	0	0	0	1	1	**93.9**
(I) correr	1	0	1	0	**48**	1	0	0	0	0	**94.1**
(D) correr	0	0	0	3	1	**46**	0	0	0	1	**90.2**
(I) boxear	0	0	0	0	0	0	**102**	0	0	0	**100.0**
(D) boxear	0	0	0	0	0	0	0	**101**	1	0	**99.0**
aplaudir	0	0	0	0	0	0	0	0	**158**	2	**98.8**
brazos	0	0	0	0	0	0	0	0	0	**119**	**100.0**

Tabla 2: **Matrices de confusión para el contorno (ciclos)**. Arriba: la matriz de confusión para el HMM con 2 estados ocultos (90.8 ± 0.9). Abajo: la matriz de confusión para el CRF con contexto $W = 2$ (97.3 ± 0.5). Las filas indican las clases verdaderas y las columnas las clases asignadas. La última columna es el % de éxito por clase.

Nuestro corpus se compone de: 200 secuencias de *andar*, *trotar* y *correr*, en cada uno de los sentidos, 313 secuencias de *boxear a la izquierda* mientras que sólo 84 de *boxear*

HMM	(I) andar	(D) andar	(I) trotar	(D) trotar	(I) correr	(D) correr	(I) boxear	(D) boxear	aplaudir	brazos	%
(I) andar	**41**	1	0	0	0	0	0	0	0	0	**97.6**
(D) andar	2	**42**	0	0	0	0	0	0	0	0	**95.5**
(I) trotar	0	0	**31**	1	0	0	0	1	0	0	**93.9**
(D) trotar	0	0	0	**33**	0	0	0	0	0	0	**100.0**
(I) correr	2	0	18	2	**21**	0	7	0	1	0	**41.2**
(D) correr	1	1	1	19	3	**22**	4	0	0	0	**43.1**
(I) boxear	0	0	0	0	0	0	**90**	8	4	0	**88.2**
(D) boxear	1	0	0	0	0	0	3	**96**	2	0	**94.1**
aplaudir	0	0	0	0	0	0	6	0	**154**	0	**96.3**
brazos	0	0	0	0	0	0	0	0	0	**119**	**100.0**
CRF											
(I) andar	**41**	1	0	0	0	0	0	0	0	0	**97.6**
(D) andar	2	**41**	0	1	0	0	0	0	0	0	**93.2**
(I) trotar	5	0	**27**	0	1	0	0	0	0	0	**81.8**
(D) trotar	0	7	0	**25**	0	0	0	0	0	1	**75.8**
(I) correr	2	0	4	2	**41**	0	0	0	0	2	**80.4**
(D) correr	0	1	0	2	0	**45**	0	0	0	3	**88.2**
(I) boxear	0	0	0	0	0	0	**102**	0	0	0	**100.0**
(D) boxear	0	0	0	0	0	0	2	**93**	3	4	**91.2**
aplaudir	0	0	0	0	0	0	0	3	**151**	6	**94.4**
brazos	0	0	0	0	0	0	0	0	0	**119**	**100.0**

Tabla 3: **Matrices de confusión para el flujo óptico (ciclos)**. Arriba: la matriz de confusión para el HMM con 2 estados ocultos (90.4 ± 0.7). Abajo: la matriz de confusión para el CRF con contexto $W = 2$ (92.9 ± 1.4). Las filas indican las clases verdaderas y las columnas las clases asignadas. La última columna es el % de éxito por clase.

a la derecha, 396 de *aplaudir* y 398 de *alzar los brazos*. En total, 2391 muestras, aproximadamente 1590 para entrenamiento y 800 para evaluación. Los resultados de este test

Shape-context				Flujo óptico			
(Q)	**HMM**	(W)	**CRF**	(Q)	**HMM**	(W)	**CRF**
2	62.8 ± 0.9	2	57.9 ± 2.8	2	88.3 ± 0.7	2	73.6 ± 1.6
3	62.0 ± 2.6			3	89.2 ± 1.1		
4	61.9 ± 2.1			4	88.8 ± 1.4		
5	62.2 ± 1.1			5	89.2 ± 1.2		
6	64.0 ± 1.1			6	89.4 ± 1.3		
7	63.8 ± 0.9			7	89.4 ± 0.7		
8	62.4 ± 4.0			8	89.2 ± 1.2		
9	62.4 ± 1.9			9	90.2 ± 1.0		
10	62.8 ± 1.7			10	90.2 ± 1.1		

Tabla 4: **HMM vs. CRF con secuencias ruidosas.** Resultados promedios de la precisión y su desviación estándar en la base KTH. Columna de la izquierda: características basadas en *shape-context*; derecha: basadas en histogramas magnitud-orientación del flujo óptico. K es el número de estados del HMM y *W* la ventana de contexto para el CRF.

son mostrados en la tabla 4. En presencia de características ruidosas los resultados del modelo discriminativo CRF caen ante los del generativo HMM. Concretamente, en el caso de las características de flujo óptico el reconocimiento en el CRF baja entorno al 20% mientras que en el HMM decae sólo un 2% respecto al uso de las secuencias no ruidosas del primer experimento. Analizando las matrices de confusión (tablas 5 y 6) se observa que las acciones *trotar* y *correr* presentan muy baja tasa de reconocimiento en el CRF, éstas se confunden principalmente con acciones similares, sin embargo, *boxear a la derecha* es la acción que ofrece peores resultados en este modelo. En el HMM también se produce una caída del reconocimiento para esta acción pero no tan acusada como en el caso de los CRF. *Boxear a la derecha* es la acción con menos número de secuencias de entrenamiento, los modelos discriminativos al estar más ligados a los datos de entrenamiento tienen menor poder de generalización que los generativos. No obstante, el HMM

HMM	(I) andar	(D) andar	(I) trotar	(D) trotar	(I) correr	(D) correr	(I) boxear	(D) boxear	aplaudir	brazos	%
(I) andar	**120**	7	49	6	11	6	0	0	1	0	**60.0**
(D) andar	18	**121**	14	34	4	8	0	0	1	0	**60.5**
(I) trotar	43	9	**79**	5	58	4	1	0	0	1	**39.5**
(D) trotar	4	26	20	**86**	2	62	0	0	0	0	**43.0**
(I) correr	18	1	80	3	**97**	1	0	0	0	0	**48.5**
(D) correr	6	15	4	53	7	**114**	1	0	0	0	**57.0**
(I) boxear	3	0	2	1	1	3	**272**	3	21	7	**86.9**
(D) boxear	0	0	1	1	3	2	3	**36**	35	3	**42.9**
aplaudir	0	3	0	1	4	0	33	2	**240**	113	**60.6**
brazos	0	2	0	0	0	0	13	0	59	**324**	**81.4**
CRF											
(I) andar	**148**	15	7	1	2	4	4	0	16	3	**74.0**
(D) andar	18	**165**	2	2	0	2	1	0	4	6	**82.5**
(I) trotar	**71**	25	51	1	34	1	5	0	9	3	**25.5**
(D) trotar	19	**73**	2	52	1	29	4	0	11	9	**26.0**
(I) correr	42	22	41	1	**86**	0	0	0	6	2	**43.0**
(D) correr	25	45	1	38	2	**80**	0	0	5	4	**40.0**
(I) boxear	4	3	2	0	4	4	**235**	2	29	30	**75.1**
(D) boxear	0	4	1	1	1	11	1	20	17	**28**	**23.8**
aplaudir	6	5	2	11	3	0	23	5	**249**	92	**62.9**
brazos	11	6	1	0	2	0	14	4	66	**294**	**73.9**

Tabla 5: **Matrices de confusión para el contorno**. Arriba: la matriz de confusión para el HMM con 7 estados ocultos (63.8 ± 0.9). Abajo: la matriz de confusión para el CRF con contexto $W = 2$ (57.9 ± 2.8). Las filas indican las clases verdaderas y las columnas las clases asignadas. La última columna es el % de éxito por clase.

necesita ahora un mayor número de estados que en el caso anterior (sección 6.2.2) probablemente para poder modelar apropiadamente los fotogramas ruidosos.

HMM	(I) andar	(D) andar	(I) trotar	(D) trotar	(I) correr	(D) correr	(I) boxear	(D) boxear	aplaudir	brazos	%
(I) andar	**197**	0	2	0	1	0	0	0	0	0	**98.5**
(D) andar	2	**192**	0	2	1	2	0	1	0	0	**96.0**
(I) trotar	3	0	**179**	0	18	0	0	0	0	0	**89.5**
(D) trotar	0	6	0	**180**	0	14	0	0	0	0	**90.0**
(I) correr	0	0	51	0	**148**	1	0	0	0	0	**74.0**
(D) correr	0	1	0	65	1	**133**	0	0	0	0	**66.5**
(I) boxear	0	0	2	0	1	0	**301**	3	6	0	**96.2**
(D) boxear	1	0	0	0	3	0	19	**47**	14	0	**56.0**
aplaudir	0	0	0	0	0	0	6	8	**374**	8	**94.4**
brazos	0	0	0	0	0	0	3	0	13	**382**	**96.0**
CRF											
(I) andar	**200**	0	0	0	0	0	0	0	0	0	**100.0**
(D) andar	0	**196**	0	2	0	0	1	0	1	0	**98.0**
(I) trotar	**141**	0	49	0	9	0	0	0	1	0	**24.5**
(D) trotar	0	**116**	0	75	0	9	0	0	0	0	**37.5**
(I) correr	31	0	51	0	**108**	2	0	0	0	8	**54.0**
(D) correr	0	38	0	54	1	**102**	0	0	0	5	**51.0**
(I) boxear	1	0	0	0	0	0	**289**	0	15	8	**92.3**
(D) boxear	0	0	0	0	0	0	3	5	**44**	32	**6.0**
aplaudir	0	0	0	0	0	0	6	1	**347**	42	**87.6**
brazos	0	0	0	0	0	0	0	0	11	**387**	**97.2**

Tabla 6: **Matrices de confusión para el flujo óptico**. Arriba: la matriz de confusión para el HMM con 7 estados ocultos (89.4 ± 0.7). Abajo: la matriz de confusión para el CRF con contexto $W = 2$ (73.6 ± 1.6). Las filas indican las clases verdaderas y las columnas las clases asignadas. La última columna es el % de éxito por clase.

En cuanto a las características, el ruido degrada en mayor grado a las basadas en contornos: el reconocimiento desciende entorno al 40% en el CRF y un 30% en el HMM

	Escenario 1	Escenario 2	Escenario 3	Escenario 4
Shape-context				
HMM ($K = 7$)	72.3 ± 4.7	51.2 ± 3.4	65.1 ± 2.9	66.7 ± 5.5
CRF ($W = 2$)	65.5 ± 2.0	51.0 ± 5.3	59.0 ± 6.0	56.0 ± 3.4
Flujo óptico				
HMM ($K = 7$)	94.6 ± 1.8	81.9 ± 1.7	90.2 ± 3.0	90.8 ± 1.1
CRF ($W = 2$)	80.0 ± 2.5	70.1 ± 1.9	72.9 ± 2.4	71.3 ± 2.1

Tabla 7: **HMM vs. CRF: Escenarios.** Resultados promedio y su desviación estándar en cada escenario.

respecto al primer experimento, frente al 20% (CRF) y al 2% (HMM) con el flujo óptico. Estudiando las matrices de confusión vemos que la principal fuente de confusión con el flujo óptico se da entre acciones similares realizadas en el mismo sentido, por ejemplo *andar* y *trotar*, en cambio, distingue razonablemente bien entre acciones con sentido opuesto. Mientras, que para los contornos se diferencia principalmente dos bloques de confusión, el de las acciones en que el movimiento es gobernado por las piernas y el de las acciones en las que sólo se mueven los brazos. Debido a esto, de aquí en adelante utilizaremos las características basadas en el flujo óptico para evaluar los distintos modelos gráficos.

La tabla 7 muestra los resultados por escenarios de KTH, descritos en la sección 5.1, para el CRF (W=2) y el HMM con 7 estados, el cual presenta los resultados con menor desviación estándar. El escenario que ofrece los peores resultados en todos los casos es el escenario 2 como era de esperar al ser el más variable. La tabla 8 desglosa los resultados de cada escenario por acción en el caso del HMM de 7 estados alimentado con las características basadas en flujo óptico, que es el modelo con el que se han conseguido el mejor resultado: El escenario 1, que es el menos ruidoso, presenta buenos resultados para todas las acciones. La acciones en las que el sujeto se desplaza, especialmente *trotar* y *correr*, se ven más afectadas por las ropas holgadas y el entorno interior de los escenarios 3 y 4 respectivamente.

		Escen. 1	Escen. 2	Escen. 3	Escen. 4
(I)	**andar**	96.0	100	98.0	100.0
(D)	**andar**	98.0	92.0	94.0	96.0
(I)	**trotar**	90.0	90.0	72.0	88.0
(D)	**trotar**	94.0	94.0	76.0	78.0
(I)	**correr**	94.0	46.0	86.0	82.0
(D)	**correr**	96.0	36.0	92.0	82.0
(I)	**boxear**	98.7	84.5	97.5	100.0
(D)	**boxear**	90.0	0.0	75.0	57.1
aplaudir		94.0	99.0	96.9	94.0
brazos		93.0	92.0	91.8	100.0

Tabla 8: **Resultados por acciones en cada escenario para el HMM.** Tanto por ciento de éxito por acciones y escenarios con un HMM de 7 estados alimentando con características basadas en flujo óptico.

En el escenario 2, las acciones *correr* en ambos sentidos y *boxear a la derecha* presentan una tasa de aciertos muy bajas, especialmente esta última donde ninguna secuencia de dicho escenario se clasifica correctamente (0%). Queremos llamar la atención sobre que *boxing a la derecha* es la acción con menos muestras de entrenamiento y que las secuencias de *correr* son las de menor duración, lo que se traduce en menos muestras por estado, y por tanto los parámetros aprendidos en ambas clases pueden manejar peor las variaciones en los patrones de los vectores de características lo que probablemente justifique sus bajas tasas de reconocimiento.

6.2.4 *Experimento 3: CRF con capa de estados ocultos (LDCRF)*

Nuestro propósito con este experimento es estudiar si la incorporación en los campos condicionales de un capa de variables de estado ocultas, LDCRF (figura 20), que infiera la estructura latente en las acciones mejora los resultados en reconocimiento de los modelos discriminativos. El número

Flujo óptico		
(K)	**HMM**	**LDCRF** ($W = 2$)
2	88.3 ± 0.7	72.8 ± 2.8
3	89.2 ± 1.1	71.9 ± 2.3
4	88.8 ± 1.4	72.3 ± 3.0
5	89.2 ± 1.2	73.5 ± 3.6
6	89.4 ± 1.3	75.5 ± 1.7

Tabla 9: **HMM vs. LDCRF.** Resultados promedios de la precisión y su desviación estándar en la base KTH para características basadas en histogramas magnitud-orientación del flujo óptico. K es el número de estados por clase en el HMM y en el LDCRF y W la ventana de contexto tomada en el LDCRF.

de parámetros del LDCRF con un contexto $W = 2$ es igual a 656 con $K = 2$ (número de estados por clase) ($2K \times 2K + 2K \times d \times (2W + 1)$) frente a los 324 parámetros del CRF usando el mismo tamaño de contexto.

Los resultados obtenidos para flujo óptico son presentados en la tabla 9, podemos ver que el LDCRF no alcanza los resultados de reconocimiento del HMM. Si analizamos la matriz de confusión (tabla 10) y los resultados por escenario (tabla 11) vemos que el LDCRF tiene los mismos puntos débiles que el CRF, las principales fuentes de error provienen de la confusión de *trotar* y *correr* con acciones similares en el mismo sentido y de la acción *boxear a la derecha* que no supera el 6% de aciertos; y los escenarios con mejores resultados son los escenarios 1 y 3. En general, con muestras ruidosas los modelos discriminativos reconocen peor que los generativos, aún modelando la composición estructural de la acción. En [101] se afirma que los modelos generativos son capaces de ajustar mejor que los discriminativos patrones complejos.

	(I) andar	(D) andar	(I) trotar	(D) trotar	(I) correr	(D) correr	(I) boxear	(D) boxear	aplaudir	brazos	%
(I) andar	**194**	0	3	0	1	0	0	0	0	2	**97.0**
(D) andar	0	**196**	0	0	0	1	2	0	1	0	**98.0**
(I) trotar	**124**	0	63	0	11	0	0	0	1	1	**31.5**
(D) trotar	0	**157**	0	36	0	7	0	0	0	0	**18.0**
(I) correr	35	0	48	0	**108**	1	0	0	0	8	**54.0**
(D) correr	1	78	0	26	2	**89**	0	0	0	4	**44.5**
(I) boxear	1	0	0	0	0	0	**288**	0	15	9	**92.0**
(D) boxear	0	0	0	0	0	0	2	5	**42**	35	**6.0**
aplaudir	0	0	0	0	0	0	4	1	**355**	36	**89.7**
brazos	0	0	0	0	0	0	0	0	16	**382**	**96.0**

Tabla 10: **Matriz de confusión del LDCRF.** Matriz de confusión para el LDCRF con 2 estados ocultos y ventana de contexto $W = 2$ alimentando con características basada en flujo óptico (73.6 ± 1.6). Las filas indican las clases verdaderas y las columnas las clases asignadas. La última columna es el % de éxito por clase.

	Escen. 1	Escen. 2	Escen. 3	Escen. 4
LDCRF (2-estados)	78.6 ± 4.7	69.5 ± 3.2	74.0 ± 4.1	69.2 ± 4.9

Tabla 11: **LDCRF: Escenarios.** Resultados promedio y su desviación estándar en cada escenario.

6.3 MRF GENERATIVOS

En esta sección queremos evaluar la aplicación al problema de HAR del MRF generativo descrito en la sección 4.5 (ver figura 22), al igual que en los CRF las relaciones entre las variables de estado son no casuales, en cambio se relacionan causalmente con las variables de observación (modelo generativo). Los valores de sus parámetros W y G nos da una medida de la influencia real entre las variables conectadas lo que nos permitirá comprobar si las suposiciones de in-

Escen1	Escen2	Escen3	Escen4	**TOTAL**
91.0 ± 6.9	78.4 ± 1.8	84.3 ± 9.4	90.3 ± 3.4	$\mathbf{86.1 \pm 3.2}$

Tabla 12: **MRF generativo.** Resultados promedio y su desviación estándar en cada escenario.

dependencia condicional realizadas sobre algunos modelos gráficos son suposiciones realistas.

6.3.1 *Marco experimental*

En nuestra implementación del modelo (ver sección 4.5.1) el número de nodos de estado H y el de nodos de observación V se corresponden con el número de fotogramas de las secuencias de la acciones ($H = V = T$). Nosotros fijamos $T = 20$ que es la longitud de la secuencia de menor duración (*running*). Para secuencias de mayor longitud tomamos subsecuencias de veinte fotogramas cada dos. Cada subsecuencia será una muestra de nuestro corpus. En la fase de entrenamiento, la etiqueta más votada por las subsecuencias derivadas de un secuencia dada asigna la clase a dicha secuencia.

Nosotros estudiamos el caso completamente conectado, en el que todos los nodos de estado están conectados entre sí y a su vez conectados a todos los nodos de observación, es decir, todos los nodos de estado serán responsables en mayor o menor grado de generar cada una de las observaciones. Los parámetros del modelo son:

$$\{G(d \times H \times V), m(d \times V), W(H \times H), b(H \times 1), R(V \times d \times H), c(H \times 1)\}$$

donde d es el tamaño del vector de características ($d = 32$), en total 26.460 parámetros, tener en cuenta que W es simétrica con diagonal cero. Sin embargo, al estar densamente conectado, el modelo tiene menos grados de libertad.

		Escen1	Escen2	Escen3	Escen4	**Todos**
(I)	**andar**	82.4	82.4	82.4	100	86.8
(D)	**andar**	97.1	88.2	88.2	100	93.4
(I)	**trotar**	82.4	88.2	70.6	82.4	80.9
(D)	**trotar**	82.4	97.1	47.1	91.2	79.4
(I)	**correr**	88.2	70.6	85.3	70.6	78.7
(D)	**correr**	94.1	32.4	94.1	82.4	75.7
(I)	**boxear**	96.1	76.8	90.4	100	90.2
(D)	**boxear**	87.5	41.7	81.3	80.0	75.0
aplaudir		94.1	79.4	94.1	86.8	88.6
brazos		92.7	91.2	86.7	100	92.6

Tabla 13: **Resultados para el MRF generativo desglosados por acciones.** Tanto por ciento de éxito por acciones y escenarios para un MRF alimentando con características basadas en flujo óptico.

6.3.2 *Discusión y resultados*

El porcentaje de clases correctamente clasificadas con este modelo es 86.1 ± 3.2 (tabla 12), por lo que efectivamente el MRF con salida generativa mejora los resultados (entorno a un 15%) de los modelos discriminativos con conexiones exclusivamente no causales. Sin embargo, no alcanza los resultados obtenidos con el HMM (89.4 ± 0.7) para un HMM con 7 estados. Además, su desviación estándar es también mayor que la del HMM y el CRF lo que indica que el MRF generativo se ve más afectado por la calidad de las muestras de entrenamiento que caen en cada test. La tabla 13 desglosa los resultados por acción en cada escenario, en ella se observa una notable mejora en *boxear a la derecha* que llega al 75% de éxito frente al 6% del CRF y el 56% de el HMM con 7 estados. Al estar densamente conectado, las variables en el MRF generativo tiene menos grado de libertad y necesitan por tanto de menos muestras de entrenamiento para trabajar adecuadamente.

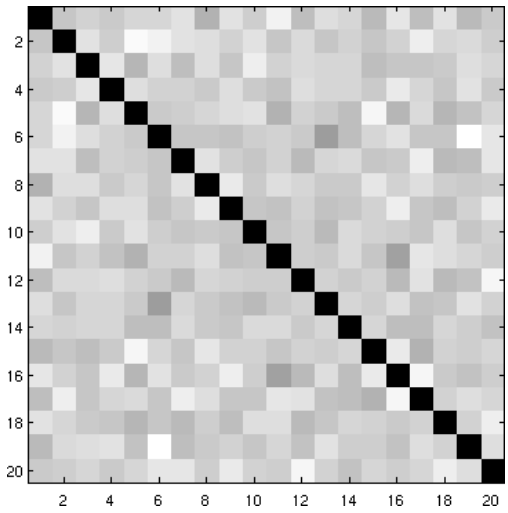

Figura 29: **W**. Matriz de pesos de las conexiones entre las variables de estado ocultas.

La figura 29 muestra los pesos W de las conexiones entre los nodos (variables) de estado. En esta figura puede observarse que cuando un nodo S_i está conectado a otro nodo S_j con un peso de conexión (w_{ij}) alto, los nodos contiguos al nodo S_i contribuyen al valor del nodo S_j en menor medida. Esto se ve mejor calculando histogramas de las distancias (en fotogramas) a los nodos que están más fuertemente ligados a cada nodo. Para ello, tomamos para cada nodo los cinco nodos con los que tiene mayor peso de conexión y medimos las distancias a la que se encuentran respecto a él, a continuación hallamos el histograma de estas distancias pesado por el valor de la conexión. Los histogramas hallados para las distintas acciones de la base KTH se muestran en la figura 30. Los nodos que más influye en uno dado son los contiguos disminuyendo esta influencia conforme aumenta la distancia. Si miramos detenidamente los histogramas observamos que para la acción *andar* hay un máximo local entorno a una distancia de 15 fotogramas, la frecuencia media de un paso para un peatón se encuentra entre los $1.56 - 2\text{Hz}$ [130] lo que equivale aproximadamente a $14 - 16$ fotogramas en secuencias grabadas a 25fps, por lo que a esta distancia tendremos de nuevo una pose similar.

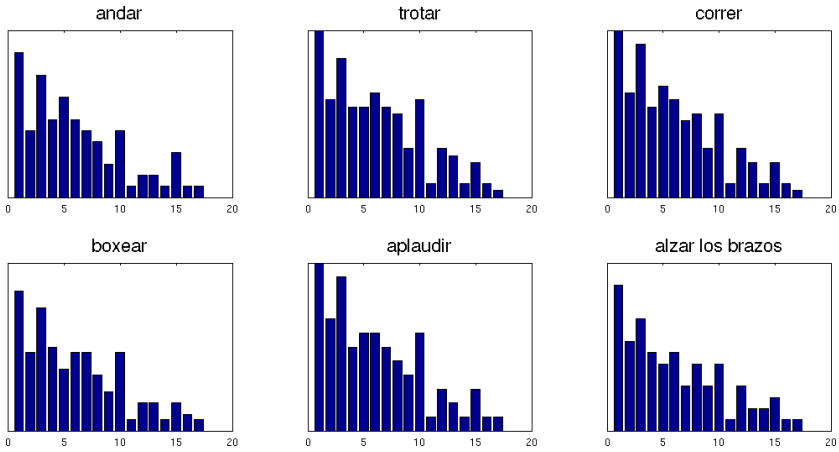

Figura 30: **Histogramas de W.** Histogramas de las distancias (en fotogramas) en que se encuentran los nodos de estado cuyas conexiones con otros nodos de estado tienen mayor peso. Los histogramas están pesados por el peso de la conexión.

Histogramas similares son también calculados para los pesos G de las conexiones con las variables de observación con el propósito de estudiar la influencia de cada nodo de estado en la generación de una observación en el instante t. Los histogramas obtenidos son similares para todas las acciones. Como era de esperar, el nodo de estado que más influye en el valor de un nodo de observación dado es el que se corresponde en el tiempo. Los resultados para W y G demuestran que las hipótesis de cadenas de primer orden y de que las observaciones sólo depende del estado de la variable oculta en el mismo instante, que hemos usado en los otros modelos gráficos, son suposiciones realistas al menos en las acciones consideradas en nuestros experimentos.

6.4 MODELOS GRÁFICOS DIRIGIDOS MULTI-CADENA

Muchos movimientos pueden descomponerse en una combinación de señales periódicas, por ejemplo, según [128] el movimiento producido por los muslos al andar puede considerarse como el movimiento de un péndulo que oscila con frecuencia f, mientras que las pantorrillas oscilarían con

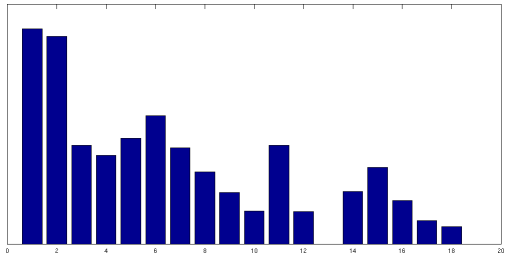

Figura 31: **Histogramas de G.** Histogramas de las distancias (en fotogramas) en que se encuentran los nodos de estado cuyas conexiones con los nodos de observación tienen mayor peso. Los histogramas están pesados por el peso de la conexión.

el doble de frecuencia $2f$. Así mismo, en la acción *alzar los brazos* estos pasan dos veces por una posición intermedia, al inicio de la acción y en el retorno, por lo que la frecuencia de esta pose se multiplica por un factor 2 con respecto a la frecuencia asociada a la acción. Según estas consideraciones en los modelos multi-cadena cada una de las cadenas puede aprender la evolución de un proceso cíclico distinto, por lo que son capaces de modelar mejor este tipo de acciones que los modelos con una única cadena de Markov.

En esta sección, vamos a examinar tres diferentes modelos multi-cadena: los FHMM, los CHMM y los PaHMM, cada uno aportando distintas propiedades relevantes a nuestro problema de HAR. Los FHMM modelan un vector de observación como el resultado de la combinación de varios procesos dinámicos independientes o débilmente acoplados (figura 11), como por ejemplo el movimiento independiente de distintas partes del cuerpo humano. En cambio, en los CHMM cada cadena de Markov modela una observación distinta pero la evolución de cada una depende también del resto, lo que permite modelar coordinación de movimientos, por ejemplo, la coordinación entre los brazos y las piernas cuando el sujeto se desplaza. En los PaHMM, al igual que los en CHMM, cada cadena de Markov da lugar a una secuencia de observación distinta (figura 12) pero sin interacción entre ellas, los procesos son totalmente independientes.

6.4.1 *Marco experimental*

Estudiamos modelos con sólo dos cadenas Markov ya que como expusimos en el capítulo 3, para los FHMM y los CHMM la complejidad de la inferencia crece exponencialmente con el número de cadenas, M, ($O(TMK^{M+1})$), por lo que para más de dos cadenas la inferencia sería computacionalmente muy costosa. De todas formas, el tipo de acciones que componen la base de datos KTH son acciones sencillas que están gobernadas principalmente por el movimiento global de las piernas y/o de los brazos, como también confirma la figura 32 que representa la potencia espectral del vector de características, así que usar dos cadenas para modelar estas acciones sería suficiente. Estos modelos se evaluaron variando el número de estados de cada una cadenas de 2 a 10, para más de 10 estados el FHMM sobreajustaría al incrementarse considerablemente el número de parámetros a aprender.

Para los modelos con distintas secuencias de observación, CHMM y PaHMM, dividimos el vector de característica en dos partes, una relativa a la parte superior del cuerpo humano (tronco, hombros, cabeza y brazos) y otra a la inferior (cadera, muslos, pantorrillas y pies) y asociamos cada una de ellas a una cadena Markov distinta con el objetivo de que cada cadena sea capaz de aprender la dinámica de estos dos grandes bloques globales de movimiento.

Los parámetros de los modelos son:

- FHMM:
  $\{\pi^{(1)}(1 \times K^{(1)}), \pi^{(2)}(1 \times K^{(2)}), A^{(1)}(K^{(1)} \times K^{(1)}),$
  $A^{(2)}(K^{(2)} \times K^{(2)}), \mu_{1:K^{(1)} \times K^{(2)}}(d \times K^{(1)} \times K^{(2)}),$
  $\Sigma_{1:K^{(1)} \times K^{(2)}}(d \times d \times K^{(1)} \times K^{(2)})\}$,
  esto es, 2198 parámetros con un número de estados $K^{(1)} = K^{(2)} = 2$ y tamaño del vector de características $d = 32$.

- CHMM:
  $\{\pi^{(1)}(1 \times K^{(1)}), \pi^{(2)}(1 \times K^{(2)}), A^{(1)}(K^{(1)} \times K^{(2)} \times K^{(1)}),$
  $A^{(2)}(K^{(1)} \times K^{(2)} \times K^{(2)}), \mu^{(1)}_{1:K^{(1)}}(d \times K^{(1)}), \mu^{(2)}_{1:K^{(2)}}(d \times K^{(2)}),$

HMM	FHMM	CHMM	PaHMM
$(7-estados)$	(4×8)	(3×7)	(6×10)
89.4 ± 0.7	91.0 ± 0.6	88.3 ± 0.6	88.2 ± 0.1

Tabla 14: **HMM vs. modelos multi-cadena.** Resultados promedios de la precisión y su desviación estándar en la base KTH para el HMM y sus extensiones multi-cadena alimentados con características basadas en flujo óptico. Los números entre paréntesis indican la combinación del número de estados con los que consiguieron los mejores resultados.

$\Sigma^{(1)}_{1:K^{(1)}}(d\times d\times K^{(1)}), \Sigma^{(2)}_{1:K^{(2)}}(d\times d\times K^{(2)})\}$,
602 parámetros con $K^{(1)}=K^{(2)}=2$ y $d=16$.

- PaHMM:
  $\{\pi^{(1)}(1\times K^{(1)}), \pi^{(2)}(1\times K^{(2)}), A^{(1)}(K^{(1)}\times K^{(1)}),$
  $A^{(2)}(K^{(2)}\times K^{(2)}, \mu^{(1)}_{1:K^{(1)}}(d\times K^{(1)}), \mu^{(2)}_{1:K^{(2)}}(d\times K^{(2)}),$
  $\Sigma^{(1)}_{1:K^{(1)}}(d\times d\times K^{(1)}), \Sigma^{(2)}_{1:K^{(2)}}(d\times d\times K^{(2)})\}$,
  598 parámetros con $K^{(1)}=K^{(2)}=2$ y $d=16$.

6.4.2 *Discusión y resultados*

La tabla 14 recoge el tanto por ciento de éxito de reconocimiento más alto conseguido experimentalmente por cada modelo y con que combinación de estados ocultos se alcanza. El FHMM es el modelo que obtiene el mejor resultado total y por escenarios (tabla 16) concretamente cuando una cadena tiene el doble de estados que la otra (4×8). Desglosando los resultados de clasificación por acciones (tabla 15) se observa que esta mejora se debe principalmente al incremento en una de las acciones más problemáticas, *correr*, que alcanza un 78%, entorno al 10% mejor que con el HMM clásico.

En cuanto al CHMM, aunque teóricamente debería ofrecer mejores resultados [84], ya que en las acciones estudiadas distintas partes del cuerpo se coordinan para producir

		HMM	FHMM	CHMM	PaHMM
estados		(7)	(4 × 8)	(3 × 7)	(6 × 10)
(I)	**andar**	98.5	99.0	96.0	97.5
(D)	**andar**	96.0	97.0	93.5	95.0
(I)	**trotar**	89.5	88.0	87.0	79.5
(D)	**trotar**	90.0	92.0	88.0	86.0
(I)	**correr**	74.0	78.0	71.5	73.5
(D)	**correr**	66.5	79.0	74.5	71.0
(I)	**boxear**	96.2	95.5	92.3	93.3
(D)	**boxear**	56.0	53.8	59.5	61.9
aplaudir		94.4	96.2	94.2	95.2
brazos		96.0	96.2	95.0	96.0

Tabla 15: **HMM vs. modelos multi-cadena por acciones.** Tanto por ciento de clases correctamente clasificadas en la base KTH para el HMM y sus extensiones multi-cadena alimentados con características basadas en flujo óptico. Los números entre paréntesis indican la combinación del número de estados para los que se consiguieron los mejores resultados.

un movimiento global (véase *aplaudir* por ejemplo) o el balanceo en los brazos para mantener el equilibrio en *andar*, *trotar* y *correr*. Sin embargo, en la práctica este modelo puede verse perturbado por el ruido de la otra secuencia de observación durante la estimación de los parámetros [131]. En el caso del PaHMM, en el que se ha aprendido un modelo para la parte superior del cuerpo y otro para la inferior independientemente, aunque cada modelo no se ve afectado por el posible ruido en las otras partes del cuerpo tampoco se ve apoyado por la información que éstas ofrecen. Por ejemplo, para clasificar una acción como *correr* o *boxear* es muy informativo si la parte inferior permanece o no en reposo.

	Escen. 1	Escen. 2	Escen. 3	Escen. 4
HMM (7-estados)	94.6 ± 1.8	81.9 ± 1.7	90.2 ± 3.0	90.8 ± 1.1
FHMM (4×8)	95.7 ± 1.1	84.5 ± 3.6	90.9 ± 3.8	92.8 ± 1.0
CHMM (3×7)	93.8 ± 1.5	80.9 ± 3.8	88.4 ± 3.0	89.9 ± 1.7
PaHMM (6×10)	94.3 ± 1.3	80.3 ± 2.3	89.5 ± 2.7	88.6 ± 1.0

Tabla 16: **HMM vs. modelos multi-cadena por escenarios.** Resultados promedios de la precisión y su desviación estándar en cada escenario KTH para el HMM y sus extensiones multi-cadena alimentados con características basadas en flujo óptico. Los números entre paréntesis indican la combinación del número de estados con los que consiguieron los mejores resultados.

6.5 EL POHMM

En los experimentos anteriores los mejores resultados han sido obtenidos con el FHMM (tabla 14), sin embargo, este modelo tiene que inferir las probabilidades para todas las combinaciones posibles de estados entre todas las cadenas que lo componen, por lo que son computacionalmente costosos y en la práctica su estructura se ve limitada a dos cadenas. Este modelo tiene una estructura similar al FHMM pero las variables de estado se relacionan con las variables de observación mediante conexiones no causales (ver figura 14). La inferencia en este modelo tiene un coste de $O(TMK^2)$ frente al $O(TMK^{M+1})$ del FHMM, podemos entonces utilizar tantas cadenas de Markov como sea necesario en nuestro problema concreto ya que el coste computacional sólo crece linealmente.

En este experimento evaluaremos el PoHMM formulado por Brown y Hinton [68] y una extensión en la que hemos incluido en el modelo distribuciones de observación compuestas por combinaciones lineales de gaussianas multivariadas (ver apéndice A), de forma que el modelo pueda

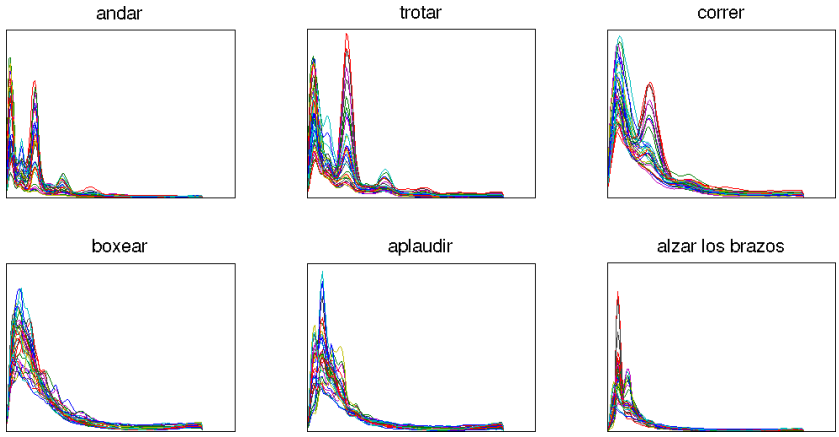

Figura 32: **Potencia espectral.** De los componentes del vector de características ($32 - D$) para cada acción.

adecuarse mejor a la gran diversidad de estilos en la ejecución de una misma acción.

También compararemos los resultados obtenidos usando la función de partición Z que normaliza la distribución de probabilidad del PoHMM estimada con dos métodos distintos: mediante una regresión logística (estimación relativa de Z) y mediante nuestro esquema basado en AIS (estimación relativa y absoluta de Z).

6.5.1 *Marco experimental*

En esta sección responderemos a algunas de las cuestiones abiertas relativas a la implementación del PoHMM, como es el caso del número óptimo de cadenas de Markov en el PoHMM y el de gaussianas modelando las observaciones en un problema concreto.

NÚMERO DE CADENAS EN EL POHMM. Para estimar el número óptimo de cadenas del PoHMM que modele adecuadamente las acciones de la base de datos KTH analizamos la potencia espectral de las componentes del vector de características. El propósito es hallar el número de frecuencias dominantes, el cual nos informa sobre las distintas dinámicas implicadas en la acción cíclica considerada. A cada frecuencia dominante se le asocia una cadena, la

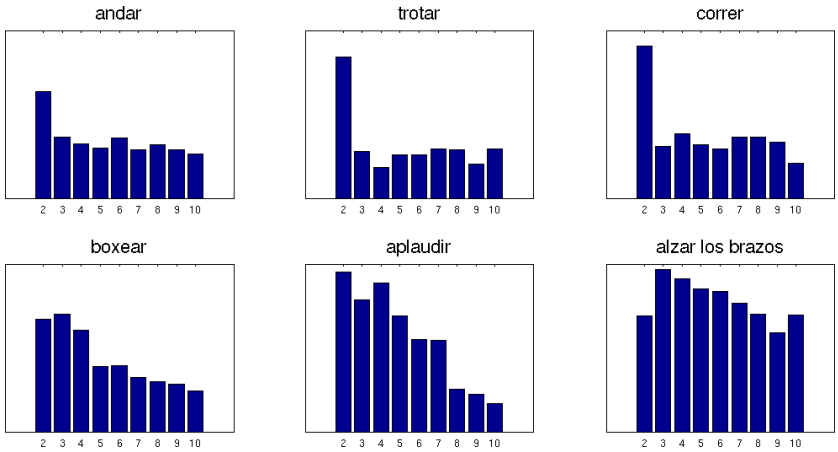

Figura 33: **Agrupamiento (*clustering*) en el espacio de características.** Valores de separación del agrupamiento para distintos números de *clusters* (de 2 a 10) en el espacio de características.

relación entre el valor de las frecuencias también nos da una idea de la relación entre el número de estados de las cadenas. Aquellas cadenas que modelen dinámicas lentas necesitarán menos estados lógicos que las que varían rápidamente que sugieren mayor número de estados para modelar adecuadamente su evolución temporal.

La figura 32 muestra que las características de *andar*, *trotar* y *correr* presentan 2 picos bien diferenciados en la frecuencia, mientras que el resto, *boxear*, *aplaudir* y *alzar los brazos* presentan un único pico, lo que concuerda con el tipo de acciones analizadas: En las primeras, el movimiento se debe a dos fuentes principales distintas, las piernas y los brazos, mientras que en las segundas sólo se debe a los brazos. Estos resultados muestran que con nuestras características es suficiente usar PoHMM formados por dos cadenas, cada una de las cuales debería adaptarse a una de las dos frecuencias dominantes. Evaluaremos este PoHMM usando distintas combinaciones del números de estados de ambas cadenas, el número de estados máximo se fijó en 20, que es el número de fotogramas de la secuencia más corta. Para tener distinta resolución temporal con cada unas de las cadenas del PoHMM, variamos una de ellas de 2 a 10 estados y la otra de 11 a 20.

NÚMERO DE GAUSSIANAS. Para estimar el número de componentes gaussianas de la distribución de observación suficiente para modelar estadísticamente la variabilidad de los vectores de características extraídos de nuestros datos construimos *clusters* en el espacio de características mediante el algoritmo K-medias, el número de gaussianas será el número de *clusters* que mejor ajuste los datos. Para ello, ajustamos a las características distinto número de *clusters* y calculamos el valor de separación entre los grupos:

$$s = \frac{1}{N}\sum_{i=1}^{N}\frac{min([b(i,1),\ldots,b(i,G)]) - a(i)}{max\big(a(i), min([b(i,1),\ldots,b(i,G)])\big)} \quad (6.1)$$

donde G es el número de *clusters* considerado, N el número de características, $a(i)$ es la distancia promedio desde el punto i^{th} a los otros puntos dentro del mismo *cluster*, y $b(i,g)$ la distancia promedio desde ese punto a puntos del *cluster* g. Cuanto mayor sea s mayor será el grado de pertenencia de las características a su grupo. La figura 33 muestra que los mejores valores de separación se obtiene para 2 *clusters* por lo que 2 gaussianas serán suficientes para modelar nuestras características.

Los parámetros de un PoHMM formado por 2 cadenas y cuya distribución de observación es una combinación lineal de 2 gaussianas son:

$$\Big\{\pi^{(m)}(1\times K^{(m)}), A^{(m)}(K^{(m)}\times K^{(m)}), [C^{(m)}_{G_1}, C^{(m)}_{G_2}]$$
$$\big[\{\mu^{(m)}_{G_1}\}_1^{K^{(m)}}, \mu^{(m)}_{G_2}\}_1^{K^{(m)}}\big](d\times K^{(m)}),$$
$$\big[\{\Sigma^{(m)}_{G_1}\}_1^{K^{(m)}}, \Sigma^{(m)}_{G_2}\}_1^{K^{(m)}}\big](d\times d\times K^{(m)})\Big\}_{m=1}^{M=2}$$

donde $C^{(m)}_{G_i}$ es el peso de la gaussiana i de la cadena m, $K^{(m)}$ su número de estados y d el vector de características $d = 32$. En la práctica estos parámetros son reformulados como funciones softmax [88], $\sigma_i(X) = e^{X_i}/\sum_k e^{X_j}$, para eliminar la necesidad de incorporar en las ecuaciones (apéndice A) las restricciones de positividad y que la suma de las probabilidades sea 1.

Finalmente, con respecto al cálculo de la función de partición Z basado en AIS, un importante asunto práctico es

la discretización de las β que definen la probabilidades intermedias y que depende del problema concreto estudiado. Nosotros examinamos tres diferentes distribuciones:

- $\beta = \{0 : 0.001 : 1\}$, 1000 distribuciones uniformemente espaciadas.
- $\beta = \{0 : 0.01 : 0.5, 0.5 : 0.001 : 0.9, 0.9 : 0.0001 : 1\}$, 1453 distribuciones con espaciado uniforme en el logaritmo de β que según [91] es el esquema óptimo.
- $\beta = \{0 : 0.001 : 0.5, 0.5 : 0.0001 : 0.9, 0.9 : 0.0001 : 1\}$, siguiendo el esquema anterior pero aumentando el número de distribuciones, en total 5503 distribuciones.

6.5.2 *Discusión y resultados*

La figura 34 muestra la estimación del $\log Z$ de cada acción y su varianza usando AIS ($\log Z_{AIS}$) con tres distribuciones distintas de β (descritas en la sección 6.5.1) para cada test de la validación cruzada 3-partes. En ella se observa que las estimaciones de $\log Z_{AIS}$ para las tres distribuciones de β son muy similares en todos los casos por lo que no influirá en nuestro problema concreto.

El PoHMM que obtiene el mayor porcentaje de aciertos con una distribución de observación formada por 2 gaussianas es un PoHMM con 7×12-estados, 90.8 ± 0.9, mientras que un PoHMM con la misma combinación de estados ocultos pero con una distribución de observación gaussiana sólo alcanza el 88.5 ± 3.6 (tabla 17), por lo que incorporar al PoHMM distribuciones de observación combinaciones lineales de varias gaussianas supone una notable mejora en el modelo. Estos resultados son obtenidos usando estimaciones de $\Delta \log Z$ halladas mediante una regresión logística ($\Delta \log Z_{reg}$) sobre las probabilidades no normalizadas. Re-estimando los resultados de este modelo (PoHMM 7×12 con 2 gaussianas) con $\Delta \log Z_{AIS}$ y $\log Z_{AIS}$, obtenemos un 89.0 ± 3.0 y 89.2 ± 2.9 respectivamente, prácticamente iguales, pero en el primer caso es necesario calcular

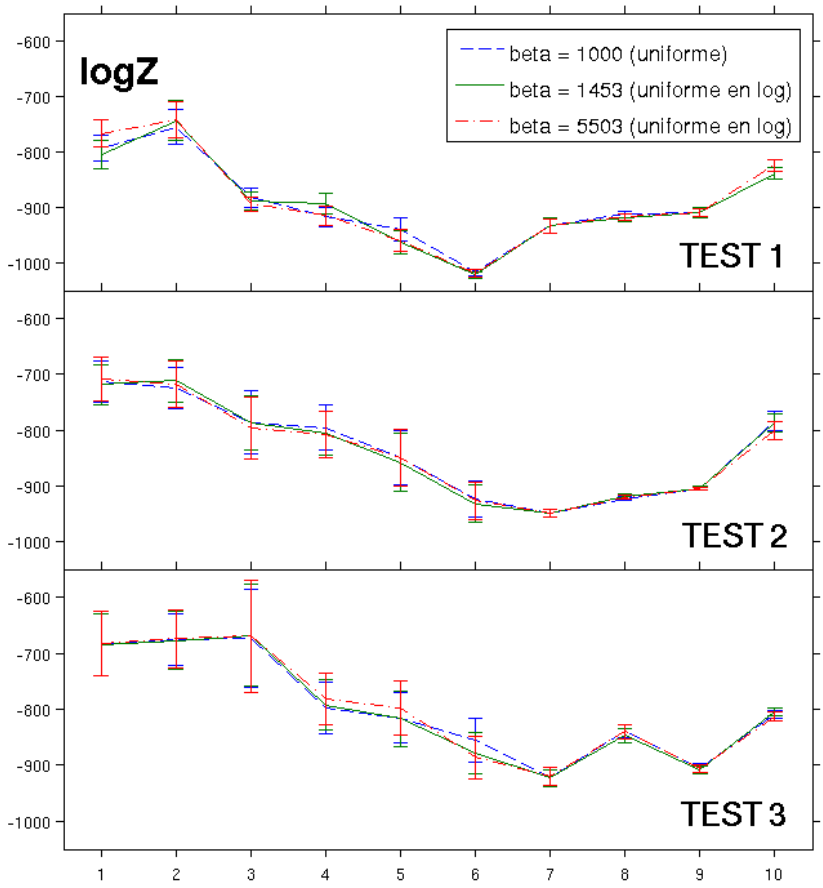

Figura 34: $\log Z_{AIS}$ **de cada acción y su varianza.** En cada uno de los test de la validación cruzada 3-partes con 3 distribuciones distintas de β. 1:(I)andar, 2:(D)andar, 3:(I)trotar, 4:(D)trotar, 5:(I)correr, 6:(D)correr 7:(I)boxear, 8:(D)boxear, 9:aplaudir, 10:alzar los brazos, con (I) indicado sentido hacia la izquierda y (D) hacia la derecha.

40 estimaciones de $\Delta \log Z_{AIS}$ frente a las 10 estimaciones de $\log Z_{AIS}$ del segundo.

La tabla 17 recoge estos resultados para cada test de la validación y el promedio. En ella se observa que el resultado promedio conseguido usando $\log Z_{AIS}$ es menor que con $\Delta \log Z_{reg}$, sin embargo, en el test 1 es al contrario. Si analizamos la figura 34 vemos que las varianzas de $\log Z_{AIS}$ estimadas para el test 1 son mucho más pequeñas que en el resto de tests, por lo que las estimaciones de $\log Z_{AIS}$ en

PoHMM (7×12-estados)	Test 1	Test 2	Test 3	**Prom.**
1-Gaussiana ($\Delta \log Z_{reg}$)	91.6	84.6	89.3	$\mathbf{88.5 \pm 3.6}$
2-Gaussiana ($\Delta \log Z_{reg}$)	91.8	90.1	90.6	$\mathbf{90.8 \pm 0.9}$
2-Gaussiana ($\Delta \log Z_{AIS}$)	92.5	87.5	87.0	$\mathbf{89.0 \pm 3.0}$
2-Gaussiana ($\log Z_{AIS}$)	92.5	88.2	86.9	$\mathbf{89.2 \pm 2.9}$

Tabla 17: **PoHMM.** Tanto por ciento de precisión para los tests de la validación cruzada 3-partes y el promedio con un PoHMM de 7×12 estados alimentado con características extraídas del flujo óptico. Primera y segunda columna, resultados con $\Delta \log Z_{reg}$ para distribuciones de 1 y 2 gaussianas respectivamente. Tercera y cuarta columna, resultados con $\Delta \log Z_{AIS}$ y $\log Z_{AIS}$ respectivamente con distribución de observación del PoHMM combinación de 2 gaussianas.

el primer test son más exactas. La acción *trotar* presenta una alta varianza en los test 2 y 3, especialmente en este último. Si observamos las matrices de confusión (tabla 18) del modelo con $\Delta \log Z_{reg}$ y $\log Z_{AIS}$ vemos que la caída del segundo caso frente al primero se debe precisamente a esta acción (*trotar*) que confunde mayoritariamente con *correr*, en cambio esta última mejora (88.5% con $\log Z_{AIS}$ frente al 74.8% con $\Delta \log Z_{reg}$).

Finalmente, queremos llamar la atención sobre que el resultado total conseguido por los PoHMM (con ambas estimaciones de Z) cae debido a la baja tasa de aciertos en *boxear a la derecha*. Dada la naturaleza de nuestra tarea es necesario un número considerable de muestras de las que aprender la gran variabilidad inherente en nuestro problema de HAR.

	(I) andar	(D) andar	(I) trotar	(D) trotar	(I) correr	(D) correr	(I) boxear	(D) boxear	aplaudir	brazos	%
(I) andar	**198**	0	1	0	1	0	0	0	0	0	**99.0**
(D) andar	1	**195**	0	1	0	2	1	0	0	0	**97.5**
(I) trotar	8	0	**178**	0	14	0	0	0	0	0	**89.0**
(D) trotar	0	1	0	**189**	0	10	0	0	0	0	**94.5**
(I) correr	0	0	56	0	**143**	1	0	0	0	0	**71.5**
(D) correr	0	1	0	43	0	**156**	0	0	0	0	**78.0**
(I) boxear	0	0	0	0	2	0	**302**	1	8	0	**96.5**
(D) boxear	2	0	1	0	2	0	17	**44**	18	0	**52.4**
aplaudir	0	0	0	0	0	0	6	2	**382**	6	**96.5**
brazos	0	0	0	0	0	0	4	0	10	**384**	**96.5**
(I) andar	**195**	0	3	0	2	0	0	0	0	0	**97.5**
(D) andar	1	**191**	0	4	0	3	1	0	0	0	**95.5**
(I) trotar	2	0	**151**	0	46	1	0	0	0	0	**75.5**
(D) trotar	0	0	0	**136**	0	64	0	0	0	0	**68.0**
(I) correr	0	0	27	0	**171**	2	0	0	0	0	**85.5**
(D) correr	0	0	0	17	0	**183**	0	0	0	0	**91.5**
(I) boxear	0	0	0	0	2	0	**301**	0	10	0	**96.2**
(D) boxear	1	0	0	0	3	0	18	**37**	25	0	**44.1**
aplaudir	0	0	0	0	0	0	6	1	**385**	4	**97.2**
brazos	0	0	0	0	0	0	1	0	16	**381**	**95.7**

Tabla 18: **Matrices de confusión para el PoHMM.** Para el caso de un PoHMM 7×12- estados con salida combinación de 2 gaussianas alimentado con las características extraídas del flujo óptico. Arriba: Resultados con $\Delta \log Z_{reg}$ (90.8 ± 0.9). Resultados con $\log Z_{AIS}$ (89.2 ± 2.9). Las filas indican las clases verdaderas y las columnas las clases asignadas. La última columna es el % de éxito por clase.

6.6 RESULTADOS PARA KTH EN LA LITERATURA

La comparación directa con trabajos previos no es posible ya que en el resto de trabajos estudiados se distingue sólo entre 6 acciones, estos no discriminan entre la misma acción pero con sentido opuesto. Sin embargo, nosotros consideramos que esta distinción es importante pues, por ejemplo, tienen un significado semántico completamente distinto que una persona se acerque o se aleje corriendo de otra persona o lugar. Debido a esta discriminación, tenemos la mitad de muestras para aquellas acciones que hemos diferenciado que el resto de trabajos. Teniendo en cuenta la necesidad de muestras de entrenamiento de los modelos de espacio de estados para modelar estadísticamente características con gran variabilidad, como ocurre en el problema de HAR, los resultados de reconocimiento se verán afectados, véase el caso de la acción *boxear a la derecha* cuya baja tasa de aciertos hace caer el resultado total de los modelos gráficos estudiados.

La tabla 19 recoge los resultados de reconocimiento para las 6 acciones de la base de datos KTH que se han publicado en los últimos años y que nos da una idea de la bondad de nuestros resultados. Una comparación cuantitativa entre estos trabajos tampoco sería precisa, los resultados pueden verse afectados por el conjunto de datos de cada test: En el test de tipo *Dejar-uno-fuera* se toma un sujeto para el test de evaluación y se entrena con los 24 restantes, se evalúa entonces con cada sujeto y finalmente se promedia. *Val. cruzada* n indica validación cruzada en n partes, la base de datos se divide en n conjuntos, un conjunto se usa para evaluar y el resto para entrenamiento, estos conjuntos se van rotando hasta completar n pruebas y se promedia. Con $16 + 9$ se toma aleatoriamente 16 sujetos para entrenar y 9 para test un número de veces y se promedia.

A pesar de tener un mayor número de acciones que clasificar, con el PoHMM obtuvimos mejor porcentaje de aciertos que muchos de los trabajos que aparecen en la tabla [132, 133, 134, 135, 136, 137, 138, 139]. Además, la mayoría de los que presentan un resultado superior emplearon un test *Dejar-uno-fuera* [140, 141, 142, 143, 144] o con mayor

muestras de entrenamiento [145], lo que influye apreciablemente en los resultados pues los modelos se construyen más robustamente, por ejemplo, en [140] la clasificación cae del 91.3% al 88.7% cuando usan la mitad de muestras para entrenamiento y la mitad para evaluación en lugar del esquema *Dejar-uno-fuera*. Cuando se toma aleatoriamente los sujetos de entrenamiento y de evaluación [146, 32, 147, 138] no podemos asegurar que se ha testeado sobre todas las muestras. Por otra parte, algunos de la técnicas requieren un preprocesamiento previo, estabilización del vídeo [148] o alineamiento espacio-temporal manual [143]. Otros suponen exacta recuperación de la pose [141] o realizan una compleja extracción de características [147]. Finalmente, cabe destacar el alto coste computacional del entrenamiento en los trabajos [145, 149].

Trabajos	Publicación	Clas. (%)	Test
Z. Wang [132]	(2009)	87.8	Dejar-1-fuera
M. J. Marín-Jiménez [149]	(2009)	95.3	Val. cruzada 5
M. Lucena [133]	(2009)	90.5	Dejar-1-fuera
Z. Zhang [140]	(2008)	91.3	Dejar-1-fuera
S. Savarese [134]	(2008)	86.8	Dejar-1-fuera
K. Schindler []	(2008)	92.7	Val. cruzada 5
I. Laptev [146]	(2008)	91.8	16 + 9 sujetos
K. Mikolajczyk [141]	(2008)	93.2	Dejar-1-fuera
A. Fathi [135]	(2008)	90.5	Dejar-1-fuera
J. Liu [142]	(2008)	94.2	Dejar-1-fuera
M. Ahmad [136]	(2008)	88.3	16 + 9 sujetos
J. Niebles [137]	(2008)	83.3	Dejar-1-fuera
S. F. Wong [139]	(2007)	86.6	Dejar-1-fuera
T. K. Kim [143]	(2007)	95.3	Dejar-1-fuera
H. Jhuang [147]	(2007)	91.7	16 + 9 sujetos
Y. Wang [144]	(2007)	92.4	Dejar-1-fuera
S. Nowozin [138]	(2007)	84.7	16 + 9 sujetos

Tabla 19: **Resultados de reconocimiento en la base KTH**. La primera columna refiere el trabajo de donde se han extraído los resultados; La segunda el año de publicación; La tercera columna presenta el máximo por ciento de acciones correctamente clasificadas en dichos trabajos; Por último, se indica el tipo de test de evaluación sobre el que se validaron los resultados. En estos trabajos no se discriminan entre la misma acción pero con sentido opuesto (6 acciones).

7

CONCLUSIONES Y TRABAJO FUTURO

En este trabajo contextualizamos en el marco del reconocimiento de acciones humanas los principales modelos de espacios de estados probabilísticos, los cuales comprenden modelos gráficos dirigidos: HMM, FHMM, CHMM y PaHMM; no dirigidos: CRF y LDCRF; y los modelos mixtos: MRF generativo y PoHMM, que nosotros aplicamos por primera vez a esta tarea. Analizando su aporte específico al problema y evaluándolos en iguales condiciones experimentales sobre una base de datos común, lo que permite extraer conclusiones globales.

La base de datos KTH ofrece interesantes características, como la baja resolución y el poco contraste de sus secuencias de vídeo, un número importante de individuos (25) con la consiguiente diversidad de estilos y apariencia, cambios en escala e iluminación y ruido debido a la presencia de sombras y a las propias ropas de los sujetos (gorros, mochilas, bufandas, batas, ...) que distorsionan la figura humana. Sobre las secuencias de esta base de datos se extrajeron características visuales de bajo nivel, flujo óptico, las cuales son más generales a cualquier problema de visión que las de alto nivel y usualmente más rápidas de extraer, lo que supone el primer paso hacia sistemas de reconocimiento que operen en tiempo real.

7.1 CONCLUSIONES

Las principales conclusiones que derivamos de nuestro trabajo son:

1. El modelo discriminativo CRF debe ser alimentado con características no ruidosas para que funcione adecuadamente en la tarea de HAR. Con datos no ruidosos los resultados de reconocimiento alcanzados con el CRF superan a los del HMM (modelo genera-

tivo de topología similar). Sin embargo, esta tendencia se invierte cuando las características empleadas contienen ruido, la tasa de éxito del CRF cae entonces drásticamente frente a la del HMM, incluso incorporando al modelo una capa de variables ocultas (LDCRF) que al igual que los HMM pretende explotar la información contenida en la estructura de la acción.

El CRF es por tanto muy sensible a la calidad de los datos. Así cuando el contorno del sujeto que la realiza la acción se extrae con precisión los resultados de clasificación son altos, entorno al 97%, mientras que con el flujo óptico, que es una característica más inestable, extraído de las mismas secuencias de vídeo se encuentran entorno al 92%.

El comportamiento del MRF generativo, efectivamente, supera al del CRF y el LDCRF, ya que su especificación no está tan ligada a las muestras concretas de entrenamiento.

2. La suposición de que el valor de la variable de estado en el instante t sólo depende del valor de las variables de estado en el instante $t-1$ y la suposición de que la variable de observación en el instante t sólo depende del valor de las variables de estado en ese mismo instante, aplicadas en la mayoría de los modelos evaluados, son simplificaciones realistas y pueden considerarse apropiadas en nuestro problema. El análisis de los pesos de las conexiones entre nodos en los MRF generativos corroboran experimentalmente las anteriores suposiciones teóricas.

3. Aquellos modelos que suponen relaciones de tipo causal entre las variables de estado (HMM, FHMM, CHMM, PaHMM y PoHMM) obtienen mejor resultado que aquellos que suponen influencias simétricas (CRF, LCDRF y MRF). Ya que estos modelos explotan el hilo temporal de las acciones, en la que una determinada pose precede a otra concreta.

   De estos modelos, el que presenta la mayor tasa de éxito es el FHMM, el cual modela las distintas dinámicas presentes en la secuencia de observación

como procesos independientes o débilmente acoplados. Aquellos que dividen las secuencias de observación según su dinámica y consideran total independencia (PaHMM) o fuerza interacción entre ellas (CHMM), pueden ser ambos demasiados restrictivos.

4. El PoHMM se ha revelado como una interesante alternativa a los modelos de espacio de estado clásicos aplicados al problema de HAR. El PoHMM combina fiabilidad y eficiencia y es capaz de modelar la estructura de la acción en múltiples escalas temporales. Los resultados muestran que nuestra reformulación del modelo para que incorpore como distribución de observación una combinación lineal de gaussianas efectivamente supone una mejora del modelo. El número de gaussianas, así como el número de cadenas de Markov en el PoHMM y la relación entre su número de estados para modelar adecuadamente un problema concreto pueden ser fijados mediante el agrupamiento en el espacio de características y el análisis espectral de la señal respectivamente.

5. El cálculo del valor de la función de partición Z permite la comparación de varios PoHMM simultáneamente, lo que supone un notable ahorro computacional. Los experimentos demuestran que el porcentaje de aciertos en clasificación es mejor cuanto más precisa sea la estimación de Z, por lo que será necesario tener un número suficiente de muestras para obtener valores estables de Z. La posibilidad de obtener una estimación no relativa para Z en los PoHMM, junto a su habilidad para modelar eficientemente procesos con ricas texturas de movimiento, posibilitará el manejo de grandes bases de datos constituidas por acciones complejas.

7.2 TRABAJO FUTURO

Para un futuro inmediato, identificamos tres principales líneas de acción:

1. La búsqueda de características más compactas que conserven su capacidad representativa y permitan combinar varias de ellas sin que el vector de características aumente su tamaño considerablemente, por ejemplo, diferencias entre histogramas. La combinación de características basadas en contornos y en flujo óptico pesadas por su relevancia en la acción concreta o bien por su robustez en la escena permitirá contrarrestar las deficiencias de unas u otras en escenarios reales.

2. Otro paso natural es la combinación de señales multimodales, como características acústicas y visuales. El ser humano es capaz de distinguir únicamente por el sonido si otra persona anda o corre sin tener contacto visual con ella, por tanto propiedades como periodicidad y/o patrón de la señal acústica son útiles características que pueden ser naturalmente integradas en los modelos descritos para explotar la información que contienen en el problema de HAR.

3. Por otro lado, queremos explorar la potencialidad del PoHMM y evaluar bases de datos compuestas por acciones de movimientos complejos (pasos de baile, etc.), así como acciones en donde intervengan varias personas.

A

ECUACIONES DEL POHMM CON DISTRIBUCIÓN DE OBSERVACIÓN COMBINACIÓN LINEAL DE GAUSSIANAS

Ahora el PoHMM se define como el producto de densidades de HMM cuyas distribuciones de observación están formadas por combinaciones lineales de G gaussianas por cada estado oculto i. La distribución de observación es entonces caracterizada por el peso de cada gaussiana g, C_{ig}, y por su media μ_{ig} y covarianza Σ_{ig}, con $g \in \{1, \ldots, G\}$ y $i \in \{1, \ldots, K\}$ siendo K el número total de estados.

Para forzar que la covarianza sea semidefinida positiva sin necesidad de incluir esta restricción en las ecuaciones de optimización parametrizamos la covarianza como $\exp(\Sigma_{ig})$. Las probabilidades de transición y de estado inicial son también parametrizadas como una función *softmax*, que además de positividad impone que estas probabilidades sumen uno:

$$A_{ij} = \sigma(a_{ij}) = \frac{\exp(a_{ij})}{\sum_k \exp(a_{ik})}$$
$$\Pi_i = \sigma(\pi_i) = \frac{\exp(\pi_i)}{\sum_k \exp(\pi_k)}$$

La probabilidad conjunta es dada entonces por:

$$\begin{aligned} p(S_t, Y_t) =& p(S_1) \prod_{t=2}^{T} p(S_t|S_{t-1}) \prod_{t=1}^{T} p(Y_t|S_t) \\ =& \prod_i \Pi_i^{\delta_{S_1,i}} \prod_{t=2}^{T} \prod_{i,j} A_{ij}^{\delta_{S_{t-1},i}\delta_{S_t,j}} \\ & \prod_{t=1}^{T} \prod_i \left(\sum_{g=1}^{G} C_{ig} \frac{b_{ig}(Y_t)}{\sqrt{(2\pi)^d}} \right)^{\delta_{S_t,i}} \end{aligned} \tag{A.1}$$

donde el último término representa la distribución de observación de una combinación lineal de las gaussianas,

$b_{ig}(Y_t)/\sqrt{(2\pi)^d}$, con pesos C_{ig}, siendo Y_t es el vector de observación (de tamaño d) en el instante t y $b_{ig}(Y_t)$ igual a:

$$b_{ig}(Y_t) = \frac{e^{-\frac{1}{2}(Y_t-\mu_{ig})' \exp(\Sigma_{ig})^{-1}(Y_t-\mu_{ig})}}{|\exp(\Sigma_{ig})|} \tag{A.2}$$

La verosimilitud logarítmica queda entonces como:

$$\begin{aligned}\mathcal{L} = &\sum_i \delta_{S_1,i} \log \Pi_i \\ &+ \sum_{t=2}^{T} \sum_{i,j} \delta_{S_{t-1},i}\delta_{S_t,j} \log A_{ij} \\ &+ \sum_{t=1}^{T} \sum_i \delta_{S_t,i} \left(\log \sum_g C_{ig} b_{ig}(Y_t) - \frac{d}{2} \log 2\pi \right)\end{aligned} \tag{A.3}$$

Las derivadas parciales de $\mathcal{L}$ con respecto a los parámetros de la matriz de transición a_{ij} y el vector de estado inicial π_i son idénticas a las enunciadas en [88]:

probabilidades de estado inicial

$$\frac{\partial \mathcal{L}}{\partial \pi_j} = \delta_{S_1,j} - \sigma(\pi_j) \tag{A.4}$$

matriz de transición

$$\frac{\partial \mathcal{L}}{\partial a_{jh}} = \sum_{t=2}^{T} \left[\delta_{S_{t-1},j}\delta_{S_t,h} - \delta_{S_{t-1},j}\sigma(a_{jh}) \right] \tag{A.5}$$

Sin embargo, las derivadas parciales respecto de las medias, las covarianzas y los pesos de la combinación de gaussianas toman ahora la forma:

medias

$$\frac{\partial \mathcal{L}}{\partial \mu_{jh}} = \sum_{t=1}^{T} \delta_{S_t,j} \frac{C_{jh} b_{jh}(Y_t)}{\sum_g C_{jg} b_{jg}(Y_t)} \exp(\Sigma_{jg})^{-1}(Y_t - \mu_{jh}) \tag{A.6}$$

covarianzas

$$\frac{\partial \mathcal{L}}{\partial \Sigma_{jh}} = \sum_{t=1}^{T} \delta_{S_t,j} \frac{C_{jh} b_{jh}(Y_t)}{\sum_g C_{jg} b_{jg}(Y_t)} * \left(\frac{1}{2} \exp (\Sigma_{jg})^{-1} (Y_t - \mu_{jh})(Y_t - \mu_{jh})' - I \right) \tag{A.7}$$

donde hemos usado que $\frac{\partial}{\partial X}|X| = |X| * X^{-T}$. I es la matriz identidad de tamaño $d \times d$.

pesos de la gaussianas

$$\frac{\partial \mathcal{L}}{\partial C_{jh}} = \sum_{t=1}^{T} \delta_{S_t,j} \frac{b_{jh}(Y_t)}{\sum_g C_{jg} b_{jg}(Y_t)} \tag{A.8}$$

REFERENCIAS BIBLIOGRÁFICAS

[1] J. K. Aggarwal and Q. Cai. Human motion analysis: A review. *Computer Vision and Image Understanding*, 73(3):428–440, 1999. (Cited on page 2.)

[2] D. M. Gavrila. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding*, 73(1):82–98, 1999. (Cited on page 2.)

[3] L. Wang, W. Hu, and T. Tan. Recent developments in human motion analysis. *National Laboratory of Pattern Recognition*, 36(3):585–601, 2003. (Cited on page 2.)

[4] T. B. Moeslund, A. Hilton, and V. Krüger. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 104(2):90–126, 2006. (Cited on page 2.)

[5] G. Lavee, E. Rivlin, and M. Rudzsky. Understanding video events: A survey of methods for automatic interpretation of semantic occurrences in video. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 39(5):489–504, September 2009. (Cited on pages 2 and 3.)

[6] A. F. Bobick and Y. A. Ivanov. Action recognition using probabilistic parsing. In *Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition*, pages 196–202, 1998. (Cited on pages 2 and 3.)

[7] L. Zelnik-Manor and M. Irani. Statistical analysis of dynamic actions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(9):1530–1535, 2006. (Cited on page 2.)

[8] E. Shechtman and M. Irani. Space-time behavior-based correlation - or - how to tell if two underlying motion fields are similar without computing them?

IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(11):2045–2056, 2007. (Cited on page 2.)

[9] M. Pittore, C. Basso, and A. Verri. Representing and recognizing visual dynamic events with support vector machines. In *Proceedings of the International Conference on Image Analysis and Processing*, pages 18–23, 1999. (Cited on page 2.)

[10] D. Cao, O. T. Masoud, D. Boley, and N. Papanikolopoulos. Online motion classification using support vector machines. In *Proceedings of The IEEE International Conference on Robotics and Automation*, volume 3, pages 2291–2296, 2004. (Cited on page 2.)

[11] Y. Guo, G. Xu, and S. Tsuji. Understanding human motion patterns. In *Proceedings of The International Conference on Pattern Recognition*, volume 2, pages B:325–329, 1994. (Cited on page 2.)

[12] H. Vassilakis, A. J. Howell, and H. Buxton. Comparison of feedforward (tdrbf) and generative (tdrgbn) network for gesture based control. In *Proceedings of the International Gesture Workshop on Gesture and Sign Languages in Human-Computer Interaction*, pages 317–321, 2001. (Cited on page 2.)

[13] K.-H. Jo, Y. Kuno, and Y. Shirai. Manipulative hand gesture recognition using task knowledge for human computer interaction. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, volume 0, pages 468–473, 1998. (Cited on page 2.)

[14] P. Hong, M. Turk, and T. S. Huang. Gesture modeling and recognition using finite state machines. In *Proceedings of the IEEE International Conference and Gesture Recognition*, pages 410–415, 2000. (Cited on page 2.)

[15] F. Lv and R. Nevatia. Single view human action recognition using key pose matching and viterbi path searching. In *Proceedings of the IEEE Conference on*

Computer Vision and Pattern Recognition, pages 1–8, 2007. (Cited on page 2.)

[16] V. D. Shet, D. Harwood, and L. S. Davis. Vidmap: Video monitoring of activity with prolog. In *Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 224–229, 2005. (Cited on page 3.)

[17] A. Kojima, M. Izumi, T. Tamura, and K. Fukunaga. Generating natural language description of human behavior from video images. In *Proceedings of the International Conference on Pattern Recognition*, volume 4, page 4728, 2000. (Cited on page 3.)

[18] M. Albanese, R. Chellappa, V. Moscato, A. Picariello, V. S. Subrahmanian, P. Turaga, and O. Udrea. A constrained probabilistic petri net framework for human activity detection in video. *IEEE Transactions on Multimedia*, 10(8):1429–1443, 2008. (Cited on page 3.)

[19] S. A. Niyogi and E. H. Adelson. Analyzing and recognizing walking figures in xyt. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, number 223, pages 469–474, 1994. (Cited on page 3.)

[20] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2, pages 1395–1402, 2005. (Cited on page 3.)

[21] T. Starner and A. Pentland. Real-time american sign language recognition from video using hidden markov models. *International Workshop on Automatic Face and Gesture Recognition*, 1995. (Cited on pages 3 and 76.)

[22] S. A. Fahmy, P. Y. K. Cheung, and W. Luk. Hardware acceleration of hidden markov model decoding for person detection. In *Proceedings of the Conference on Design, Automatic and Test in Europe*, volume 3, pages 8–13, 2005. (Cited on page 3.)

[23] R. D. Green and L. Guan. Continuous human activity recognition. In *Proceedings of the IEEE Conference on Control,Automation,Robotics and Vision*, volume 1, pages 706–711, 2004. (Cited on page 3.)

[24] I. Reid N. Robertson. A general method for human activity recognition in video. *Computer Vision and Image Understanding*, 104(2,3):232–248, 2006. (Cited on page 3.)

[25] J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequential images using hidden markov model. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, pages 379–385, 1992. (Cited on page 3.)

[26] C. Bregler. Learning and recognizing human dynamics in video sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 568–574, 1997. (Cited on page 3.)

[27] X. Feng and P. Perona. Human action recognition by sequence of movelet codewords. In *Proceedings of the International Symposium on 3D Data Processing Visualization and Transmission*, pages 717–721, 2002. (Cited on page 3.)

[28] V. Kellokumpu, M. Pietikäinen, and J. Heikkilä. Human activity recognition using sequences of postures. In *Proceedings of the Conference on Machine Vision Applications*, pages 570–573, 2005. (Cited on page 3.)

[29] N. Jin and F. Mokhtarian. A non-parametric hmm learning method for shape dynamics with application to human motion recognition. In *Proceedings of the International Conference on Pattern Recognition*, volume 2, pages 29–32, 2006. (Cited on page 3.)

[30] X. Sun., C.-W. Chen, and B. S. Manjunath. Probabilistic motion parameter models for human activity recognition. In *Proceedings of the International Conference on Pattern Recognition*, volume 1, pages 443–446, 2002. (Cited on page 3.)

[31] F. Niu and M. Abdel-Mottaleb. View-invariant human activity recognition based on shape and motion features. In *Proceedings of the IEEE International Symposium on Multimedia Software Engineering*, pages 546–556, 2004. (Cited on page 3.)

[32] M. Ahmad and S-W. Lee. Human action recognition using multi-view image sequences features. In *Proceedings of the International Conference on Automatic Face and Gesture Recognition*, pages 523–528, 2006. (Cited on pages 3, 76, and 102.)

[33] C. Chen, J. Liang, H. Hu, L. Jiao, and X. Yang. Factorial hidden markov models for gait recognition. In *Proceedings of the International Conference on Biometrics*, pages 124–133, 2007. (Cited on page 3.)

[34] C. Vogler, H. Sun, and D. Metaxas. A framework for motion recognition with applications to american sign language and gait recognition. In *Proceedings of the Workshop on Human Motion*, pages 33–38, 2000. (Cited on page 4.)

[35] C. Chen, J. Liang, H. Zhao, H. Hu, and J. Tian. Factorial hmm and parallel hmm for gait recognition. *IEEE Transactions on System, Man, and Cybernetics*, 39(1):114–123, January 2009. (Cited on page 4.)

[36] M. Brand, N. Oliver, and A. Pentland. Coupled hidden markov models for complex action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 994–999, 1997. (Cited on page 4.)

[37] N. Oliver, E. Horvitz, and A. Garg. Layered representations for human activity recognition. In *Proceedings of the IEEE International Conference on Multimodal Interfaces*, pages 3–8, 2002. (Cited on page 4.)

[38] S. Fine, Y. Singer, and N. Tishby. The hierarchical hidden markov model: Analysis and applications. *Machine Learning*, 32(1):41–62, 1998. (Cited on page 4.)

[39] L. Atallah and G-Z. Yang. The use of pervasive sensing for behaviour profiling – a survey. *Pervasive and Mobile Computing*, 5(5):447–464, October 2009. (Cited on page 4.)

[40] V. Pavlovic, J. M. Rehg, and J. Maccormick. Impact of dynamic model learning on classification of human motion. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, volume 1, pages 788–795, 2000. (Cited on page 4.)

[41] N. Moënne-loccoz, F. Brémond, and M. Thonnat. Recurrent bayesian network for the recognition of human behaviors from video. In *Proceedings of the International Conference on Computer Vision Systems*, pages 68–77, 2003. (Cited on page 4.)

[42] N. Oliver, B. Rosario, and A. Pentland. A bayesian computer vision system for modeling human interactions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):831–843, 2000. (Cited on page 4.)

[43] C. Fanti, L. Zelnik-Manor, and P. Perona. Hybrid models for human motion recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 1166–1173, 2005. (Cited on page 4.)

[44] S. Hongeng, R. Nevatia, and F. Bremond. Video-based event recognition: activity representation and probabilistic recognition methods. *Computer Vision and Image Understanding*, 96(2):129–162, 2004. (Cited on page 4.)

[45] R. Díaz-León and L. E. Sucar. Recognition of continuous activities. In *Proceedings of the Ibero-American Conference on Artificial Intelligence*, pages 875–881, 2002. (Cited on page 4.)

[46] R. Szeliski. Bayesian modeling of uncertainty in low-level vision. *International Journal of Computer Vision*, 5(3):271–301, 1990. (Cited on page 4.)

[47] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the International Conference on Machine Learning*, pages 282–289, 2001. (Cited on pages 4, 55, and 61.)

[48] D. L. Vail, M. M. Veloso, and J. D .Lafferty. Conditional random fields for activity recognition. In *Proceedings of The International Joint Conference on Autonomous Agents and Multiagent Systems*, pages 1–8, 2007. (Cited on page 4.)

[49] M. Martínez and L. E. Sucar. Learning dynamic naive bayesian classifiers. In *Florida Artificial Intelligence Research Symposiume*, pages 655–659, 2008. (Cited on page 4.)

[50] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas. Conditional models for contextual human motion recognition. *Computer Vision and Image Understanding*, 104(2-3):210–220, 2006. (Cited on pages 5 and 47.)

[51] M. A. Mendoza and N. Pérez de la Blanca. Applying space state models in human action recognition: A comparative study. In *Proceedings of the International Conference on Articulated Motion and Deformable Objects*, pages 53–62, 2008. (Cited on page 5.)

[52] C. I. Connolly. Learning to recognize complex actions using conditional random fields. In *Proceedings of the International Symposium on Visual Computing*, volume 2, pages 340–348, 2007. (Cited on page 5.)

[53] S. B. Wang, A. Quattoni, L. P. Morency, D. Demirdjian, and T. Darrel. Hidden conditional random fields for gesture recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 1521–1527, 2006. (Cited on pages 5 and 47.)

[54] L. P. Morency, A. Quattoni, and T. Darrell. Latent-dynamic discriminative models for continuous gesture recognition. In *Proceedings of the IEEE Conference*

on Computer Vision and Pattern Recognition, pages 1–8, 2007. (Cited on pages 5, 6, 50, and 51.)

[55] H. Ning, W. Xu, Y. Gong, and T. Huang. Latent pose estimator for continuous action recognition. In *Proceedings of the European Conference on Computer Vision*, number 2, pages 419–433, 2008. (Cited on page 5.)

[56] P. Natarajan and R. Nevatia. View and scale invariant action recognition using multiview shape-flow models. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008. (Cited on page 5.)

[57] L. Han, X. Wu, W. Liang, Y. Jia, and G. Hou. Discriminative human action recognition in the learned hierarchical manifold space. *Image and Vision Computing*, August 2009. (Cited on page 5.)

[58] L. Wang and D. Suter. Recognizing human activities from silhouettes: Motion subspace and factorial discriminative graphical model. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007. (Cited on page 5.)

[59] H. Kjellström, J. Romero, D. Martínez, and D. Kragić. Simultaneous visual recognition of manipulation actions and manipulated objects. In *Proceedings of the 10th European Conference on Computer Vision*, pages 336–349, 2008. (Cited on page 5.)

[60] C. Sutton, A. McCallum, and K. Rohanimanesh. Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data. *The Journal of Machine Learning Research*, 8:693–723, 2007. (Cited on page 5.)

[61] L. Liao, D. Fox, and H. Kautz. Hierarchical conditional random fields for gps-based activity recognition. In *Proceedings of the International Symposium of Robotics Research*, volume 28, pages 487–506, 2007. (Cited on page 5.)

[62] T. Wu, C. Lian, and J. Y. Hsu. Joint recognition of multiple concurrent activities using factorial conditional random fields. In *Proceedings of the Workshop on Plan, Activity, and Intent Recognition*, 2007. (Cited on page 5.)

[63] T. G. Dietterich, A. Ashenfelter, and Y. Bulatov. Training conditional random fields via gradient tree boosting. In *Proceedings of the International Conference on Machine learning*, 2004. (Cited on page 5.)

[64] X. Sun and J. Tsujii. Sequential labeling with latent variables: an exact inference algorithm and its efficient approximation. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, pages 772–780, March 2009. (Cited on page 5.)

[65] Z. Ghahramani and M. I. Jordan. Factorial hidden markov models. *Machine Learning*, 29(2-3):245–273, 1997. (Cited on pages 6, 27, and 41.)

[66] A. Quattoni, S. Wang, L.-P. Morency, M. Collins, and T. Darrel. Hidden-state conditional random fields. In *Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007. (Cited on pages 6 and 54.)

[67] G. E. Hinton, S. Osindero, and K. Bao. Learning causally linked markov random fields. In *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics*, pages 128–135, 2005. (Cited on pages 6, 55, 56, 57, 59, and 61.)

[68] A. D. Brown and G. E. Hinton. Products of hidden markov models. In *Proceedings of Artificial Intelligence and Statistics*, pages 3–11, 2001. (Cited on pages 6, 31, 32, 35, and 95.)

[69] S. L. Lauritzen. *Graphical Models*. Oxford Science Publications, 2003. (Cited on pages 11, 12, 13, 14, and 16.)

[70] F. V. Jensen. *Introduction to bayesian Networks*. Springer, 1996. (Cited on pages 12 and 21.)

[71] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., 1988. (Cited on pages 12, 20, 47, and 61.)

[72] J. A. Bilmes. *Mathematical Foundations of Speech and Language Processing*, chapter Graphical models and automatic speech recognition. SpringerVerlag, 2003. (Cited on pages 13 and 76.)

[73] M. I. Jordan. *Learning in Graphical Models*. MIT Press, 1998. (Cited on page 14.)

[74] P. Smyth, D. Heckerman, and M. Jordan. Probabilistic independence networks for hidden markov probability models. *Neural Computation*, 9(2):227–269, 1997. (Cited on pages 14 and 19.)

[75] S. C. Zhu. Statistical modeling and conceptualization of visual patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(6):691–712, 2003. (Cited on page 16.)

[76] W. L. Buntine. Operations for learning with graphical models. *Journal of Artificial Intelligence Research*, 2:159–225, 1994. (Cited on page 16.)

[77] L. R. Rabiner and B. H. Juang. An introduction to hidden markov models. *IEEE Acoustic, Speech and Signal Processing Magazine*, 3(1):4–16, 2003. (Cited on pages 17 and 26.)

[78] A. J. Viterbi. Error bounds for convolutional codes and an asymptotically optimun decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269, 1967. (Cited on pages 17 and 18.)

[79] F. V. Jensen, S. L. Lauritzn, and K. G. Olensen. Bayesian updating in recursive graphical models by local computations. *Computational Statistical Quartely*, 4:269–2282, 1990. (Cited on page 19.)

[80] J. A. Bilmes. What hmms can do. *IEICE Transactions on Information and Systems*, 89-D(3):869–891, 2006. (Cited on pages 20, 22, and 76.)

[81] U. Kjaerulff. Triangulation of graphs - algorithms giving small total state space. Technical Report R 90-09, Department of Mathematics and Computer Science, Aalborg university, 1990. (Cited on page 20.)

[82] W. X. Wen. Optimal decomposition of belief networks. In *Proceedings of the Uncertainty in Artificial Intelligence*, pages 209–224, 1990. (Cited on page 20.)

[83] F. V. Jensen, K. G. Olesen, and S. K. Andersen. An algebra of bayesian belief universes for knowledge-based systems. *Networks*, 20(5):637–659, 1990. (Cited on page 22.)

[84] M. Brand. Coupled hidden markov models for modeling interacting processes. Technical Report 405, Massachusetts Institute of Technology Media Lab Perceptual Computing, 1997. (Cited on pages 30, 60, and 93.)

[85] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, Serie B 39(1):1–38, 1977. (Cited on page 31.)

[86] L. E. Baum, T. Petrie, G. Soules, and N. Weis. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The Annals of Mathematical Statistics*, 41(1):164–171, 1970. (Cited on page 31.)

[87] G. E. Hinton. Products of experts, 1999. (Cited on pages 32 and 36.)

[88] A. D. Brown. *Product Models for Sequences*. PhD thesis, University of Toronto, 2002. (Cited on pages 33, 37, 97, and 110.)

[89] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002. (Cited on pages 35, 38, and 59.)

[90] V. Gómez, H. J. Kappen, and M. Chertkov. Approximate inference on planar graphs using loop calculus and belief propagation. In *Proceedings of the Annual Conference on Uncertainty in Artificial Intelligence*, January 2009. (Cited on page 37.)

[91] R. M. Neal. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001. (Cited on pages 37, 38, 39, 40, and 98.)

[92] R. Salakhutdinov and I. Murray. On the quantitative analysis of deep belief networks. In *Proceedings of the International Conference on Machine Learning*, pages 872–879, 2008. (Cited on page 38.)

[93] C. Sutton and A. McCallum. *Introduction to Conditional Random Fields for Relational Learning*. MIT Press, 2006. (Cited on page 45.)

[94] A. McCallum. Efficiently inducing features of conditional random fields. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pages 403–410, 2003. (Cited on pages 45 and 60.)

[95] A. Quattoni, M. Collins, and T. Darrel. Conditional random fields for object recognition. In *Proceedings of the Neural Information Processing Systems*, pages 1097–1104, 2004. (Cited on page 48.)

[96] S.F. Chen and R. Rosenfeld. A gaussian prior for smoothing maximum entropy models. Technical report, Carnegie Mellon University, 1999. (Cited on page 52.)

[97] H. Wallach. *Efficient training of conditional random fields*. PhD thesis, University of Edinburgh, 2002. (Cited on page 53.)

[98] R. M. Neal and G. E. Hinton. A view of the em algorithm that justifies incremental, sparse, and other variants. *Learning in Graphical Models*, pages 355–368, 1998. (Cited on page 57.)

[99] M. Weling and G. E. Hinton. A new learning algorithm for mean field boltzmann machines. In *Proceedings of the International Conference on Artificial Neural Networks*, pages 351–357, 2002. (Cited on page 59.)

[100] A. Y. Ng and M. I. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in Neural Information Processing Systems*, 2002. (Cited on page 61.)

[101] I. Ulusoy and C. M. Bishop. Generative versus discriminative methods for object recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 258–265, 2005. (Cited on pages 61 and 85.)

[102] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. In *Proceedings of the International Conference on Pattern Recognition*, volume 3, pages 32–36, 2004. (Cited on page 65.)

[103] G. R. Bradski. Computer vision face tracking for use in a perceptual user interface. *Intel Technology Journal*, 1998. (Cited on page 66.)

[104] C. Stauffer and W. E. L. Grimson. Adaptive background mixture models for real-time tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 246–252, 1999. (Cited on page 68.)

[105] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 886–893, 2005. (Cited on page 68.)

[106] P. Felzenszwalb, D. Mcallester, and D. Ramanan. A discriminatively trained, multiscale, deformable part

model. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008. (Cited on page 68.)

[107] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the Computer Vision and Pattern Recognition*, pages 511–518, 2001. (Cited on page 68.)

[108] C. Thurau. Behavior histograms for action recognition and human detection. In *Workshop on Human Motion – Understanding, Modeling, Capture and Animation*, pages 299–312, 2007. (Cited on page 69.)

[109] R. Polana, R. Nelson, and A. Nelson. Low level recognition of human motion (or how to get your man without finding his body parts). In *Proceedings of IEEE Workshop on Motion of Non-Rigid and Articulated Objects*, pages 77–82, 1994. (Cited on page 69.)

[110] B. Leibe and B. Schiele. Analyzing appearance and contour based methods for object categorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 409–415, 2003. (Cited on page 69.)

[111] S. Belongie and J. Malik. Matching with shape contexts. In *Proceedings of the IEEE Workshop on Content-Based Access of Image and Video Libraries*, pages 20–26, 2000. (Cited on page 69.)

[112] G. Mori, S. Belongie, and J. Malik. Efficient shape matching using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(11):1832–1837, 2005. (Cited on page 69.)

[113] J. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6), 1986. (Cited on page 71.)

[114] A. A. Efros, A. C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *Proceedings of the IEEE*

International Conference on Computer Vision, volume 2, pages 726–733, 2003. (Cited on page 71.)

[115] J. L. Barron, D. J. Fleet, S. S. Beauchemin, and T. A. Burkitt. Performance of optical flow techniques. In *Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition*, pages 236–242, 1992. (Cited on page 71.)

[116] G. Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Proceedings of The Scandinavian Conference on Image Analysis*, pages 363–370, 2003. (Cited on page 71.)

[117] G. Farnebäck. *Polynomial Expansion for Orientation and Motion Estimation*. PhD thesis, Linköping University, Sweden, 2002. (Cited on pages 71 and 72.)

[118] R. Polana and A. Nelson. Recognizing activities. In *Proceedings of the International Conference on Pattern Recognition*, volume 1, pages 815–818, 1994. (Cited on page 71.)

[119] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the DARPA Image Understanding Workshop*, pages 121–130, 1981. (Cited on page 72.)

[120] D. J. Fleet and A. D. Jepson. Computation of component image velocity from local phase information. *International Journal of Computer Vision*, 5(1):77–104, 1990. (Cited on page 72.)

[121] M. J. Black and P. Anandan. The robust estimation of multiple motions: parametric and piecewise-smooth flow fields. *Computer Vision and Image Understanding*, 63(1):75–104, 1996. (Cited on page 72.)

[122] R. Szeliski and J. Coughlan. Spline-based image registration. *International Journal of Computer Vision*, 22(3):199–218, 1997. (Cited on page 72.)

[123] K. P. Murphy. The bayes net toolbox for matlab. *Computing Science and Statistics*, 33:2001, 2001. (Cited on page 75.)

[124] L. P. Morency, A. Quattoni, C. M. Christoudias, and S. Wang. *Hidden-state Conditional Random Field Library*, 2007. User Guide. (Cited on page 75.)

[125] C. S. Wallace and M. P. Georgeff. A general objective for inductive inference. Technical report, Department of Computer Science, Monash University, 1983. (Cited on page 76.)

[126] R. V. Babu, B. Anantharaman, K. R. Ramakrishnan, and S. H. Srinivasan. Compressed domain action classification using hmm. *Pattern Recognition Letters*, 23(11):1203–1213, 2002. (Cited on page 76.)

[127] R. Cutler and L. S. Davis. Robust real-time periodic motion detection, analysis, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):781–796, 2000. (Cited on page 77.)

[128] C. Curio, J. Edelbrunner, T. Kalinke, C. Tzomakas, and W. von Seelen. Walking pedestrian recognition. *IEEE Transactions on Intelligent Transportation Systems*, 1(3):155–163, 2000. (Cited on pages 77 and 90.)

[129] R. Goldenberg, R. Kimmel, E. Rivlin, and M. Rudzsky. Behavior classification by eigendecomposition of periodic motions. *Pattern Recognition*, 38(7):1033–1043, 2005. (Cited on page 80.)

[130] B. Heisele and C. Wohler. Motion-based recognition of pedestrians. In *Proceedings of The International Conference on Pattern Recognition*, volume 2, pages 1325–1330, 1998. (Cited on page 89.)

[131] Z. Ghahramani. Learning dynamic bayesian networks. *Adaptive Processing of Sequences and Data Structures, International Summer School on Neural Networks*, pages 168–197, 1997. (Cited on page 93.)

[132] Z. Wang and B. Li. Human activity encoding and recognition using low-level visual features. In *Proceedings of the International Joint Conference on Artificial Intelligence*, July 2009. (Cited on pages 102 and 103.)

[133] M. Lucena, N. Pérez de la Blanca, J. M. Fuertes, and M. J. Marín-Jiménez. Human action recognition using optical flow accumulated local histograms. In *Proceedings of the Iberian Conference on Pattern Recognition and Image Analysis*, pages 32–39, June 2009. (Cited on pages 102 and 103.)

[134] S. Savarese, A. DelPozo, J. C. Niebles, and L. Fei-Fei. Spatial-temporal correlatons for unsupervised action classification. In *Proceedings of the IEEE Workshop on Motion and Video Computing*, pages 1–8, 2008. (Cited on pages 102 and 103.)

[135] A. Fathi and G. Mori. Action recognition by learning mid-level motion features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 23–28, 2008. (Cited on pages 102 and 103.)

[136] M. Ahmad and S-W. Lee. Human action recognition using shape and clg-motion flow from multi-view image sequences. *Pattern Recognition*, 41(7):2237–2252, 2008. (Cited on pages 102 and 103.)

[137] J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision*, 79(3):299–318, 2008. (Cited on pages 102 and 103.)

[138] S. Nowozin, G. H. Bakir, and K. Tsuda. Discriminative subsequence mining for action classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1–8, 2007. (Cited on pages 102 and 103.)

[139] S. F. Wong and R. Cipolla. Extracting spatiotemporal interest points using global information. In *Procee-*

dings of the IEEE International Conference on Computer Vision, pages 1–8, 2007. (Cited on pages 102 and 103.)

[140] Z. M. Zhang, Y. Q. Hu, S. Chan, and L. T. Chia. Motion context: A new representation for human action recognition. In *Proceedings of the European Conference on Computer Vision*, pages 817–829, 2008. (Cited on pages 102 and 103.)

[141] K. Mikolajczyk and H. Uemura. Action recognition with motion-appearance vocabulary forest. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 0, pages 1–8, 2008. (Cited on pages 102 and 103.)

[142] J. Liu and M. Shah. Learning human actions via information maximization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 0, pages 1–8, 2008. (Cited on pages 102 and 103.)

[143] T. K. Kim, S. F. Wong, and R. Cipolla. Tensor canonical correlation analysis for action classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007. (Cited on pages 102 and 103.)

[144] Y. Wang, P. Sabzmeydani, and G. Mori. Semi-latent dirichlet allocation: A hierarchical model for human action recognition. In *Proceedings of the Workshop on Human Motion Understanding, Modeling, Capture and Animation*, pages 240–254, 2007. (Cited on pages 102 and 103.)

[145] K. Schindler and L. J. van Gool. Combining densely sampled form and motion for human action recognition. In *Proceedings of the German Association for Pattern Recognition (DAGM) Annual Symposium*, pages 122–131, 2008. (Cited on page 102.)

[146] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Proceedings of the IEEE Conference on In Computer Vision*

and Pattern Recognition, pages 1–8, 2008. (Cited on pages 102 and 103.)

[147] H. Jhuang, T. Serre, L. Wolf, and T. Poggio. A biologically inspired system for action recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1–8, 2007. (Cited on pages 102 and 103.)

[148] Y. Wang, Z. Q. Liu, and L. Z. Zhou. Learning hierarchical non-parametric hidden markov model of human motion. In *Proceedings of International Conference on Machine Learning and Cybernetics*, volume 6, pages 3315–3320, 2005. (Cited on page 102.)

[149] M. J. Marín-Jiménez, N. Pérez de la Blanca, M. A. Mendoza, M. Lucena, and J. M. Fuertes. Learning action descriptors for recognition. In *Proceedings of the Workshop on Image Analysis for Multimedia Interactive Services*, pages 5–8, May 2009. (Cited on pages 102 and 103.)