Representación, acceso y descripción de vídeo para futuros servicios multimedia*

M. Pardàs, J.R. Casas, A. Gasull, F. Marqués, A. Oliveras, P. Salembier, E. Sayrol, C. Chang, D. Comas, C. Ferran, X. Giró, J.L. Landabaso, M. León, J.R. Morros, J. Ruiz, J. Solé, V. Vilaplana

Departament de Teoria del Senyal i Comunicacions, Universitat Politècnica de Catalunya,
e-mail: { montse, josep, toni, ferran, albert, philippe, elisa, camilo, comas, cferran, xgiro, jl, mleon, morros, jrh, joel, verónica}@gps.tsc.upc.es

***Abstract*—En este artículo se presenta la investigación que el grupo "Processament d'Imatge" del Departament de Teoria del Senyal i Comunicacions de la UPC está llevando a cabo en la actualidad en el área de la representación, el acceso y la descripción de vídeo orientado a futuros servicios multimedia.**

I. INTRODUCCIÓN

Históricamente, los aspectos de representación (funcionalidad de compresión) e identificación (funcionalidades de indexación, búsqueda, navegación, localización) del contenido multimedia se han considerado de forma independiente. Sin embargo, para la mayoría de los futuros servicios multimedia, los conceptos de representación e identificación del contenido se tendrán que considerar dentro de un marco unificado. El ritmo de crecimiento del volumen de información disponible es tal que la identificación del contenido deberá tenerse en cuenta en la representación.

En este artículo presentamos el desarrollo de nuevas herramientas y técnicas para crear y manipular una representación única que dé soporte al mismo tiempo a la representación (funcionalidad de compresión) y a la identificación (funcionalidades de indexación, búsqueda, navegación, localización) del contenido en secuencias de vídeo. El trabajo se organiza en dos líneas de investigación complementarias que dan lugar a diversos objetivos:

1) Una línea de carácter fundamental y básico: Desarrollo de herramientas teóricas y básicas. Su finalidad principal es mejorar y extender el conjunto de herramientas disponibles para definir algoritmos adaptados al procesado de las secuencias de vídeo.

2) Una línea de carácter más aplicado que se apoya sobre los resultados de la primera línea: Desarrollo de nuevos algoritmos. Aquí se han identificado dos retos importantes: 1) crear un puente entre el análisis de bajo nivel de la señal y su interpretación semántica y 2) definir un marco común para la representación y gestión de la información vídeo.

* Este trabajo se ha realizado en el contexto del proyecto TIC2001-0996 del MCyT

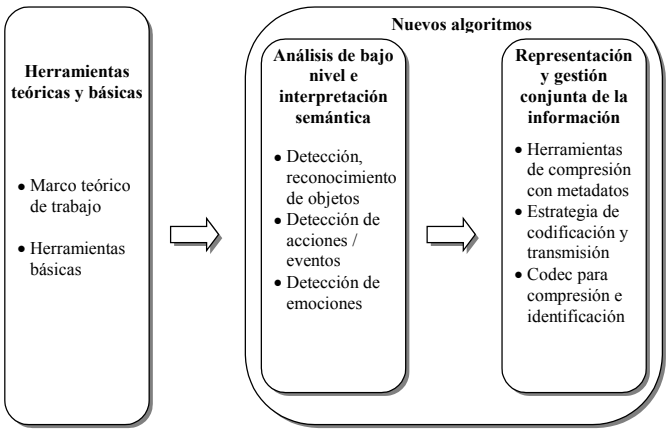

Figura 1: Esquema de las diferentes líneas de investigación que se van a presentar.

II. DESARROLLO DE HERRAMIENTAS TEÓRICAS Y BÁSICAS

Gran parte de las herramientas de análisis que utilizamos se basan en operadores conexos, teoría de grafos y/o técnicas avanzadas de estimación de movimiento y trayectorias y/o modelos ocultos de Markov. En esta sección se han desarrollado conceptos y técnicas relacionadas con estas herramientas. Además, también se ha pretendido extender el marco de trabajo anterior para dar respuesta a nuevas necesidades.

A. Extensión y desarrollo del marco de trabajo

La necesidad de nuevas técnicas de modelado está relacionada con el número elevado de nuevos servicios multimedia. De forma tradicional, las imágenes se representan como una tabla rectangular de píxeles y las secuencies de video se interpretan como un flujo continuo de imágenes discretas. Los nuevos servicios multimedia se apoyan sobre una representación más cercana al mundo

real. Las técnicas de compresión y de indexación basadas en el contenido son dos ejemplos típicos donde nuevas herramientas de modelado y procesado son necesarias. En estos ejemplos, el modelado tiene que reflejar, como mínimo parcialmente, el proceso de creación visual: la imagen se crea a partir de la proyección de una escena compuesta de objetos 3D sobre un plano bidimensional. En estos casos, la idea de región (temporal, espacial o espacio-temporal) juega un papel fundamental.

A parte del problema de modelado, podemos constatar que la inmensa mayoría de las herramientas de procesado de imagen no se adaptan bien al procesado orientado a regiones. Por ejemplo, los filtros de bajo nivel tienen en general una estrecha relación con la representación en píxeles de las imágenes: convolución lineal a partir de una respuesta impulsional, filtro de mediana, filtros morfológicos basados en erosión y dilatación con un elemento estructurante, etc. En todos estos casos, la estrategia de procesado consiste en modificar el valor individual de los píxeles a partir de una función de los píxeles vecinos dentro de una ventana. No se adaptan bien a un procesado al nivel de la región.

Los primeros ejemplos de procesado basado en regiones se pueden encontrar en el campo de la segmentación. De forma más reciente, un subconjunto de filtros morfológicos, denominados operadores conexos, ha tenido un desarrollo importante. Los operadores conexos son herramientas de procesado basados en la noción de región porque no modifican los valores individuales de los píxeles sino que interactúan al nivel de las zonas donde la señal es constante. Estas zonas son las zonas planas. De forma intuitiva, los operadores conexos pueden eliminar las fronteras entre las zonas planas pero no pueden crear o desplazar los contornos. En nuestro grupo estamos analizando varias formas de estructurar las zonas planas para poder actuar de forma eficiente sobre ella. Una respuesta que nos parece particularmente interesante es la técnica basada en árboles binarios de particiones.
El árbol representa un conjunto de regiones que se pueden extraer de la imagen (ver Figura 2). Son todas zonas planas o uniones de zonas planas. Las hojas del árbol representan las zonas planas iniciales y los demás nodos representan regiones obtenidas por la fusión de varias regiones. El árbol se puede considerar como una representación multi-escala de un amplio conjunto de regiones. Las regiones grandes aparecen cerca de la raíz del árbol y los detalles cerca de las hojas.

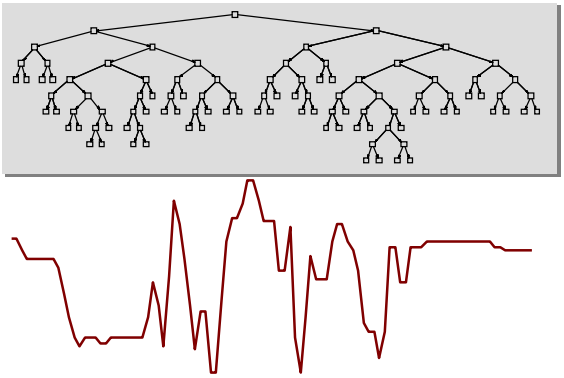

Figura 2: Árbol binario de partición (ejemplo para una señal monodimensional)

El árbol de particiones tiene que crearse de tal forma que las regiones más relevantes para la aplicación estén representadas. Estamos analizando y estudiando un amplio abanico de criterios de creación del árbol (criterio de homogeneidad: color, textura, movimiento, etc.) así como de técnicas de filtrado de esta representación (criterio de filtrado: tamaño, contrate, verosimilitud de presencia de un objeto, etc.). Tal como se refleja en la Figura 3, las técnicas de filtrado se basan en una poda del árbol. El procesado global tiene tres pasos: Creación del árbol, poda del árbol y restitución de una imagen.

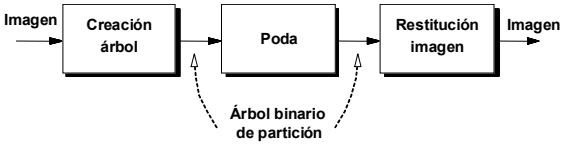

Figura 3: Estrategia de filtrado con árboles de partición.

B. Operadores Conexos

Como se ha comentado anteriormente, los operadores conexos son operadores morfológicos que interactúan con las zonas planas de la imagen. Este tipo de operadores permite pasar del estudio de la secuencia a nivel del píxel al nivel de región. Una manera de implementar estos filtros es mediante la estructura de árbol binario de particiones. En este punto se ha tratado de mejorar la estructura de árbol binario de particiones para poder actuar de manera eficiente sobre las zonas planas (regiones) que definen la imagen. Se han desarrollado técnicas que permiten flexibilizar y extender dicha estructura. Para ello, se ha definido el concepto de nodo extendido que permite analizar características de una unión de regiones vecinas, aunque esta unión no esté directamente representada en la estructura de árbol.

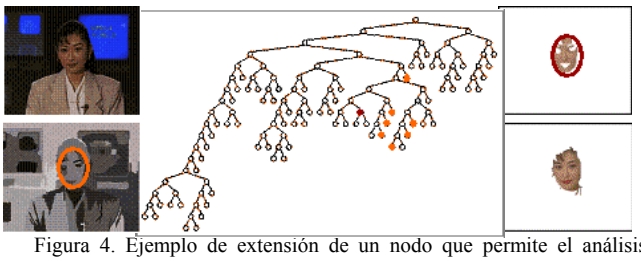

Figura 4. Ejemplo de extensión de un nodo que permite el análisis completo de un objeto en la imagen. La región marcada en granate es extendida utilizando información de la forma del objeto (elipse) hasta crear una nueva región que contiene totalmente la cara, aunque ésta no está presente en el árbol

C. Retículo condicional: introducción de la profundidad

A parte de desarrollar técnicas que trabajan directamente sobre la idea de regiones se está analizando a su vez un nuevo marco para la morfología matemática en presencia de información de profundidad. Este tipo de información no está presente para aplicaciones tipo televisión o video-teléfono, pero en el futuro un número muy amplio de aplicaciones implicarán la visualización de la escena por un conjunto de cámaras. En este caso, técnicas similares a la estimación de movimiento son útiles para estimar la distancia a la cámara de cada píxel o de cada zona plana de la imagen.

En marco de trabajo clásico de la morfología matemática es el retículo binario o de funciones. El primero se utiliza para imágenes binarias y el segundo para imágenes a nivel de gris. La creación de un retículo se apoya sobre una relación de orden y dos operadores: Sup e Inf. Esta relación de orden y los dos operadores se relacionan directamente con la idea que tenemos de los objetos pertinentes para la aplicación. Por ejemplo, en el retículo de funciones el orden es el nivel de gris de los píxeles. En consecuencia, las componentes de interés son los máximos y mínimos. Los operadores resultantes tendrán la capacidad de procesar los máximos y/o mínimos.

Aquí también, pretendemos representar de forma más fiel el proceso de creación de las imágenes. En lugar de utilizar la idea de región como en el apartado anterior, queremos crear un retículo incluyendo la información de distancia de los píxeles (o regiones) a la camera. Esta información es fundamental porque la mayoría de los objetos son opacos y lo que observamos es la proyección del objeto más próximo a la cámara de manera independiente a su nivel de gris.

Una vez estimada la profundidad de cada píxel (o zona plana) de la imagen, podemos desarrollar un retículo "condicional" capaz de tratar la información de profundidad y de nivel de gris. Podemos destacar que este estudio no es similar a los trabajos sobre morfología multidimensional donde cada una de las dimensiones juega un papel más o menos similar. Aquí, la información de distancia predomina sobre la de nivel gris. En consecuencia, el retículo condicional utiliza un orden léxicográfico entre la profundidad y el nivel de gris.

D. Wavelets no lineales

A su vez, se está estudiando el uso de wavelets no lineales en la creación de dichos operadores las cuales permiten una descomposición rápida y eficiente de la imagen. El marco clásico de la morfología matemática está basado en la representación en píxeles de la imagen, requiriendo una extensión para poder incluir el concepto de regiones o, a más alto nivel, el de objetos. Una extensión natural es vía descomposición multi-resolución para tener así un modelo jerárquico, al estilo Árbol binario de partición, muy útil para los distintos procesados posteriores.

Figura 5: Ejemplo de descomposición de la imagen Lenna mediante wavelets morfológicos

Siguiendo la idea de representación multi-escala, se han desarrollado estrategias de análisis mediante wavelets morfológicos que permiten una descomposición rápida y eficiente de la imagen. Además, permiten una representación a distintos niveles de resolución, de manera que las imágenes quedan divididas explícitamente por contornos. Por su parte, los detalles, que son almacenados a otro nivel, posibilitan la segmentación de regiones para la extracción de objetos o la eliminación de detalles o texturas irrelevantes (sino perniciosos) para su detección y reconocimiento. La diferencia principal con el Árbol binario de partición es la noción de escala. En el caso de los Árboles binarios de partición la noción de escala es geométrica (unión de regiones). En el caso de wavelets morfológicas, la noción de escala se apoya sobre las propiedades espacio-frecuencial de los píxeles.

Finalmente, se está estudiando también el caso de wavelets morfológicas adaptativas que consiguen al mismo tiempo tasas de compresión elevadas, a vistas de su inclusión en el codificador/descodificador de compresión.

E. Teoría de Grafos

Para la descripción de los objetos complejos presentes en una secuencia a partir de las regiones anteriormente ob-

tenidas, se ha optado por explotar la teoría de grafos. Los objetos son entidades semánticas por definición, a menudo compuestos de distintas partes visualmente distinguibles. Ello compromete la efectividad de los algoritmos de segmentación cuando se aplican directamente a la detección y extracción de objetos complejos y estructurados. Para la detección de objetos complejos es necesario considerar ciertas propiedades (compacidad, adyacencia, regularidad, inclusión, quasi-inclusión, simetría axial, central y quasi-simetría) que definen su estructura, reflejando la naturaleza física del objeto en la señal visual.

Las propiedades anteriores las denominamos sintácticas, en contraposición a las propiedades que se derivan de los objetos como instancias de conceptos semánticos. Al igual que los criterios habitualmente considerados para la segmentación (homogeneidad en textura espacial o en el movimiento de los objetos), las propiedades sintácticas no son específicas de un tipo determinado de objetos. Se pueden emplear de modo genérico para cualquier objeto complejo natural o artificial, y no tienen el inconveniente de restringir el dominio de aplicación como en el caso del empleo de modelos de objetos específicos.

Las propiedades sintácticas se evalúan mediante técnicas de análisis sintáctico o estructural, basadas en formas y las configuraciones espaciales de las distintas partes de los objetos. Forma y estructura son difícilmente evaluables directamente a partir de los píxeles como criterios de segmentación. Por ello, el análisis estructural se lleva a cabo a partir de regiones simples de textura homogénea obtenidas en una segmentación inicial de la imagen (sobre-segmentación), mediante la medida de las distintas propiedades estructurales (regularidad, simetría, inclusión...).

Figura 6: Análisis estructural mediante simetría, compacidad, regularidad e inclusión a partir de a) una partición inicial de 57 regiones que se estructuran progresivamente en, b) 24, c) 8, d) 3 y e) 2 regiones

En el ejemplo de la Figura 6, ciertas propiedades estructurales permiten la diferenciación de los objetos en una partición inicial. La partición inicial de la imagen se estructura progresivamente en conjuntos de regiones quasi-simétricas o que resultan parcialmente incluidas unas en otras. En el tercer paso del algoritmo de estructuración (figuras c y d), se puede observar como la propiedad de inclusión permite agrupar las regiones 1 y 2 con el fondo, y la propiedad de simetría parcial con respecto a un elemento central permite la agrupación de las regiones 4 y 5 con la región 3 (elemento central). La información de textura y la fuerza de la transición entre dos regiones también se considera en cada paso de estructuración. Ello explica porqué las dos últimas regiones inferiores son agrupadas con por inclusión en el fondo y no por simetría con un elemento central (3+4+5).

La teoría de grafos ha permitido también definir modelos de entidades semánticas a partir de los denominados grafos de descripción. En un grafo de descripción, las entidades semánticas y las relaciones se asignan a los vértices del grafo para modelar una entidad semántica de nivel semántico superior. Los documentos multimedia contienen diversas instancias de entidades semánticas que pueden ser detectadas a partir de un modelo definido por un grafo de descripción. La similitud semántica de la instancia respecto al modelo se estima a partir del cálculo de un valor de confianza c(SE). Para ello, se ha desarrollado una expresión de cálculo de la c(SE) que combina los valores de relevancia (R) proporcionados por el modelo con los valores de confianza (c) calculados a partir las instancias:

$$c(se) = \frac{\sum_{i=1}^{M} R_i c_i}{\sum_{i=1}^{M} R_i} = \frac{\sum_{i=1}^{N} R_i c_i + \sum_{i=N+1}^{M} R_i c_i}{\sum_{i=1}^{N} R_i + \sum_{i=N+1}^{M} R_i}$$

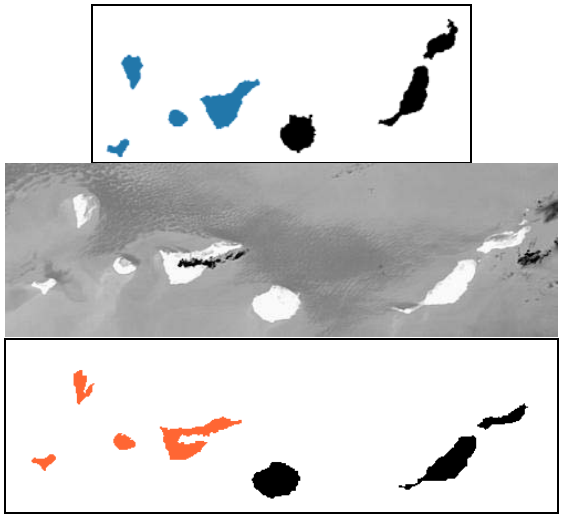

Figura 7: Ejemplo de reconocimiento del conjunto de islas que forman la provincia de Tenerife (en naranja en la imagen) en una imagen con distorsión geométrica y oclusiones debidas a nubes. La imagen superior presenta el modelo utilizado.

Los vértices de los grafos se etiquetan como necesarios u opcionales, de tal manera que la expresión de cálculo de c(SE) tan sólo tiene en cuenta aquellos vértices opciona-

les cuya confianza aumenta el valor final de c(SE). Este tipo de análisis se ha aplicado a ámbitos distintos como la detección de caras en escenarios no controlados o el reconocimiento de archipiélagos concretos en imágenes de teledetección, como se ilustra en la Figura 7.

F. Estimación de movimiento y trayectoria

La detección de eventos requiere la estimación robusta de la trayectoria de los objetos en la escena. Para la estimación de movimiento y trayectoria, se han desarrollado diversas técnicas destinadas a mejorar la robustez y eficiencia en la estimación del movimiento. La mayoría de los esquemas de estimación de movimiento adopta un modelo de movimiento, el cual relaciona la intensidad de la imagen con los parámetros de movimiento deseados, a través de algún criterio de estimación tal como el de error cuadrático medio (MSE). El problema central consiste en la búsqueda de una solución para dicho criterio, con el objetivo de determinar los vectores de movimiento (flujo óptico) que mejor relacionen la posición de los píxeles en tramas de imágenes sucesivas. Muchas soluciones a este problema ya han sido propuestas e incluyen ya sea métodos de correlación, o bien basados en cálculo de gradiente así como también otros enfoques estocásticos. Todos estos métodos de estimación, sin embargo, enfrentan una dificultad en común la cual consiste en el mal condicionamiento de la medición del flujo óptico y, adicionalmente, la propagación de estimaciones erróneas

Para estos efectos nos hemos concentrado en técnicas basadas en cálculo de gradiente. Estas técnicas pueden proveer, con rapidez y precisión, información sobre el flujo óptico denso con una definición a nivel sub-pixel. Además, esta información puede servir de entrada para aplicaciones de más alto nivel. Estas técnicas han sido implementadas en un esquema con resolución múltiple, permitiendo así una mayor flexibilidad en la búsqueda del movimiento, a través de los diversos niveles de resolución espacial.

Por medio del seguimiento de la diferencia media absoluta entre tramas sucesivas, ha sido posible detectar grandes diferencias en la intensidad de la imagen (la mayoría de las veces debido a casos particulares de oclusión, cambios de iluminación o grandes movimientos) descartando de esta manera estimaciones dudosas. Otra evaluación de la calidad de estimación de movimiento se ha realizado analizando la información de gradiente espacial. Regiones de mucha textura, así como regiones que contienen esquinas o una elevada frecuencia espacial, producen sistemas bien condicionados, los cuales están caracterizados por autovalores grandes y resultados en consecuencia confiables. En general, una vez que las falsas estimaciones son descartadas, éstas pueden ser correctamente reemplazadas, utilizando un filtrado de mediana sobre las estimaciones vecinas. Finalmente, la utilización de ventanas temporales extendidas, basadas en hipótesis de continuidad del movimiento, se han propuesto para proveer información sobre el gradiente temporal, obteniendo así una mejoría en la estimación de movimiento.

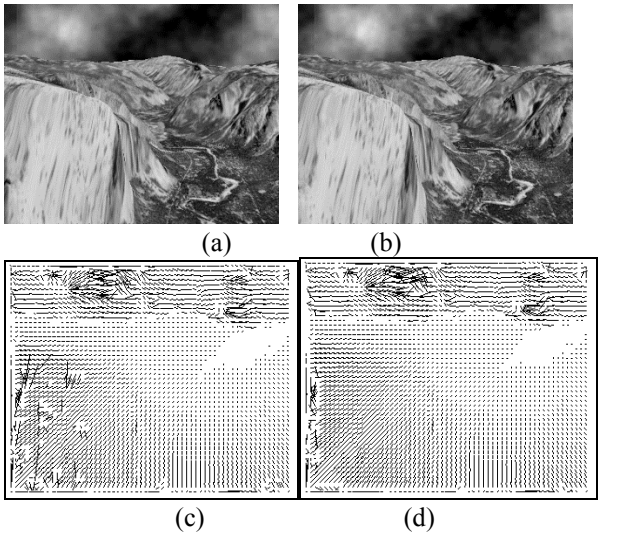

Figura 8: Ejemplo de campo denso de movimiento obtenido mediante la utilización de técnicas robustas.
(a) y (b) Cuadros originales #7 y #9 de la secuencia sintética Yosemite, respectivamente.
(c) y (d) Campos de movimiento usando 1 y 3 niveles jerárquicos, respectivamente

G. Modelos ocultos de Markov

Los modelos ocultos de Markov (HMM), herramienta muy utilizada en el procesado de voz, permiten clasificar señales con evolución temporal y, por tanto, pueden ser útiles para la detección y clasificación de escenas. En nuestro grupo trabajamos en la adaptación al caso de procesado de vídeo analizando básicamente los aspectos relacionados con el diseño de su topología y su entrenamiento.

En el diseño de la topología de los HMM se ha tenido en cuenta que, mientras que en las aplicaciones de reconocimiento del habla la topología de los HMM es normalmente del tipo izquierda-derecha, ya que permite caracterizar señales que evolucionan en el tiempo de manera sucesiva, no podemos generalizar este comportamiento para las evoluciones que se pretende caracterizar en análisis de secuencias de vídeo. Así por ejemplo, se ha podido comprobar que para describir la evolución temporal de las expresiones faciales en secuencias de imágenes es más versátil utilizar modelos que permitan transiciones en los dos sentidos (izquierda-derecha y derecha-izquierda).

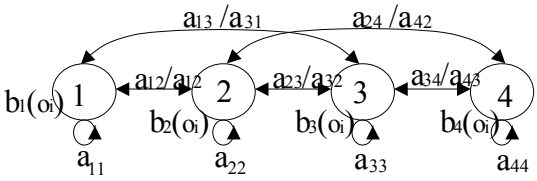

Figura 9: Topología del modelo utilizado en análisis de vídeo

En el entrenamiento de los modelos, y dado que las bases de datos de entrenamiento de que disponemos son reducidas, se ha utilizado Modelos Ocultos de Markov Semicontinuos. Además, utilizando información a priori sobre el comportamiento temporal de los eventos que queremos caracterizar es posible entrenar los nuevos modelos sin requerir la grabación de nuevas bases de datos. Es decir, se puede aplicar esta información a las bases de datos disponibles para crear bases de datos extendidas en las que se disponga de los diferentes tipos de evoluciones que se prevén posibles.

III. ANÁLISIS DE BAJO NIVEL E INTERPRETACIÓN SEMÁNTICA

En esta línea desarrollamos un conjunto de algoritmos que, basándose en las herramientas descritas en la línea anterior, permiten analizar las secuencias y detectar la presencia de entidades semánticas que ayuden a su indexación y codificación.

A. Detección de caras humanas

La detección de caras humanas en la escena se ha estudiado mediante dos técnicas. La primera ha contemplado la descripción de un objeto mediante sus componentes más simples y las relaciones semánticas entre ellas. La segunda se ha basado en un esquema de segmentación no supervisada del espacio de color YUV. Básicamente, la primera técnica se caracteriza por su mayor precisión (mejor localización de la(s) cara(s) en la imagen y menor tasa de error) mientras que la segunda técnica presenta una mayor simplicidad de cálculo, sin un gran detrimento de la calidad de los resultados. Por tanto, ambas técnicas son válidas, dependiendo de la aplicación final.

La primera técnica se basa en la utilización de operadores conexos y, en ella, la verosimilitud de cara de una región se estima asumiendo una distribución Gaussiana de la clase y utilizando el análisis de componentes principales (PCA) en la construcción del estimador. Esta técnica ha sido mejorada introduciendo información sobre la forma del objeto buscado. Se ha extendido la utilización de la PCA con dos modelos de forma del objeto cara: un modelo muy simple, elíptico (EPCA), y otro más complejo basado en el uso de la PCA ponderada (WPCA). Ambas extensiones aumentan notablemente la robustez del método. Como resultado de estas extensiones, se ha conseguido mejorar notablemente el sistema, aumentando la tasa de detecciones correctas y disminuyendo el número de falsas alarmas. Para la caracterización de la clase de las caras se utiliza la base de datos XM2VTS.

La segunda técnica se basa en un esquema de segmentación no supervisada del espacio de color YUV. Se ha dotado el esquema de gran robustez a través de la utilización de sistemas multiresolución en la etapa de normalización de la cara, que contribuyen a elegir el mejor tamaño de la cara a normalizar. A su vez, se ha incorporado también técnicas de localización de ojos, nariz y boca que confieren al sistema una robustez adicional.

Figura 10: Ejemplos de detección de caras

B. Detección de texto

En algunos casos, la información textual presente en la escena aporta un gran conocimiento semántico de la misma. Por tanto, es necesario estudiar algoritmos de detección de texto. Para ello, se ha analizado las técnicas actuales, clasificándolas en dos grupos: métodos en el dominio transformado y métodos en el dominio espacial. El primer grupo sirve para localizar las zonas en la imagen con alta probabilidad de contener texto. Esta aproximación se vislumbra como una herramienta de pre-procesado excelente y para este fin se está estudiando tanto la DCT como la transformada Wavelet.

Figura 12: Ejemplo de análisis mediante técnicas de *wavelets* para la extracción de texto

El segundo grupo se basa en algoritmos de análisis espacial de texturas, contornos o componentes conexas, entre otros. En el ámbito de la creación de nuevos operadores conexos, se está investigando la robustez de distintos parámetros extraídos de las componentes conexas. Se ha comprobado que no existe un único parámetro que sea suficientemente robusto a variaciones de la fuente y del tamaño del texto. Sin embargo, hay indicios de que el parámetro de complejidad junto a un filtro de contraste será la base que permita una buena detección de las regiones de texto sobre cualquier tipo de entorno. Para este trabajo de estudio de robustez de los parámetros se ha generado una extensa base de datos. Ahora se está analizando cuáles de estas características definen mejor el texto y con qué grado de fiabilidad lo hacen.

C. Detección de objetos modelados

El modelado de objetos es una estrategia muy útil para detectar su presencia en la escena. Para la detección de objetos modelados se ha analizado diversas técnicas. Inicialmente, se ha dividido el problema en modelos simples (incluyendo un único objeto básico) y modelos complejos (donde el objeto a detectar está compuesto por varios objetos básicos).

Un modelo simple que se ha usado se basa en la forma del objeto. Para explotar este tipo de modelo se ha desarrollado una técnica que compara versiones transformadas del modelo con los contornos obtenidos de la segmentación de la imagen. La comparación se basa en una aproximación morfológica de la distancia de Chamfer. Así se compara un mapa de distancias que mide la distancia euclídea entre cada uno de los puntos de la imagen y los contornos de la partición. El contorno de referencia puede ser representado como una curva parametrizada o no parametrizada. En el caso de las curvas parametrizadas, las posibles transformaciones vienen dadas por la propia parametrización, mientras que para curvas no parametrizadas, se reducen a rotación, escalado y traslación. Los resultados muestran que el empleo de la función distancia permite suavizar la información de contornos, proporcionando detecciones correctas incluso cuando la información de contorno no es completa o no coincide exactamente con el contorno de referencia. Además, el uso de segmentación para calcular los contornos presenta ventajas con respecto a las técnicas clásicas de detección de contornos, pues permite un análisis más global de la imagen. De esta forma, se pueden extraer los contornos más relevantes incluso cuando se encuentran presentes objetos con mucha textura. Esta técnica se ha aplicado a la detección de caras humanas (modelo de elipse) así como a otros estructuras sencillas (estrellas de mar, círculos,…).

También se ha desarrollado un segundo modelo que combina información de homogeneidad de textura y de las características del contorno (regular, continuo y cerrado). Estas informaciones han sido combinadas usando una técnica de propagación de snakes implementadas mediante curvas de nivel. Esta técnica se ha aplicado a la detección de figuras humanas en escenas complejas. El proceso se divide en dos etapas: primero se realiza una segmentación en la imagen inicial y a continuación se realiza un seguimiento del objeto hallado en las imágenes sucesivas. La primera etapa comienza con un modelado estadístico de los objetos presentas en la imagen. En particular, se han buscado objetos cuya luminancia sea homogénea aunque no sean necesariamente conexos. Por ejemplo, la topología humana presenta contornos regulares, continuos y cerrados. Por tanto, se ha optado por modelarlos con snakes que permiten modelar bien estas propiedades.

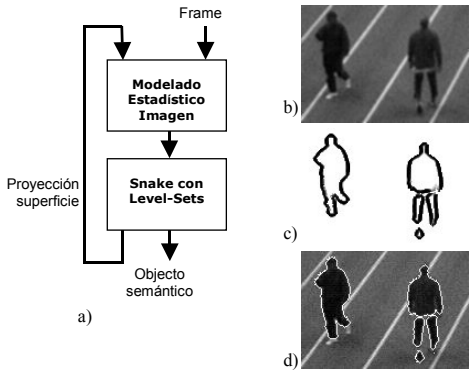

Figura 13: (a) Esquema del sistema, (b) imagen original, (c) contorno obtenido a partir del modelo y (d) segmentación final de los cuerpos humanos

Por otra parte, con el objetivo de permitir cambios en la topología de los objetos segmentados, se ha realizado una implementación en level-sets. Dicho método interpreta el snake como la curva de nivel cero de una superficie. Para aumentar la robustez del algoritmo, la superficie se propaga controlada por un término asociado a la información de contorno y otro asociado a la información de textura del modelo del objeto. El seguimiento del objeto a lo largo del tiempo se hace proyectando la superficie final a la imagen siguiente y propagándola de nuevo.

También se ha desarrollado una técnica específica para la aplicación de vigilancia. Este trabajo consiste en el diseño de un sistema robusto para la detección y seguimiento temporal de objetos con vistas a la automatización de los sistemas de video vigilancia. Las tareas desarrolladas incluyen aprendizaje adaptativo del fondo de

escena, segmentación de los objetos de primer plano, eliminación de sombras proyectadas y extracción de la representación de cada objeto en forma de velocidad, proporciones, color y tamaño para un seguimiento efectivo. El sistema desarrollado es capaz de etiquetar cada objeto, seguirlo y clasificarlo en persona/vehículo incluso bajo oclusiones. Los resultados se han obtenido tanto para secuencias propias como para las secuencias estandarizadas PETS 2001.

Figura 14: Ejemplo de la segmentación y eliminación de sombras

Figura 15: Ejemplos del seguimiento en una escena compleja con oclusiones

Finalmente, para tener en cuenta la posible complejidad de los objetos, se ha estudiado técnicas basadas en grafos de descripción. En estos grafos se han definido nodos necesarios y nodos opcionales y, en base a ellos, se ha definido una función de confianza que combina de manera no lineal las confianzas en la detección de cada uno de los elementos que forman el grafo de descripción. Como se ha comentado anteriormente, este sistema se ha utilizado en ámbitos tan distintos como la detección de caras frontales o archipiélagos de islas. De esta manera se ha probado la flexibilidad de los modelos en dos problemas visualmente dispares:

- En el caso de caras frontales, la expresión de C(SE) se aplica a los valores de confianza proporcionados por un algoritmo independiente encargado de buscar por separado los diferentes rasgos faciales que componen una cara. A partir de dichos valores, la expresión de C(SE) proporciona un valor global para la entidad completa "cara frontal". De esta manera es posible obtener una lista ordenada de candidatos a cara frontal presentes en una imagen.
- Los DGs se han utilizado también para la detección y etiquetado de un conjunto de islas en imágenes de teledetección a partir de los descriptores MPEG-7 de forma y posición relativa de las islas. El algoritmo permite superar los problemas de oclusión debidos a la presencia de nubes, gracias a la información adicional que proporciona la posición relativa de las islas.

D. Detección de acciones simples de un ser humano

Para detectar acciones simples de un ser humano se ha desarrollado una técnica basada en el análisis de la silueta de la persona mediante técnicas de distancia geodésica y de correspondencia entre grafos por si se dispone de varias cámaras y se desea obtener la información tridimensional. Así, se detecta y etiqueta los puntos críticos de la silueta (cabeza, manos y brazos), lo que lleva a crear un modelo sencillo del cuerpo humano que permite analizar acciones simples.

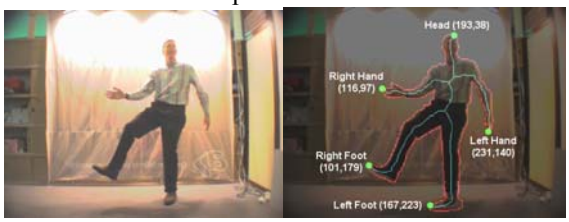

Figura 16: Ejemplo de análisis de la silueta humana y detección de los puntos críticos

Además, también se ha trabajado en el seguimiento específico de los brazos de una persona. Este seguimiento debe facilitar el reconocimiento de acciones simples (saludos) o más complejas (el lenguaje de signos). Las herramientas que se han desarrollado para la realización de este trabajo incluyen la creación de un modelo en tres dimensiones de la estructura de los brazos mediante un modelo kinemático de cadenas. Este modelo 3D es proyectado sobre un plano 2D para realizar el seguimiento de los brazos a partir de una posición conocida. A partir de este seguimiento, será necesaria la realización de modelos de movimiento para clasificar las diferentes acciones que pueden ser detectadas.

E. Clasificación de escenas

Con la clasificación de escenas se pretende asignar de manera automática una etiqueta caracterizando el género de la secuencia. Para ello, se ha extendido el concepto de árbol binario de particiones al caso de actuar en secuencias, tomando cada imagen como el elemento mínimo de la señal a analizar. Una vez segmentada la secuencia, se ha usado árboles de decisión, formulándose el problema como N clasificaciones binarias (una para cada género posible) seguido de una decisión posterior global. En esta clasificación, se ha utilizado la estructura jerárquica de índice o tabla de contenidos de la secuencia para mejorar la fiabilidad de la clasificación. Así se ha obtenido una clasificación multi-escala del género local de la secuencia. Las prestaciones actuales de este sistema son del orden de 95% de éxito.

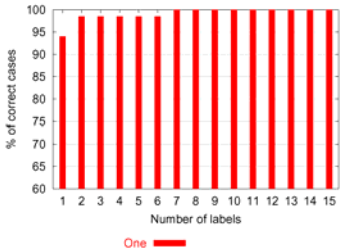

Figure 17: Tasa de éxito de determinación del género dominante de una secuencia. El eje horizontal representa el número de géneros diferentes asignados y el eje vertical representa la tasa de éxito.

La Figura 18 ilustra como el género local puede ir evolucionando a lo largo de la secuencia. El Árbol de Partición Binaria representa la tabla de contenido de la secuencia. La secuencia completa corresponde al nodo raíz del árbol. Los demás nodos representan los capítulos, secciones, sub-secciones, etc. al mismo modo que la tabla de contenido de un libro. Los nodos verdes (gris claro) del primer árbol representan las partes de la secuencia con el género de Acción. Los nodos rojos (gris oscuro) corresponden a partes de la secuencia con el género de No Acción. De la misma forma el segundo árbol indica la sub-partes de la secuencia con genero de Documental (Verde: gris claro) y de No Documental (Rojo: gris oscuro).

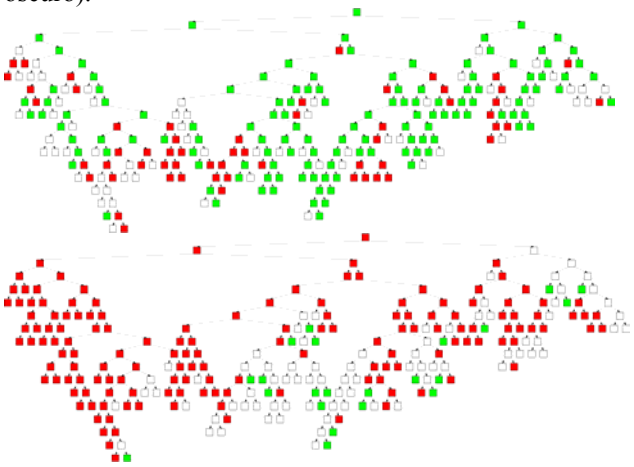

Figure 18: Ejemplo de propagación de la clasificación temática.

F. Reconocimiento de expresiones faciales y emociones

Se ha trabajado también ern el desarrollo de un sistema de reconocimiento de expresiones faciales basado en secuencias de video. Se ha partido de un sistema basado en HMM. Inicialmente se desarrolló para su aplicación al reconocimiento de expresiones faciales en intervalos de video previamente marcados. Es decir, se debe conocer a priori que la secuencia que se va a analizar representa una expresión facial y que se parte del estado facial neutro hasta evolucionar a un máximo de la expresión. De hecho, las referencias que se han encontrado sobre este tema siempre trabajan de este modo y por tanto no pueden aplicarse en un entorno real. El sistema que se ha desarrollado en nuestro grupo utiliza como características a analizar los parámetros estandarizados en MPEG-4 para la animación de caras, FAP (Face Animation Parameters). Estos parámetros se extraen de la secuencia de video utilizando técnicas automáticas de análisis de imagen y a continuación se utilizan en un reconocedor basado en HMM. Para el entrenamiento de los HMM se ha utilizado la base de datos de Cohn-Kanade.

Para poder aplicar el sistema de reconocimiento en un entorno real, se ha diseñado un sistema que permite su aplicación en largas secuencias de video a través de la extracción de forma eventual o continua de porciones de estas secuencias. Para ello se ha trabajado en dos líneas principales: el diseño de una topología para los HMM adaptada a este nuevo contexto y la implementación de un proceso de rechazo de secuencias que permita rechazar aquéllas que no pueden clasificarse como pertenecientes a ninguna expresión (por ejemplo, cuando la persona está hablando y por tanto sus movimientos faciales no corresponden a una expresión sino a la articulación de las palabras).

Los modelos utilizados en el reconocimiento de emociones sufren una evolución típica desde el estado neutro al máximo de la emoción. Para ello se utilizan HMM con una estructura izquierda-derecha. Sin embargo, si el algoritmo de reconocimiento se aplica a una parte de la secuencia elegida de manera aleatoria, podemos encontrar diferentes comportamientos: neutro a máximo, máximo a neutro, de neutro a máximo y regreso a neutro, o evoluciones que no llegan o no parten del máximo. Para este tipo de estructuras es más conveniente adoptar topologías donde no se impone la restricción izquierda-derecha. Además, no se debe imponer que el estado de entrada sea el inicial, ya que se puede iniciar la expresión en cualquier punto de la secuencia. Por tanto, la probabilidad inicial se equi-distribuye entre todos los estados.

Utilizando esta nueva topología, se ha comparado el sistema utilizando la misma base de datos pero creando previamente (con las mismas imágenes) secuencias que siguen las diferentes evoluciones temporales citadas. El porcentaje de reconocimiento medio obtenido para las diferentes expresiones es del 82%, mientras que utilizando modelos de izquierda a derecha con entrada únicamente en el primer estado y con secuencias con evolución de neutro a máximo, el resultado era del 84%. Como puede observarse, la disminución de la tasa de reconocimiento es muy pequeña en comparación con la mejora que se obtiene de la posibilidad de aplicar el sistema

a todo tipo de secuencias con diferentes evoluciones.

Figura 19: Ejemplos de detección de emoción

Además, se ha introducido un proceso de validación que permite reducir la probabilidad de error en secuencias que no corresponden a emociones. Para ello se ha utilizado un factor de confianza:

$$R = P_{\mathrm{mod}elo elegido} - P_{\mathrm{mod}elo siguiente}$$

Se descartará aquellas secuencias con un factor R por debajo de un umbral. Este umbral se ha ajustado empíricamente para conseguir una probabilidad de error baja descartando el mínimo número de secuencias correctas. Los experimentos han mostrado que se puede descartar aproximadamente un 60% de las secuencias que corresponden a habla y no representan ninguna emoción clara.
Se ha investigado también la posibilidad de combinar la información proporcionada por los FAPs con la que puede extraerse de una estimación densa de movimiento en la región facial. Para ello, se ha utilizado el sistema de estimación de movimiento de Kanade-Lucas-Tomasi. La información de movimiento extraída se ha comprimido utilizando Análisis de Componentes Principales (PCA). Sin embargo, añadir esta información adicional no ha mejorado los resultados del reconocedor.

Finalmente, se ha elaborado una base de datos audiovisual que permitirá combinar la información extraída de un módulo de análisis de voz con la información extraída del análisis de vídeo, para mejorar el reconocimiento de las expresiones mediante un sistema multimodal, también basado en los Modelos Ocultos de Markov.

IV. Representación y gestión conjunta de la información

En esta línea se ha definido un conjunto de algoritmos de compresión que aprovechan la información descrita mediante los metadatos. De hecho, se ha analizado tanto herramientas básicas de codificación, como definido estrategias de codificación y transmisión. La finalidad de este trabajo es proponer un sistema codificador-descodificador que combine la compresión e identificación de los contenidos.

A. Herramientas de codificación mediante metadatos

Se ha investigado la posible mejora en codificación mediante el uso de metadatos que han sido extraídos con otra funcionalidad (por ejemplo, funcionalidades de indexación, búsqueda y adquisición, navegación, etc.). Para ello se ha definido diversos escenarios donde es posible la utilización de estos descriptores para mejorar la codificación. En concreto, se ha investigado los estándares de indexación MPEG-7 y SMPTE para la mejora de los codificadores híbridos actuales. Se ha dedicado particular atención a la codificación de caras, por su importancia en sistemas de videoconferencia, videotelefonía y multimedia. Los diversos esquemas de codificación investigados son:

- Reordenación de tramas: Las diferentes tramas de una secuencia de video pueden ser reordenadas antes de la codificación para generar una secuencia (reordenada) más eficiente para codificar. Esta secuencia reordenada puede contener, por ejemplo, todas las tramas similares juntas en el mismo segmento. Los resultados obtenidos son de ganancias de hasta el 8% en tasa de datos para la misma calidad visual.
- Codificación de transiciones: Existen descriptores estándar que representan las transiciones típicas de secuencias de video. Estas transiciones almacenan información de la localización de la transición así como de un modelo de la misma. Esta información puede ser usada para mejorar la predicción dentro de la transición. Mejoras de hasta el 80% (en tasa de datos para la misma calidad visual) para transiciones cortas y del 20% para transiciones graduales largas.
- Selección de tramas de referencia: Este esquema de codificación pretende reformular el problema de estimación de movimiento en uno de búsqueda y adquisición. En este caso, se pretende aprovechar la mejora de utilizar varias tramas de referencia en codificación para seleccionar estas tramas entre un conjunto de tramas mucho mayor. Como el conjunto de estas tramas puede ser muy elevado (de incluso miles de ellas) se usan metadatos para realizar la preselección de tramas de referencia. En este caso, mejoras alrededor del 10% en tren de datos has sido obtenidas mediante el uso de metadatos.
- Control de la tasa de bits con metadatos: Los parámetros de los codificadores híbridos actuales son generalmente fijados al principio de la codificación. Este esquema de codificación pretende generar diferentes reglas para actualizar dinámicamente los parámetros del codificador para ajustarse a la secuencia de entrada. En concreto se han investigados nuevas herramientas para seleccionar la estructura de GoP mediante diferentes metadatos de movimiento. Resultados preliminares presentan mejoras del 6% en tasa de bits para secuencias de entrada con diferentes shots.
- Codificación de caras: Se ha desarrollado un sistema de codificación de caras utilizando los coeficientes de la transformada Karhunen-Loeve de la cara a codificar. Dicha transformada implica una generación

adaptativa del auto-espacio. A tal fin se ha desarrollado un algoritmo que permite adaptar el auto-espacio a las caras de la secuencia de entrada. El sistema de codificación tiene un sistema básico basado en el estándar H-264 que permite la codificación de caras cuando la calidad de la imagen codificada con el auto-espacio no es de suficiente calidad. Como complemento al sistema de codificación, se han introducido los metadatos de caras descritos en el estándar MPEG-7 en el tren de datos generado, que permite tareas de búsqueda e indexación. Los resultados obtenidos igualan y en algunos casos mejoran los obtenidos en el estándar H.264.

Figura 20: Ejemplo de modelado de una transición en una secuencia de vídeo. Una mejor caracterización del tipo de transición permite mejorar la predicción dentro de la misma

B. Estrategias de codificación y transmisión

Dentro de esta tarea se incluye el estudio de la asignación conjunta de bits para datos y metadatos. En este estudio se pretende validar las estrategias de codificación usando metadatos de la tarea anterior. Se debe analizar escenarios para cada esquema y decidir si es necesario el envío de la información de los metadatos en recepción. En caso afirmativo, se debe realizar una estimación de la cantidad de bits necesaria para enviar estos metadatos y si esa cantidad representa un tanto por ciento elevado de la ganancia obtenida al usar los metadatos en codificación.

- Reordenación de tramas: Los metadatos usados para la reordenación son también necesarios en recepción. Si el decodificador no tiene acceso a los mismos deberán ser enviados. Se han investigado técnicas de codificación de esta información dando como resultado que el envío de esta información adicional únicamente representa el 1% de la ganancia obtenida al usar los metadatos
- Codificación de transiciones: Del mismo modo que el esquema anterior, los metadatos utilizados en la codificación de transiciones son también necesarios en decodificación. Resultados preliminares muestran que la transmisión de estos datos representa el 0.5% de la ganancia para transiciones largas y el 2% de la ganancia para transiciones cortas.
- Selección de tramas de referencia: En el caso de selección de tramas, resultados preliminares muestran que el envío de la información de referencia supone el 9% de la ganancia.
- Control del flujo de bits con metadatos: En este esquema de codificación únicamente se modifican los parámetros en el codificador, de este modo no son necesarios en recepción.

A su vez, en el ámbito de estrategias de codificación y transmisión, también se ha analizado técnicas de codificación con múltiples descripciones. En la técnica propuesta, los datos se codifican en una descripción de alta resolución mediante un codificador H.263. Además, una descripción de baja resolución se genera mediante la duplicación de las partes relevantes de la descripción de alta resolución.

C. Codificador/Descodificador para compresión e identificación

Finalmente, las ideas anteriores se han combinado para crear un codec para compresión e identificación. Para lograr este objetivo se han desarrollado dos campos principales:

- Creación de una estructura para la generación y uso de descriptores: Se ha desarrollado un marco común para la generación y acceso a descriptores MPEG-7 sobre imágenes y secuencias de vídeo. De esta manera, es posible la ampliación futura a nuevos descriptores más eficientes. Estas herramientas ya han sido integradas en el software común del grupo de imagen.
- Implementación de diferentes algoritmos de mejora de la codificación sobre un mismo codificador: En particular se ha elegido el codificador H.264/AVC como base para la implementación de las diferentes herramientas de codificación con metadatos.

Por último será necesario la puesta en común de ambas estructuras mediante un codificador "inteligente" que guarde información de los metadatos que han sido utilizados para indexación e identificación del contenido y utilice esa información para maximizar su eficiencia a la hora de codificar y enviar la información de vídeo.

REFERENCES

[1] B.S. Manjunath, P. Salembier, T. Sikora, editors, Introduction to MPEG-7: Multimedia Content Description Interface. Wiley, 2002. ISBN: 0-471-48678-7.

[2] O. Avaro, P. Salembier, Systems Architecture, En B. S. Manjunath, P. Salembier, T. Sikora, Introduction to MPEG-7: Multimedia Content Description Interface, Chapter 3, Wiley, 2002.

[3] P. Salembier, J. Smith, Overview of MPEG-7 multimedia description schemes and schema tools, En B. S. Manjunath, P. Salembier, T. Sikora, Introduction to MPEG-7: Multimedia Content Description Interface, Chapter 6, Wiley, 2002.

[4] A. B. Benitez, J. M. Martinez, H. Rising, P. Salembier, Description of a Single Multimedia Document, En B. S. Manjunath, P. Salembier, T. Sikora, Introduction to MPEG-7: Multimedia Content Description Interface, Chapter 8, Wiley, 2002.

[5] R. Mech, F. Marqués, "Objective evaluation criteria for 2-D shape estimation results of moving objects", en el Special Issue on Multimedia Signal Processing, Journal on Applied Signal Processing, vol. 2002, n. 4, pp. 401-409, abril 2002.

[6] P. Salembier, “Overview of the MPEG-7 Standard and of Future Challenges for Visual Information Analysis”, EURASIP Journal on Applied Signal Processing, Volumen 4, pp. 1-11, abril 2002.

[7] ¨D. Comás, R. Singh, A. Ortega, F. Marqués, “Unbalanced multiple description video coding based on a rate-distortion optimization”, en el Special Issue on Image Analysis for Multimedia Interactive Services, Journal on Applied Signal Processing, vol. 2003, n. 1, pp. 81-90, enero 2003.

[8] P. Salembier and J. Ruiz, “On Filters by Reconstruction for Size and Motion Simplification”, Sixth ISMM'2002, pp. 425-434, Sydney, Australia, abril 2002.

[9] M. Pardas, A. Bonafonte, J.L. Landabaso, “Emotion recognition based on MPEG-4 facial animation parameters”, Proc. ICASSP’02, pp. IV: 3624 - 3627, Orlando, mayo 2002.

[10] P. Salembier and J. Ruiz, “Connected Operators Based on Reconstruction Process for Size and Motion Simplification”, Proc. ICASSP’02, Orlando, mayo 2002.

[11] F. Marqués, C. Sobrevals, “Facial Feature Segmentation from Frontal View Images”, Proceedings EUSIPCO-2002, pp, 33-36 Toulousse, Francia, septiembre 2002.

[12] L. Garrido, P. Salembier, “A framework for the retrieval of multiple regions using Binary Partition Trees and low level descriptors”. Proceedings EUSIPCO-2002, Toulousse, Francia, septiembre 2002.

[13] F. Marqués, M. Pardàs, R. Morros, “Object-matching based on partition information”, Proc. IEEE ICIP-02, pp. II.829 – II.832, Rochester, USA, septiembre 2002.

[14] X. Giró, F. Marqués, “Semantic entity detection using description graphs”, Workshop on Image Analysis for Multimedia Services WIAMIS-03, pp. 39-42, Londres, abril 2003.

[15] D. Douxchamps, X. Marichal, T. Umeda, P. Correa, F. Marqués, “Automatic body analysis for mixed reality applications”, Workshop on Image Analysis for Multimedia Services WIAMIS-03, pp. 423-426, Londres, abril 2003.

[16] G. Stamou, Y Avrithis, F. Marqués, S. Kollias, P. Salembier, “Semantic unification of heterogenous multimedia archives”, Workshop on Image Analysis for Multimedia Services WIAMIS-03, pp. 573-577, Londres, abril 2003.

[17] J.L. Landabaso, M. Pardas, A. Bonafonte, “HMM recognition of expressions in unrestrained video intervals”, Proc. IEEE ICASSP’03, pp. III: 197 - 200, Hong Kong, China, abril 2003.

[18] J. Ruiz Hidalgo, P. Salembier, “Metadata based coding tools for hybrid codecs”, Picture Coding Symposium, Saint-Malo, Francia, abril 2003

[19] Noel O'Connor, Sorin Sav, Tomasz Adamek, Vasileios Mezaris, Ioannis Kompatsiaris, Tsz Ying Lui, Ebroul Izquierdo, Christian Ferran Bennström and Josep R Casas, “Region and object segmentation algorithms in the Quimera segmentation platform”, CBMI, 2003.