Artículos

Indización automática de vídeo

Por **Toni Navarrete** y **Josep Blat**

Resumen: *Tras comentar los problemas que presenta la indización de imagen y vídeo con respecto al texto, se describen algunas de las técnicas básicas para la indización automática de vídeo. Se presenta el paradigma de recuperación basada en el contenido y se exponen los métodos automáticos de segmentación e identificación de fotogramas clave, además de introducir los parámetros de bajo nivel que pueden identificar una imagen. Se muestran también las deficiencias básicas de los métodos automáticos basados únicamente en la imagen y algunos ejemplos de proyectos que utilizan información adjunta, tales como el audio o el texto sobreimpreso. Por último se apunta que el uso de estándares internacionales para la descripción del contenido, como por ejemplo Mpeg-7, puede facilitar la aparición de aplicaciones más ricas basadas en vídeo.*

Toni Navarrete Terrasa es ingeniero en informática por la Universitat de les Illes Balears. Es profesor de los estudios de informática en la Universitat Pompeu Fabra desde 1999, donde está realizando su tesis doctoral en sistemas de recuperación de vídeo e información geográfica en el marco de la web semántica, que ya ha resultado en diferentes publicaciones. Ha participado también en varios proyectos de investigación europeos.
http://www.tecn.upf.es/~tnavarrete
http://www.tecn.upf.es/gti

Josep Blat es catedrático y director del Departamento de Tecnología de la Universitat Pompeu Fabra. Ha dirigido proyectos europeos de investigación en diversas áreas de multimedia y sistemas interactivos (como entornos cooperativos, portales web con inteligencia, enseñanza telemática, juguetes computacionales educativos o gráficos 3D avanzados), así como en modelización y análisis matemático de imágenes. Es co-director de dos másters en animación y en diseño y programación de videojuegos.
http://www.tecn.upf.es/~jblat
http://www.tecn.upf.es/gti

Palabras clave: *Vídeo, Imagen en movimiento, Indización automática de vídeo, Recuperación basada en contenido, Procesamiento de imagen, Segmentación de vídeo, Fotograma clave.*

Title: Automatic indexing of video

Abstract: *After an initial discussion of the problems presented by image and video indexing as compared to text indexing, the authors describe some of the basic techniques for automatic video indexing. The content-based retrieval paradigm and some automatic methods for segmentation and key-frame identification are further described. Certain low-level parameters for identifying an image are also introduced. The authors discuss the drawbacks of such automatic methods based solely on the image and give examples from projects using accompanying information as well, such as audio and captions. The article concludes by pointing out that the use of standards, like Mpeg-7, can promote the development of new and richer applications based on video.*

Keywords: *Video, Motion images, Automatic indexing of video, Content-based retrieval, Image processing, Video segmentation, Key-frame.*

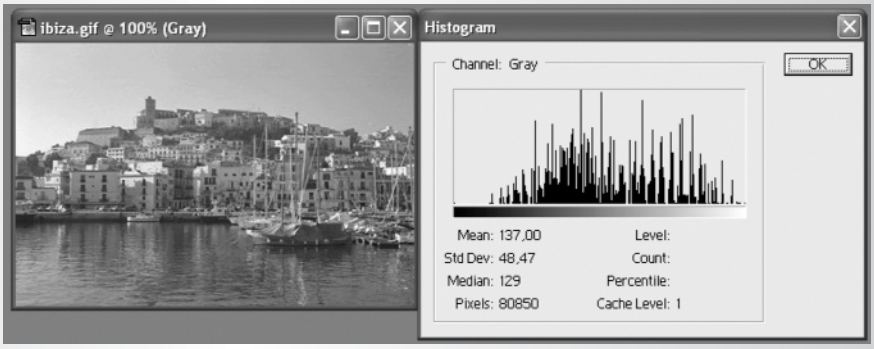
***Navarrete, Toni; Blat, Josep**. "Indización automática de vídeo". En: El profesional de la información, 2003, noviembre-diciembre, v. 12, n. 6, pp. 430-442.*

1. Introducción

Tanto desde un punto de vista semántico como técnico (formatos de representación y métodos de procesamiento) la indización automática de imágenes es mucho más complicada que la de texto. Además, el vídeo —o imagen en movimiento— añade más elementos a esa complejidad. Todo esto hace que los resultados en el campo de la indización automática de vídeo, a pesar de los avances, estén aún lejos de los conseguidos con texto. Vale la pena mencionar aquí la distinción existente entre el proceso de indexación (generación de los índices informáticos de un campo o de un fichero con sus estructuras apropiadas como diferentes

Artículo recibido: 13-06-03
Aceptación definitiva: 24-08-03

¿Qué es el histograma de una imagen?



Representa la cantidad de píxeles de la imagen para cada posible valor de color. En la figura 4 se muestra el histograma de una imagen con 256 niveles de gris, numerados del 0 (negro) al 255 (blanco). Así, el histograma tiene 256 valores, uno para cada valor posible. El 0 representa el número de píxeles negros en la imagen, el 255 el número de píxeles blancos, etc. Vemos en la imagen cómo predominan los valores medios, habiendo muy pocos valores bajos (oscuros) y siendo el valor medio de 137. Cuando se trabaja con imágenes en color se pueden utilizar tres histogramas, uno para cada uno de los tres componentes (rojo, azul y verde) o bien uno para la combinación de los tres, estableciendo grupos de valores (nótese que sería poco útil un histograma con 16 millones de entradas).

tipos de árboles) y la indización automática (asignación automática de términos para la representación del contenido). Es a esto segundo a lo que se consagra fundamentalmente este artículo. Señalemos que en *EPI* se han publicado varios artículos sobre diversos aspectos del tratamiento documental de imágenes estáticas, entre los que se cuentan las dos revisiones de **Mari Carmen Marcos** (1998, v. 7, n. 11 y 1999, v. 8, n. 7-8), así como la de **Jesús Muñoz** sobre bancos de imágenes (2001, v. 10, n. 3).

«Los resultados en el campo de la indización automática de vídeo, a pesar de los avances, están aún lejos de los conseguidos con el texto»

Mientras que los métodos de indización automática de texto toman la palabra como unidad a partir de la cual se realiza la indización y búsqueda, tras unas fases de extracción de palabras vacías de significado (como preposiciones, artículos y demás) y una normalización o lematización, esta unidad mínima de significado no es tan clara al tratar con imágenes.

Intentaremos comparar la indización de una frase con la de una imagen para comprender las diferencias. Por un lado, la oración "El profesional de la información es una publicación bimestral" es relativamente sencilla para un método automático. Simplificando el proceso, se extrae "el", "de", "la" y "una", se normaliza el resto de palabras de alguna manera y se añaden al índice las siguientes entradas: "profesional", "información", "ser" (muchos sistemas también lo eliminarían), "publicación" y "bimestral". Al hacer una búsqueda, por ejemplo por la palabra "publicación", se compara ésta con las entradas en el índice, normalmente no en orden secuencial sino utilizando una estructura en árbol para mejorar la eficiencia. No entramos aquí en temas de compresión del índice ni de búsquedas dentro de estructuras comprimidas.

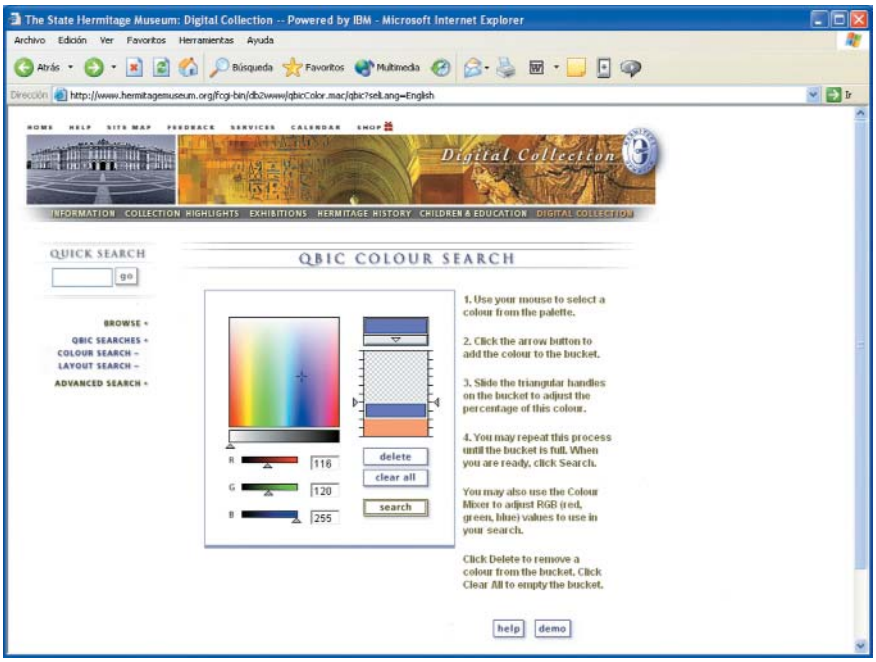
Pero al intentar indizar la imagen de la figura 1 (página 432), en primer lugar habría que plantearse cuáles serían las entradas que deberían aparecer en el índice. En segundo lugar, cómo podría el usuario especificar su consulta.

Los métodos manuales (a veces también denominados intelectuales) de indización de imágenes se basan en asignar una lista de descriptores a la imagen e introducirlos en el índice. Por ejemplo se podrían emplear los siguientes: *Ibiza*, *Baleares*, *velero*, *puerto*, *catedral*, *patrimonio Unesco*, *mar*, *atardecer*, etc. En su lugar, también es posible utilizar una descripción textual que se indizaría siguiendo el proceso antes citado. Por ejemplo: "*veleros anclados en el puerto de Ibiza (Baleares), ciudad patrimonio de la Unesco, en un atardecer de un día soleado*" podría en cierto modo describir esta imagen.

En cualquiera de los dos casos la búsqueda se especificará mediante una cadena de texto ya que en realidad se trata de una consulta no sobre la imagen en sí, sino sobre su descripción (meta-información). Como es lógico, el uso de tesauros conlleva una sustancial mejora en las búsquedas.

Los métodos automáticos no siguen este enfoque, en parte debido a que se encuentran muy lejos de poder extraer una lista de descriptores significativos, y ni mucho menos tan completa como la que haría un documentalista. En su lugar se utiliza el paradigma de recuperación de imágenes basado en el contenido y ve-

Figura 1. Puerto de Ibiza



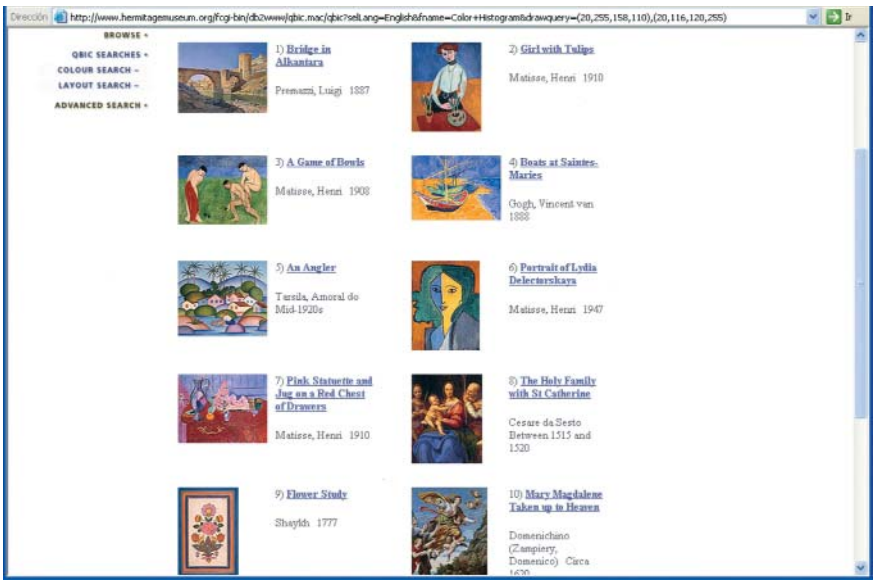


Figura 2. Definición de una búsqueda por color con Qbic y página de resultados. Imágenes capturadas de la web del Museo del Hermitage

remos cómo la forma habitual de expresar la consulta no es a través de términos, sino mediante muestras (otras imágenes o bocetos). Esto se describirá en el apartado 2. A continuación, en los puntos 3, 4 y 5 se presentarán las fases de la indización de vídeo siguiendo este modelo: segmentación, selección de fotogramas clave e indización. En el apartado 6 se enumeran algunas aplicaciones basadas en este paradigma, así como sus restricciones. El séptimo recoge una serie de sistemas que introducen un mayor grado de semántica a partir de elementos externos a la imagen en sí, como por ejemplo el audio. El artículo acaba con una recapitulación final en la que se destaca la importancia de los estándares de descripción del contenido audiovisual.

2. Recuperación basada en el contenido

Como hemos comentado, los métodos automáticos han venido siguiendo fundamentalmente un enfoque diferente al basado en descriptores. La idea es extraer un conjunto de parámetros de la imagen que la puedan llegar a identificar de forma unívoca. Veremos que esos parámetros, todos ellos de bajo nivel, están relacionados básicamente con el color, la forma y la textura. Además, y muy importante, cambia la manera de especificar las búsquedas a un paradigma basado en lo que se conoce por búsqueda por ejemplos. Así, la consulta típica no se hace mediante una cadena de términos sino introduciendo una imagen o dibujando un esbozo, a partir de lo cual el sistema buscará otras con características similares de color, forma y/o textura.

El más conocido de los modelos que siguen este enfoque es probablemente *Qbic*[1] (*Query by image content*) de *IBM* (en una versión más moderna se denomina *CueVideo*), que además de imagen estática también soporta secuencias de vídeo. Se está utilizando actualmente como base para un buscador online de las

cas tienen aún poca aplicabilidad en sistemas reales. Pocos usuarios están interesados en hacer búsquedas según patrones de formas o colores. Una excepción podrían ser los sistemas de ayuda a los publicistas, ya que en su trabajo sí pueden necesitar peticiones de ese tipo. Pero, desde luego, aún no son aplicables a modelos de búsqueda generales o a bibliotecas digitales.

No obstante, hay ciertos casos en los que sí es posible una aproximación basada únicamente en parámetros de bajo nivel de la imagen extraídos automáticamente. Se trata de problemas reducidos y en los que todas las imágenes tienen ciertas propiedades comunes. Un ejemplo claro, si bien no podría considerarse propiamente un sistema de indización de vídeo, es el de las aplicaciones de detección y reconocimiento de matrículas de coche que podemos encontrar en las puertas de numerosos aparcamientos. Todas las matrículas (al menos las de un mismo país) presentan un mismo tamaño, colores y formas unificadas que facilitan el proceso.

«En otros casos se asocia el texto del guión de una película o serie de televisión con las imágenes, típicamente a nivel de plano o escena»

Más interesante y complejo es el sistema de **Wang** (et al.)[16]. A partir de una base de imágenes con las más de 2.000 especies de peces de Taiwán, identifica cuáles son los que aparecen en secuencias de vídeo, siendo tolerante a cambios del ángulo de visión respecto al del patrón y a oclusiones parciales (por parte de otros peces u objetos). El sistema tampoco sigue el paradigma de búsqueda por contenido, sino que es más bien una herramienta de etiquetado de vídeo, pero muestra las posibilidades en un contexto reducido.

Otro problema, en parte similar al anterior aunque se resuelva por técnicas muy distintas y al que se han dedicado muchos esfuerzos de investigación, es el de detección y reconocimiento de caras humanas. En realidad se trata de dos cuestiones distintas: detección y reconocimiento, que se abordan con técnicas diferentes. En lo que se refiere a la detección, se pretende localizar dónde hay caras en una imagen, y suele abordarse analizando parámetros de color y buscando áreas con los colores típicos de la piel, cabello, ojos, etc. En cuanto al reconocimiento se pretende que, dada una cara, se recuperen los datos de esa persona a partir de un banco de fotos (preferentemente desde varios ángulos). La mayoría de los algoritmos de reconocimiento está basada en la localización de una serie de puntos determinados tales como la punta de la nariz, las comisuras de los labios, las pupilas, la punta de la barbilla y así hasta entre 15 a 80 puntos singulares según el sistema, y que llegan a identificar de forma prácticamente unívoca a una persona. El hecho de utilizar parámetros morfológicos permite que los algoritmos sean independientes de las razas, de la presencia de barba o incluso de gafas (siempre que los ojos sean visibles). Algunos sistemas comerciales consiguen una efectividad muy cercana al 100% en posición frontal, si bien ésta baja sensiblemente al trabajar con diferentes poses.

7. Sistemas basados en información adjunta

Así pues, con la excepción de estos casos comentados del soporte al publicista y detección y reconocimiento de formas muy concretas como matrículas, peces o caras, la información que puede extraerse de una imagen de forma automática no es suficiente para construir un sistema más genérico, como una biblioteca digital de propósito general o un buscador de imágenes o vídeos en internet. Pero de igual forma resulta imposible plantearse un etiquetado manual de todas las imágenes y vídeos de la Red. Y más teniendo en cuenta que para indizar manualmente un vídeo de una hora de duración suele necesitarse mucho más de una hora. La solución que muchos sistemas han seguido pasa por utilizar información "adjunta" al vídeo (no la imagen propiamente dicha) y extraer de ahí la semántica que permita unas descripciones y búsquedas más ricas. El problema pasa así a ser a menudo de recuperación de texto, formulándose la consulta mediante términos. Veremos a continuación varios sistemas que extraen esta semántica de fuentes diversas.

1. Código html.

Tanto *Webseek*, de la *Columbia University*, como *Google* indizan imágenes en la web, aunque el primero también lo hace con clips de vídeo. En el caso de *Webseek* se asume que muchas veces en el propio nombre del fichero se describe en parte una imagen. Igualmente puede ser en el nombre del directorio donde se halle alguna referencia semántica. Así, cuando busquemos por la palabra "Barcelona", lo que hará será recuperarnos todas las imágenes que contengan esta palabra en el nombre del fichero o path. Al hacer la indización, esta información se extrae de la etiqueta html utilizada para insertar la imagen.

Además de que la suposición de que en el path de un fichero hay información relevante es bastante cuestionable, este mecanismo de indización reduce al vídeo a una unidad indivisible, sin considerar segmentos, lo cual es excesivamente simplista. Como contrapartida *Webseek* combina este método con una herramienta de búsqueda por contenido basada en el histograma.

Figura 6. Resultado de la consulta "Chirac Aznar". Imagen capturada de Google

Google también utiliza una técnica similar pero más sofisticada. Parte de la idea de que alrededor de la imagen aparece un texto que la describe. Así, lo que se indiza en relación a esa imagen, además del path, es ese texto próximo a ella. El mecanismo de búsqueda es similar al llevado a cabo cuando es texto, pero recuperando las correspondientes imágenes, como en la figura 6.

Cuanto mayor sea la colección de imágenes, más fácil será que las diez primeras recuperadas sean relevantes. Aún así, no siempre el texto cercano a una imagen habla de ella, por lo que pueden obtenerse resultados que nada tienen que ver con la cadena de búsqueda. En el ejemplo se buscó "Chirac Aznar" en *Google* y, a pesar de que la mayoría sí muestran a ambos presidentes, o al menos a uno de ellos, también aparece una "curiosa" imagen de Bush en solitario. A veces los resultados son completamente diferentes a lo que se buscaba. Nótese cómo casi ninguna de las recuperadas contiene en su url las palabras de la cadena de búsqueda. Hay que aclarar que *Google* únicamente indiza imágenes estáticas y no vídeos y, aunque el método sería también aplicable para este tipo de documentos, nos encontraríamos de nuevo con la simplificación de no considerar segmentos.

2. Notas de producción.

La cadena de televisión japonesa *NHK*, *Japan Broadcasting Corporation*, utiliza para su archivo de noticias las notas de producción. Así, al vídeo de cada noticia se le asigna el texto preparado durante la confección de la misma —con unos valores temporales asociados— siendo ése el que se utiliza para la indización, si bien también se emplean técnicas de procesamiento de lenguaje natural para determinar sujeto y acción de cada oración y mejorar así el proceso (pueden encontrarse más detalles en la nota 17).

Un enfoque similar se utiliza en otros casos al asociar el texto del guión de una película o serie de televisión con las imágenes, típicamente a nivel de plano o escena, según se pretendan implementar posteriormente las búsquedas. Evidentemente, en un enfoque como éste son aplicables todas las técnicas de recuperación de texto.

3. Audio.

En muchos entornos la semántica está en el audio, y más en concreto en la voz. Los noticiarios de televisión son un claro ejemplo. Así, se aplican técnicas de procesamiento de habla para obtener el texto asociado a cada segmento. Un ejemplo de herramienta que sigue este procedimiento es el *CueVideo* de *IBM*, que es una continuación de *Qbic*. Para encontrar más información sobre este proyecto puede visitarse la siguiente url:

http://www.almaden.ibm.com/projects/cuevideo.shtml

No obstante, los algoritmos de procesamiento del habla no son del todo fiables y menos aún cuando intervienen varios hablantes en la conversación y el sistema no ha podido ser entrenado. Aun así, ofrecen una sustancial mejora a los sistemas basados sólo en imagen.

4. Audio y texto sobreimpreso.

Además de en el audio, y especialmente en las noticias de televisión, es frecuente que pueda encontrarse información sobre la imagen en el texto que se sobreimprime. Un claro ejemplo es la biblioteca digital *Informedia* de la *Carnegie-Mellon University* y su sistema de recuperación de noticias bajo demanda. En un proyecto concreto de *Informedia*[18,19] se asocian lugares geográficos a las noticias a partir tanto de la voz como del texto sobreimpreso. Para la extracción del texto sobreimpreso se ha desarrollado un sistema de reconocimiento óptico de caracteres en vídeo (*vocr*). El resultado es que puede visionarse un mapa de la zona asociada a la noticia, además de buscar otras asociadas a lugares concretos (también se puede usar un mapa como interfaz). Como curiosidad hay que decir que es la mayor videoteca digital del mundo, con varios terabytes de vídeo:

http://www.informedia.cs.cmu.edu

obras del *Museo del Hermitage* en San Petersburgo. En la figura 2 vemos cómo se ha especificado una búsqueda por color, donde la muestra es una zona horizontal azulada sobre otra zona rojiza.
http://www.hermitagemuseum.org

También puede definirse una búsqueda a partir de una disposición de los elementos más compleja. En la figura 3 (página 436) vemos cómo puede especificarse el layout que se toma como muestra para la consulta. Otros conocidos sistemas de este tipo, son *Swim*[2], de la *National University of Singapore* y los proyectos de la *Columbia University VideoQ*[3] y *VisualSeek*, todos ellos con soporte para vídeo.
http://www.ctr.columbia.edu/VideoQ/
http://www.ctr.columbia.edu/VisualSEEk/

Por otra parte, la indización de vídeo o imagen en movimiento presenta un problema añadido respecto a la imagen estática, ya que no es viable indizar todos y cada uno de los fotogramas. Téngase en cuenta que un vídeo de 30 minutos de un programa de televisión, a razón de 25 fotogramas por segundo tiene un total de 45.000. Además, sería computacionalmente costosísimo analizarlos e indizarlos todos y tampoco tendría sentido, ya que la variación de un fotograma al siguiente suele ser mínima. Así pues, el problema que se plantea es: qué vamos a elegir como unidad para la indización en vez del fotograma. A esa unidad le denominamos segmento y al proceso para obtenerlos segmentación.

3. Segmentación automática

Como se ha explicado, la segmentación es el proceso mediante el cual se divide el vídeo en unidades más pequeñas que serán la base para la posterior indización y a las que se les denomina segmentos. Normalmente, se toma como unidad el plano (o *shot* en inglés) entendiéndose por tal una serie de fotogramas contiguos filmados sin interrupción, en los que no hay cambios de cámara y que representan una acción continua en el tiempo y el espacio.

Actualmente existen numerosas aplicaciones, tanto de investigación como comerciales, de segmentación automática, y la mayoría se basa en el color. Así, analizando el histograma (ver cuadro) de cada fotograma, se asume que un cambio brusco de color supone un cambio de contexto y por tanto un nuevo segmento. Sin embargo, los cambios de plano no siempre son rígidos, ya que en la edición se utilizan frecuentemente efectos de fundido (u otros) para pasar de un plano al siguiente. Esto provoca que las técnicas basadas en el color (y en el histograma en particular) generen en las zonas de transición multitud de segmentos muy pequeños, incluso de un único fotograma. Esta microsegmentación también aparece típicamente cuando la cámara hace un *zoom*.

Para solventar este inconveniente, algunos sistemas simplemente fijan una duración mínima del segmento (por ejemplo *Físchlár*[4] no los acepta cuando son menores de un segundo); otros simplemente tienen controles manuales que permiten corregir la microsegmentación en zonas conflictivas. Algunos sistemas, además del color, utilizan otras propiedades de la imagen con el fin de extraer posibles movimientos de cámara como *zooms*, *pans* (movimientos en el plano horizontal) o *tilts* (cuando se producen en el plano vertical), efectos de fundido, etc., y tenerlos en cuenta a la hora de decidir si se está ante un cambio de plano o no. Este proceso suele estar basado en la extracción de regiones de la imagen y en el posterior análisis de su movimiento.

«El problema que se plantea es: qué vamos a elegir como unidad para la indización en vez del fotograma. A esa unidad la denominamos segmento y al proceso para obtenerla segmentación»

Por otra parte, algunos sistemas, por ejemplo los antes citados *Swim* o *VideoQ*, realizan la segmentación directamente sobre el vídeo comprimido en *Mpeg* o *Jpeg*, con la consiguiente mejora en la eficiencia. Existen otros modelos de segmentación que no toman el plano como base (ver notas 5, 6, 7, 8, 9 y 10 entre las más relevantes) pero por razones de espacio no entraremos en más detalles.

4. Selección automática de los fotogramas clave

Una vez que se han identificado los segmentos, siguiendo con el enfoque de recuperación por contenido, la próxima etapa es la de seleccionar uno o varios fotogramas (*key-frames* o fotogramas clave) que identifiquen cada segmento. Serán éstos los que posteriormente se indizarán y servirán de base para las búsquedas. Existen varios métodos para su selección: el más simple es tomar el primer fotograma del segmento (otros utilizan en vez del primero el último o el del medio); también existen casos donde se toma un *key-frame* cada cierta cantidad de fotogramas (a veces ese número es configurable por el usuario).

Sin embargo sería más inteligente seleccionar el fotograma que mejor represente a la serie que forma el segmento. Para ello, algunos sistemas (como por ejemplo el citado *Físchlár*) toman aquel con el histograma más cercano al promedio. *Qbic* genera en ocasiones un

fotograma clave "virtual". En un movimiento de *pan* (cuando la cámara se mueve a lo largo del plano horizontal) es difícil que un único fotograma identifique al segmento, así que *Qbic* genera una imagen mosaico y la toma como *key-frame*.

La mayoría de los sistemas utilizan estos fotogramas clave no sólo como base para la indización sino también como forma de presentar los resultados de las búsquedas. Algunos también los usan agrupados en varios niveles para generar una navegación jerárquica del vídeo, como el caso de *Swim* o *Físchlár*[11] entre otros. En este sentido, es también interesante la interfaz de visualización rápida de un vídeo basada en fotogramas clave de *Mbase*[12], del *Fuji Xerox Palo Alto Laboratory*.

5. Indización automática de los fotogramas clave

Una vez que tenemos un *key-frame* hay que indizarlo. Para ello, siguiendo el enfoque de recuperación basada en contenido, se extrae una serie de parámetros de bajo nivel de la imagen, que serán utilizados en las búsquedas. Para ser más concreto, la imagen se describe mediante tres componentes: color, forma y textura. En este apartado vamos a comentar brevemente algunos de los parámetros más utilizados para cada componente. No obstante, debido a su complejidad matemática, no entraremos en detalles acerca del proceso de extracción. Lógicamente, no todos los sistemas utilizan todos ellos, sino ciertas combinaciones.

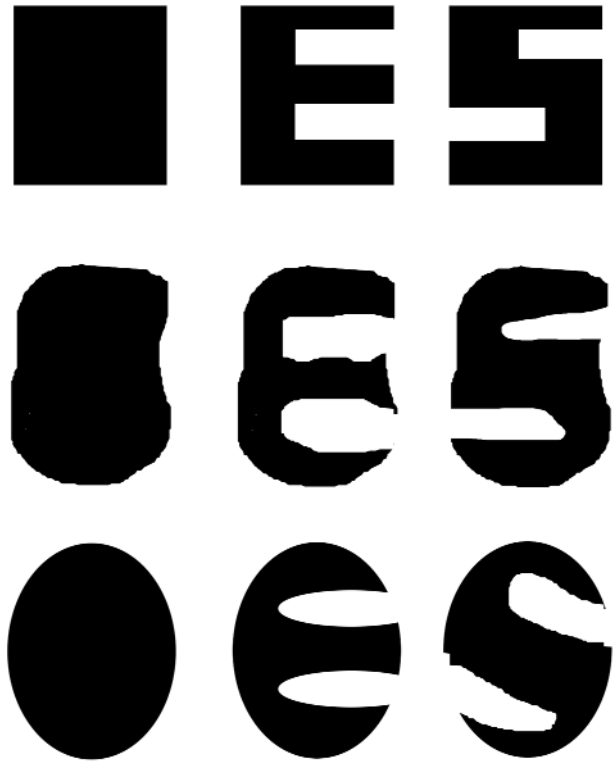


Figura 5. Diferencia entre región y contorno

La manera más fácil de indizar una imagen es a partir del color y, al igual que en el proceso de segmentación, el histograma constituye una útil herramienta. La ventaja de usar el color respecto de otros componentes como la forma, es que el histograma de una imagen apenas varía si se efectúan pequeñas operaciones de rotación, traslación o escalado, con lo cual el mecanismo es en cierta medida tolerante a cambios del ángulo de visión, todo lo contrario que le ocurre a la forma. Además, téngase en cuenta que es difícil que dos imágenes distintas presenten un histograma igual.

«La ventaja de usar el color respecto de otros componentes como la forma, es que el histograma de una imagen apenas varía si se efectúan pequeñas operaciones de rotación, traslación o escalado»

Hay que hacer notar que hay otras maneras de representar el color de un píxel además de usando los valores de rojo, verde y azul (RGB). Estas representaciones, al igual que RGB, son lo que se denominan espacios de color. De hecho, tanto la señal de televisión (ya sea *PAL* o *Ntsc*) como el vídeo digital se codifican utilizando otro espacio de color llamado YUV, donde la "Y" representa la luminancia (la información de niveles de gris), mientras que la "U" y la "V" se utilizan para identificar la crominancia (la parte del color). Cada sistema de codificación construye U y V siguiendo unas fórmulas diferentes. Conviene aclarar que a menudo se utiliza el acrónimo YCbCr en vez de YUV. Otros espacios de color sobre los que aquí no profundizaremos son el conocido *cmyk* (*cyan*, *magenta*, *yellow*, *black*) de los colores complementarios de RGB, utilizado en el contexto de impresión sobre papel, *HSV* (cuyo nombre proviene de *hue*, *saturation*, *value*), y *hmmd*, (*hue*, *max*, *min*, *difference*).

Algunos sistemas utilizan otros mecanismos para representar la información de color, ya sea sustituyendo o complementando al histograma. Por ejemplo, es frecuente utilizar una lista de los colores predominantes en vez del histograma, ya que es más flexible que éste. Otros incorporan información acerca del *layout* de colores de la imagen. Normalmente, las técnicas de extracción que se utilizan están diseñadas para funcionar con un espacio de color concreto.

Los otros dos componentes que pueden ser utilizados son la textura y la forma. En cuanto a la primera, existen varios modelos matemáticos para representarla, basados en la frecuencia de repetición, la orientación o el contraste. Es destacable que *Mpeg-7* ha definido y estandarizado lo que se ha denominado des-

criptor homogéneo de textura (*HTD*).

Aunque la forma es la componente que parece intuitivamente más clara para un humano, y probablemente la que más utilicemos, su descripción no es fácil. La forma se representa a partir tanto de su región como de su contorno. En la imagen de la figura 5 vemos la diferencia entre ambos conceptos. Observamos que los elementos de cada fila son similares entre sí siguiendo un criterio basado en la región que ocupan (la primera fila un rectángulo, la segunda una forma irregular y la tercera una elipse). Por contra, también los elementos que están en una misma columna son similares entre sí en lo que respecta al contorno (una E la segunda columna y una S la tercera).

En según qué aplicaciones, se pone el enfoque en una, otra, o ambas. Hay diferentes maneras de extraer y representar regiones y contornos. Un método frecuente de representación es hacerlo a partir de los ángulos que presentan, ya que así se consigue una mayor independencia de la traslación, rotación y escalado; por otra parte, el cálculo no resulta excesivamente complejo. Además de almacenar las formas destacables de la imagen, también pueden ser de especial relevancia las relaciones espaciales entre ellas y cómo éstas van variando a lo largo del segmento.

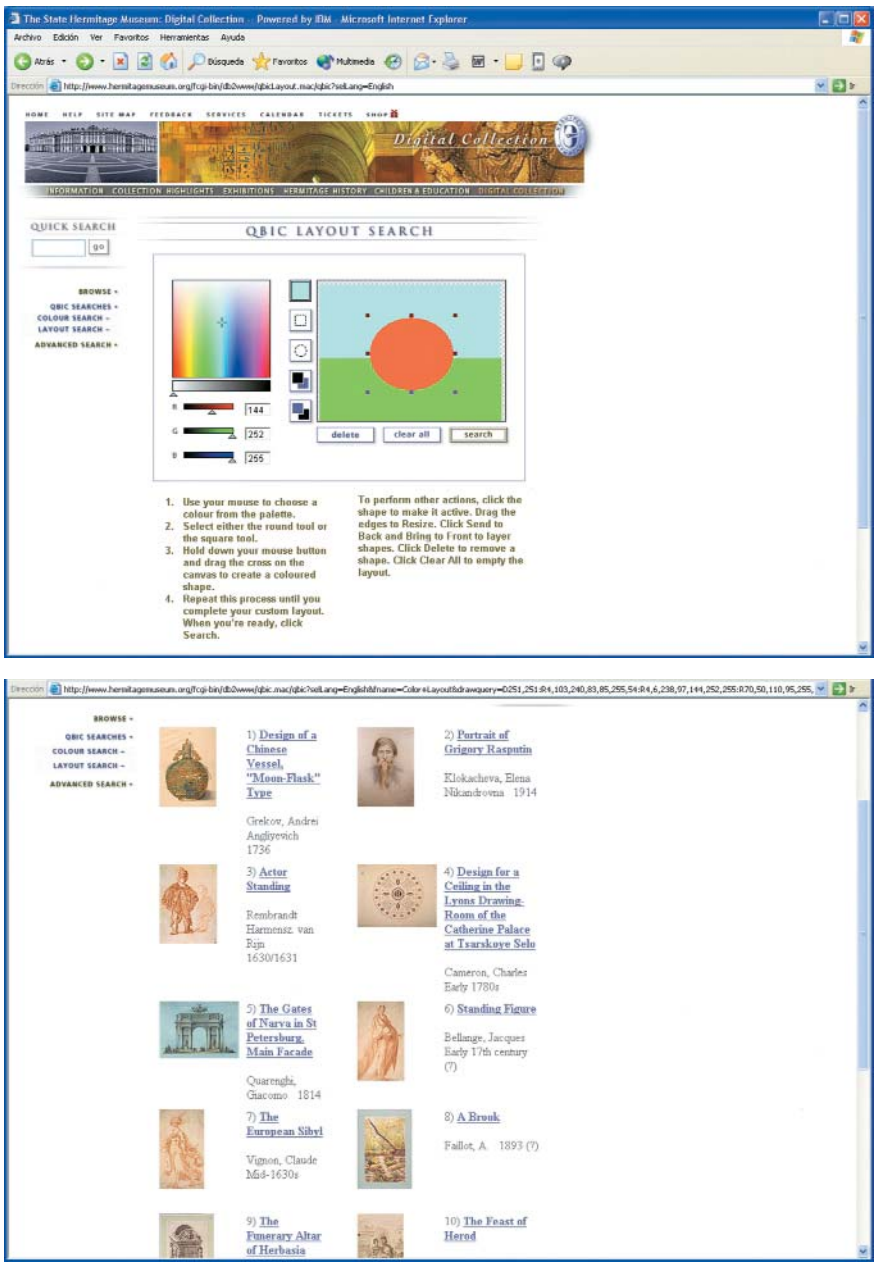


Figura 3. Definición de una búsqueda por layout con Qbic y página de resultados. Imágenes capturadas de la web del Museo del Hermitage

Por último, además de estos parámetros del fotograma clave, algunos sistemas también registran la información acerca del movimiento a lo largo del segmento, ya sean de las regiones (objetos) dentro de la imagen, como movimientos de la cámara. *Pan*, *tilt*, *roll*, *zoom*, *track* o *dolly* son los posibles movimientos de una cámara.

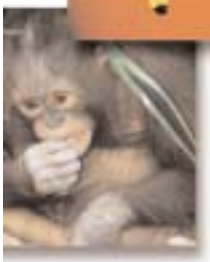
Si el lector desea obtener más información acerca de la indización de imágenes y vídeo, en las notas 13, 14 y 15 encontrará tres buenas revisiones sobre esta materia.

6. Ejemplos de aplicaciones basadas en contenido

Ya hemos visto anteriormente la interfaz del *Museo del Hermitage* que, aunque dedicado únicamente a imagen estática, sirve como un claro ejemplo de sistema de búsqueda por contenido. Interfaces similares para la consulta pueden encontrarse en los sistemas *Swim*, *VideoQ* y *VisualSeek*, entre otros.

Sin embargo, la efectividad de estos sistemas automáticos es relativa. En primer lugar, la herramienta para representar nuestra consulta (por colores o *layout*) no permite excesivos detalles. En segundo lugar, los resultados no siempre son todo lo precisos que cupiera esperar, al menos para la visión de un humano.

Además, y más importante aún si cabe, estos enfoques presentan un gran inconveniente: el contenido semántico con el que se describe la imagen es prácticamente nulo y, como hemos visto, se reduce a una serie de parámetros de bajo nivel. A pesar de la importancia desde un punto de vista de investigación, estas técni-



Bountiful Research

CAB ABSTRACTS® via *ISI Web of Knowledge*℠ provides fertile ground for rich, robust research in agriculture and the applied life sciences. When you search CAB ABSTRACTS through the *ISI Web of Knowledge* platform, you'll see the fruits of your labor from:

- Full integration of resources and cross-search discovery tools
- Simultaneous searching of multiple resources, with de-duplicated results
- Inter-product links, links to full-text
- Direct links to citation information in *Web of Science*®

With CAB ABSTRACTS via *ISI Web of Knowledge*, a dynamic, easy-to-use interface gives you a richer research experience. Yet you also reap the high quality CAB ABSTRACTS features researchers have come to depend on:

- Global coverage (from 125+ countries) back to 1973
- Thousands of journals, books, and conference proceedings
- CAB Thesaurus and CABICODES

Cultivate rich, robust research at your institution.

CAB ABSTRACTS is produced by CABI Publishing

THOMSON ISI

For more information, please visit: **www.isinet.com/isi/forms/cabab**.
Or call +1 800 336 4474 or e-mail sales@isinet.com.

Un enfoque similar, utilizando la transcripción del audio y del texto sobreimpreso es el del proyecto *AT&TV*, del laboratorio de *AT&T* en Cambridge (Inglaterra), y el de la videoteca digital *Vision*[20] de la *University of Kansas*.

5. Coordenadas geográficas.

Terminamos este apartado comentando un caso algo diferente como es el del proyecto *VideoGIS*[21]. Su objetivo es combinar el vídeo con la información geográfica que en él aparece. Así, a partir de una colección de documentos filmados desde el aire o desde tierra, se está desarrollando un sistema que soporta dos operaciones básicas:

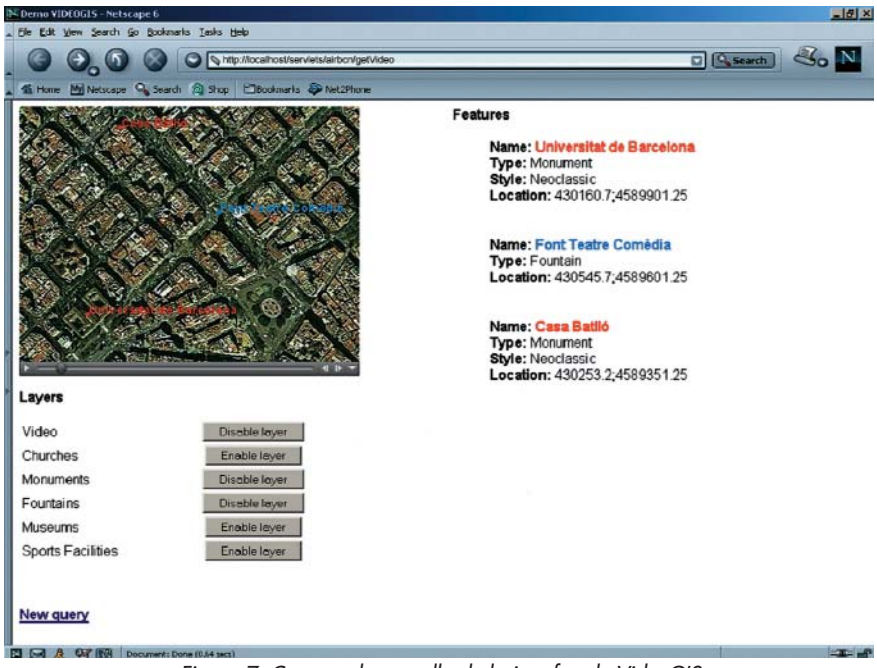


Figura 7. Captura de pantalla de la interfaz de VideoGIS

—Sobreimpresionar los elementos geográficos de interés sobre la secuencia de vídeo. El usuario puede hacer un clic sobre ellos para obtener más información de los mismos.

—Construir de forma automática itinerarios guiados a partir de una consulta del tipo "genera un vídeo que muestre los monumentos modernistas del Ensanche de Barcelona" o "genera un vídeo que muestre los monumentos a menos de un kilómetro de cierto punto" (figura 7).

Para filmar el vídeo se utiliza también un receptor de *GPS* adosado a la cámara, además de otros sensores de inclinación y orientación. A partir de estos valores puede determinarse la posición de la cámara y hacia dónde está enfocada. De esta forma la base para la indización automática es un fichero con estos datos de situación. El sistema realiza sucesivas consultas a un sistema de información geográfica que devuelve qué hay en ese espacio visible, es decir, qué elementos geográficos aparecen, que son los que se utilizarán para describir e indizar el vídeo. Al realizar la consulta, de nuevo se utiliza el sistema de información geográfica para recuperar qué elementos son los que satisfacen la consulta, para después recuperar las secuencias donde aparecen.

Figura 8. Información acerca de los pendientes de la protagonista. Imagen tomada de la web de HyperSoap

Nótese que el hecho de trabajar con largas secuencias de vídeo continuo no permite un enfoque basado en planos (*shots*). Por ello un segmento en *VideoGIS*, es decir la unidad mínima, será la parte de vídeo en la que aparecen los mismos elementos, o dicho de otra forma, una serie de fotogramas contiguos con características geográficas uniformes.

La generación automática de los itinerarios guiados sigue unas reglas de composición que pretenden darle una cierta estructura narrativa. Por ejemplo, si de la consulta de los monumentos del Ensanche resulta un vídeo de 5 horas, probablemente éste no será útil para el usuario, por lo que el sistema debe decidir qué fragmentos tienen más relevancia y cómo componerlos para que se siga una línea argumental y el resultado se ajuste a un tiempo adecuado.

8. Conclusiones y expectativas de futuro

Hemos visto cómo las técnicas para la indización automática de vídeo basada en el contenido han avanzado considerablemente en los últimos años. No obstante, excepto en contextos muy concretos y reducidos, los parámetros a bajo nivel de la imagen que se utilizan no aportan la semántica necesaria para construir un sistema genérico. Así, varios modelos utilizan la semántica de cierta información adjunta (audio y/o

texto sobreimpreso en la mayoría de casos) para ofrecer soluciones más ricas.

Tener una buena descripción del vídeo hace posible implementar sistemas dotados de una mayor "inteligencia" que los clásicos mecanismos de búsqueda. El sistema de itinerarios guiados de *VideoGIS* es un ejemplo, pues no es sólo un sistema de recuperación sino más bien de generación automática de presentaciones, algo así como un editor automático de documentales.

En una línea similar ha habido numerosos trabajos dedicados a la generación automática de noticiarios "a la carta" o de resúmenes de programas de televisión. Además, la televisión interactiva (TVi) permite interesantes aplicaciones como *HyperSoap*[22] del *MIT*: una telenovela para TVi donde prácticamente todo lo que aparece en la imagen se puede comprar (ropa, muebles, etc.); el usuario hace un clic sobre un elemento y ve su descripción, incluyendo el precio. Al acabar el capítulo, puede ver una lista de los productos consultados y comprar los que desee (figura 8).

http://www.media.mit.edu/hypersoap/

Un factor importante en la aparición de nuevas aplicaciones de este tipo será la adopción de estándares para el desarrollo del contenido audiovisual. Varios estándares han surgido a partir de distintas organizaciones internacionales y con objetivos diferentes. Algunos de los más relevantes son <*indecs*> orientado a la gestión de la propiedad intelectual; *Advanced authoring format* (*AAF*), enfocado a la descripción del proceso de autoría; *TV-Anytime*, *Smef* (de la *BBC*) y *Pmeta* (de la *Unión Europea de Radiodifusión*), que se centran principalmente en el intercambio de programas de televisión; *Smpte 335M*, de propósito general; y *Mpeg-7*, estándar *ISO*, el más amplio y fácilmente extensible de todos (mediante esquemas xml). Para saber más acerca de *Mpeg-7* las notas 23 y 24 ofrecen una buena introducción.

Notas

1. **Flickner, M.**; **Sawhney, H.**; **Niblack, W.**; **Ashley, J.**; **Huang, Q.**; **Dom, B.**; **Gorkani, M.**; **Hafner, J.**; **Lee, D.**; **Petkovic, D.**; **Steele, D.**; **Yanker, P.** "Query by image content: the Qbic system". En: *Ieee computer*, 1995, septiembre, pp. 23-31.

2. **Zhang, H. J.**; **Low, C. Y.**; **Smoliar, S. W.** "Video parsing, retrieval and browsing: an integrated and content-based solution". En: *ACM multimedia*, 1995, pp. 15-24.

3. **Chang, S.**; **Chen, W.**; **Meng, H.**; **Sundaram, H.**; **Zhong, D.** "VideoQ: an automated content based video search system using visual cues". En: *ACM international conference on multimedia*, 1997, pp. 313-324.

4. **Lee, H.**; **Smeaton A. F.**; **O'Toole, C.**; **Murphy, N.**; **Marlow, S.**; **O'Connor, N. E.** "The Físchlár digital video recording, analysis, and browsing system". En: *Riao: content-based multimedia information acces*s, 2000.

5. **Davenport, G.**; **Aguirre, S.**; **Pincever, N.** "Cinematic primitives for multimedia". En: *Ieee computer graphics & applications*, 1991, julio.

6. **Aguierre Smith, T. G.** *If you could see what I mean*. MIT MS thesis. Cambridge, Massachusetts, EUA, 1992.

7. **Hjelsvold, R.**; **Midtstraum, R.** "Modelling and querying data". En: *20th International conference on very large data bases*, 1994.

8. **Adali, S.**; **Candan, K. S.**; **Chen, S.**; **Erol, K.**; **Subrahmanian, V. S.** "Advanced video information system: data structures and query processing". En: *ACM-Springer multimedia systems journal*, 1996.

9. **Subramanian, V. S.** *Principles of multimedia database systems*. San Francisco: Morgan Kaufman Publishers, 1997. Isbn 1558604669.

10. **Tran, D. A.**; **Hua, K. A.**; **Vu, K.** "Semantic reasoning based video database systems". En: *11th International conference on databases and expert systems applications*, 2000.

11. **Lee, H.**; **Smeaton, A. F.**; **Furner, J. "**User interface issues for browsing digital video". En: *21st BCS Irsg colloquium on IR*, 1999.

12. **Foote, J.**; **Boreczky, J.**; **Girgensohn, A.**; **Wilcox, L.** "An intelligent media browser using automatic multimodal analysis". En: *ACM multimedia*, 1998, pp. 375-380.

13. **Aigrain, P.**; **Zhang, H.**; **Petkovic, D. "**Content-based representation and retrieval of visual media: a state of the art review". En: *Multimedia tools and applications*, 1996, v. 3, pp. 179-202.

14. **Smeulders, A. W. M.**; **Worring, M.**; **Santini, S.**; **Gupta, A.**; **Jain, R. "**Content-based image retrieval at the end of the early years". En: *Ieee transactions on pattern analysis and machine intelligence*, 2000, diciembre, v. 22, n. 12.

15. **Brunelli, R.**; **Mich, O.**; **Modena, C. M.** "A survey on the automatic indexing of video data". En: *Journal of visual communication and image representation*, 1999, pp. 78-112.

16. **Wang, C. H.**; **Lin, H. C.**; **Shih, C. C.**; **Tyan, H. R.**; **Lin, C. F.**; **Mark Liao, H. Y.** "Querying image database by video content". En: *Advances in multimedia information processing, PCM 2002. Third Ieee Pacific rim conference on multimedia*, 2002.

17. **Kim, Y. B.**; **Shibata, M.** "Content-based video indexing and retrieval – a natural language approach". En: *Ieice transactions on information and systems*, 1996, E79-D (6), pp. 695-705.

18. **Haupmann, A. G.**; **Witbrock M. J.** "Informedia: news-on-demand multimedia information acquisition and retrieval". En: **Bradbury, M. T.**: *Intelligent multimedia information retrieval*. Cambridge; Massachusetts: MIT Press, 1997. Isbn 0-262-63179-2.

19. **Christel, M. G.**; **Olligschlaeger, A. M. "**Interactive maps for digital video library". En: *Ieee international conference on multimedia computing and systems*, 1999, pp. 381-387.

20. **Gauch, S.**; **Li, W.**; **Gauch, J.** "The Vision digital video library". En: *Information processing & management*, 1997, v. 33, n. 4, abril, pp. 413-426.

21. **Navarrete, T.**; **Blat, J.** "VideoGIS: segmentación y modelado de vídeo basado en información geográfica". En: *3er Congreso interacción persona-ordenador*, 2002.

22. **Bove Jr., V. M.**; **Dakss, J.**; **Agamanolis, S.**; **Chalom, E.** "Adding hyperlinks to digital television". En: *Smpte 140th technical conference*, 1998.

23. **Martínez, J. M.** (ed.). "Mpeg-7 overview (version 8)". *Moving picture experts group. ISO/IEC JTC1/SC29/WG11N4980*.

24. **Manjunath, B. S.**; **Salembier, P.**; **Sikora, T.** *Introduction to Mpeg-7: multimedia content description interface*. John Wiley & Sons, 2002. Isbn 0-471-48678-7.

***Toni Navarrete** y **Josep Blat,** Departament de Tecnologia, Universitat Pompeu Fabra.*
toni.navarrete@upf.edu
josep.blat@upf.edu