Bases de datos para Multimedia: Recuperación por Contenido

Manuel Agustí i Melchor, Jose Miguel Valiente González
Dept. de Informática de Sistemas y Computadores
Universidad Politécnica de Valencia
46022 Valencia, España
e-mail: magusti@disca.upv.es, jvalient@disca.upv.es

Resumen

El avance de servicios multimedia acaecido con el progreso tecnológico y la posibilidad de compartir y distribuir datos de tipo imagen a través de las redes de comunicación han enfatizado la importancia de herramientas para la recuperación de información visual. Las bases de datos de imágenes se emplean en un vasto abanico de áreas como son el entretenimiento, el arte, la publicidad, la medicina y la industria entre otros. En todos estos contextos, el problema principal está relacionado con la necesidad de un acceso eficiente a la información de contenido visual. Generalmente, la información que se busca ha de obtenerse a partir de lo que los humanos recuerdan después de observar durante un par de minutos una determinada imagen. Puede incluir las formas de los objetos más relevantes, distribuciones o áreas de un determinado color, texturas o la disposición visual de los elementos en la imagen. En este contexto se sitúa la implicación de la temática de Recuperación de Imágenes Basada en Contenido (CBIR, *Content Based Image Retrieval*) como método para llevar a cabo esta tarea.

El usuario prefiere realizar las peticiones mediante términos relativos al contenido, que estén fuertemente orientadas a la presencia de objetos definidos de forma abstracta. La creación de índices para éstas, de forma manual, así como la dificultad de predecir todas las posibles futuras consultas para crear los índices correspondientes en la etapa de catalogación que hagan innecesarias futuras reindizaciones está lejos de una solución efectiva y eficiente.

En el presente trabajo se describen las aproximaciones utilizadas en el campo de las imágenes que permiten la clasificación de las mismas de forma automática. Así como también se plantea el caso del audio y el vídeo.

Palabras clave: Bases de datos multimedia, bases de datos de imágenes, búsqueda por contenido, indexación, multiresolución.

1. Introducción: elementos definitorios del problema.

Disponer de una base de datos es muy interesante para un gran número de aplicaciones. Sin embargo, la generalización de los esquemas tradicionales de gestión de las bases de datos impone restricciones demasiado severas para un uso eficiente en determinados contextos, como es el caso de los sistemas Multimedia. Los problemas existentes en sistemas Hipermedia a la hora de realizar búsquedas, como por ejemplo la Web, muestran que cuando el tamaño de esa colección de documentos es elevado y su tipología diversa, resulta poco interesante para el usuario la existencia de una base de datos que no defina mecanismos adecuados para su examen, consulta y recuperación. Para centrar el tema de discusión tomemos por ejemplo las bases de datos de imágenes (colecciones de imágenes).

Las aproximaciones convencionales para clasificar las características visuales en términos de descripciones textuales han demostrado ser inadecuadas para indexar imágenes [5]. De hecho, la menor riqueza expresiva del texto, respecto a las características visuales no permite explotar de forma plena las habilidades de la memoria del ser humano y los resultados de una consulta pueden no ser relevantes a a las expectativas del usuario. Esta es la razón por la que actualmente, las tendencias se encaminan a utilizar contenidos visuales como descriptores.

Las limitaciones de una aproximación basada en descripciones textuales y la oportunidad de apoyar la labor de los diseñadores sugieren incorporar al sistema la capacidad de incluir un sistema de consulta y recuperación basado en descriptores sintácticos y/o semánticos que permitan una aproximación mayor al contenido de las imágenes. Este proporcionará un método de exploración y

recuperación de las imágenes en función de, por ejemplo, el motivo de partida que ofrezca el diseñador.

Es por este motivo que se hace necesario utilizar otros descriptores de mayor nivel expresivo. En una primera aproximación pueden ser derivados del análisis estructural de las imágenes, basándose en la utilización de elementos como el color, la forma (contornos) y la textura que ya se obtienen en la etapa de análisis de la imagen. O bien agrupaciones u organizaciones de estos, de mayor entidad.

En un segundo nivel, la exploración se realiza a partir de la elección de un motivo, como se ilustra en la Figura 1; una imagen de pequeñas dimensiones pero con gran nivel de detalle (información usualmente obtenida a partir de un mapa de bits) es objeto de análisis de los elementos que componen la base de datos sin restricciones de descriptores precalculados. El resultado es la obtención de una secuencia de coincidencias que, para su discriminación, pueden ir acompañadas de un valor numérico que indica la presencia y cercanía del motivo en cada una de las imágenes mostradas.

Hablando en general, un sistema eficiente de recuperación de información visual debe ser capaz de:

- Definir elementos de recuperación que sean significativos en el contexto de la aplicación. A este subproblema se le denomina Modelado de las Imágenes. La importancia de esos elementos, descriptores, radica en su capacidad para describir las imágenes y permitir su procesado. .
- Proporcionar un Método de Consulta que permita al usuario especificar de forma natural características selectivas así como información imprecisa: se denominan técnicas de búsqueda.
- Definir métricas de igualdad o importancia que sean satisfactorias para la percepción del usuario, denominadas Medidas de Similitud.

Los métodos de búsqueda basados en descriptores son extremadamente eficientes ([17]). Es por este motivo que es la forma más utilizada de enfocar las operaciones de búsqueda en el ámbito de las imágenes. Pero hay dos problemas principales en la aproximación basada en descriptores. Por un lado el determinar qué descriptores hay que utilizar y por otro lado la representación de una determinada base de datos en forma de descriptores.

Si el usuario realiza, de forma natural, las consultas basándose en el contenido de las imágenes, ¿por qué no utilizar técnicas de descripción del contenido de las mismas para realizar el análisis de las imágenes? Esta aproximación es posible en el marco de trabajo apropiado: será aceptable si es posible en función del coste espacial y temporal del sistema de consultas. Estas aplicaciones son difíciles de llevar a cabo por las técnicas clásicas de reconocimiento de objetos utilizadas en Visión por Computador, por la complejidad de la tarea de encontrar objetos generales en contextos abiertos. Así, es necesario profundizar en la temática de determinar el contenido de las imágenes, de forma que sea posible la clasificación de objetos.

El método de consulta por ejemplos visuales es el paradigma de interacción que explota las capacidades naturales del ser humano en lo que se refiera al análisis de imágenes y su interpretación. Esto requiere que las características visuales sean extraídas de las imágenes y utilizadas como índices en la fase de recuperación. Las consultas por ejemplos visuales permiten la imprecisión e incompletitud de la expresión, al permitir a los usuarios esbozar la imagen que tienen en la memoria (por ejemplo formas coloreadas dispuestas siguiendo un determinado patrón).

Figura 1. Resultado de la exploración basada en el contenido de una imagen

La interactividad con la información que introduce el usuario es esencial para la fase de recuperación de información. Cuando los resultados obtenidos no son plenamente satisfactorios, el proceso habitual es mejorar la calidad de la respuesta manteniendo, tan bajo como sea posible, el número de fallos a expensas que se obtengan un número mayor de imágenes erróneas. De esta forma, la consulta inicial se refina o modifica sobre la base de las imágenes encontradas en cada caso.

La importancia de elementos visuales depende de la subjetividad del observador (que no se conoce a priori, cuando se crea la base de datos) y del contexto de la aplicación (que sí es conocido). Para abarcar la subjetividad del usuario, es necesario desarrollar criterios de búsqueda capaces de manejar imprecisiones (instancias no exactamente iguales a la de muestra) y falta de información en las descripciones que propongan los usuarios. El hecho que el contexto de la aplicación se conozca de antemano puede ayudar a escoger los algoritmos de reconocimiento y las funciones de similitud que deben ser incluidas en el sistema.

En los últimos años, varios sistemas prototipos se han propuesto que implementan sistemas de búsqueda por contenido (*retrieval by content*) en bases de datos de imágenes abordando diferentes aspectos de la información contenida en las imágenes, como son las texturas, similitud de formas y relaciones semánticas entre objetos de la imagen. En el siguiente punto se recoge la problemática y los enfoques que se han utilizado en el campo de la consulta (recuperación) de imágenes por contenido. El problema central es encontrar métodos de gran velocidad para encontrar el(los) elemento(s) que mejor corresponden a una determinada operación de consulta en grandes colecciones de imágenes.

2. Técnicas para la Recuperación de Imágenes por Contenido

Con el objetivo de ser capaces de describir los contenidos de una imagen, sobre la base de las características y/o representaciones de interés con la pretensión de abordar el diseño de un sistema de consultas por contenido en bases de datos de imágenes, se han venido utilizando diferentes aproximaciones a la temática de reconocimiento de objetos en imágenes, en la que es posible distinguir ([4]): búsqueda de primitivas visuales de bajo nivel (incluyendo características como color y texturas), pasando por estrategias de aprendizaje y agrupamiento jerárquico e incluyendo la capacidad de manipular objetos genéricos en contextos y configuraciones no predeterminadas. En la Figura 2 se relacionan las diferentes técnicas o aproximaciones que se emplean para esta temática y que se comentan a continuación.

Es posible distinguir tres tipos de descriptores que permiten un análisis de las imágenes a tres niveles CON diferente capacidad de abstracción y que son:

- Nivel 1: primitivas básicas. En éste, la posibilidad de establecer relaciones entre imágenes ser realiza en base a valores de color de los puntos, a las texturas que se pueden observar y a los contornos que es posible detectar con diferentes operadores;
- Nivel 2: sintáctico. Denominado "reconocimiento de patrones" por que aquí se unen las anteriores informaciones para buscar regiones de mayor entidad y comprobar si se ajustan a una reglas de descripción de elementos tipo;
- Nivel 3: semántico. Llamado "reconocimiento de objetos" en el que la complejidad es un poco mayor, al permitir que existan imprecisiones entre la definición del objeto de la realidad a buscar y sus posibles transformaciones sufridas en el paso a 2D, las variaciones de las condiciones de adquisición de la imagen, etc.

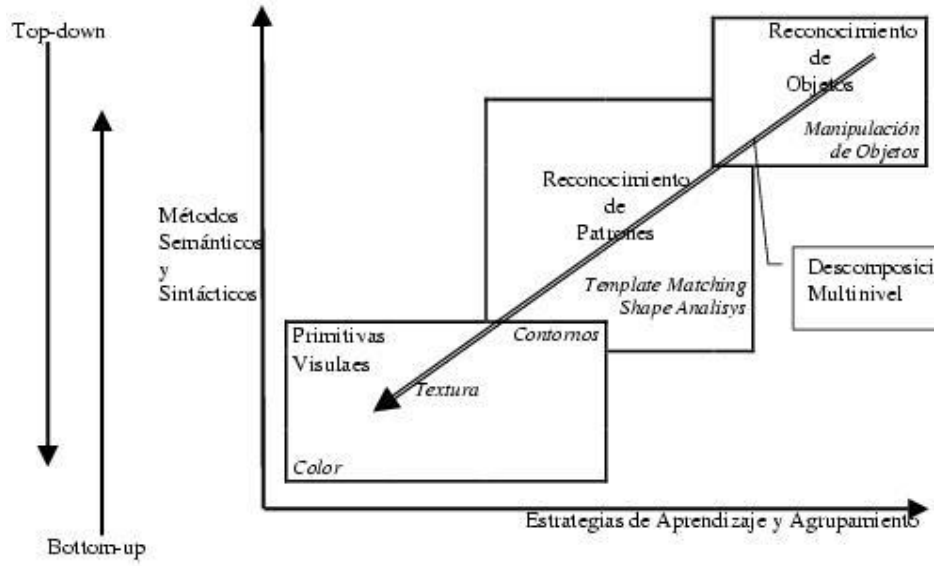

Figura 2. Técnicas para la descripción del contenido de una imagen

Los sistemas de consultas en bases de datos de imágenes se basan, tradicionalmente, en que realizan un preproceso en la imagen encaminado a obtener una conjunto de descriptores y utilizar estos para la búsqueda en la base de datos de imágenes. De esta forma se elimina el coste de realizar el análisis de contenidos de los elementos que componen la base de datos en cada consulta, pero limita las búsquedas a aquellas se pueden realizar con la combinación de los descriptores utilizados. De manera que este tipo de análisis es efectivo para conjuntos de imágenes en los que los descriptores satisfacen cualquier consulta permitida como por ejemplo en bases de datos de caracteres, caras, representantes de categorías (sólo gatos, peces, sillas, ...). Todas estas se caracterizan por que el motivo de las mismas es único, ocupa gran parte de la escena, es fácilmente distinguible del fondo y se puede normalizar los valores de color, tamaño del objeto o su posición.

Pero no es suficiente en aquellas colecciones (o librerías) de imágenes con gran riqueza de detalles que no es representable de forma compacta por los descriptores, cuyas características son la falta de control sobre las condiciones con que se obtienen las imágenes (posición de la cámara, iluminación, movimiento de la escena, etc.), la gran variabilidad de las características de los objetos de interés (color, tamaño, posición, número, ...) y existencia de partes ocultas. En ocasiones se pueden considerar como descriptores las propiedades de los materiales (descriptores de bajo nivel).

En los últimos años [15], varios sistemas prototipos se han propuesto abordando diferentes aspectos de la información contenida en las imágenes, como son las texturas, similitud de formas y relaciones semánticas entre objetos de la imagen. El objetivo de las técnicas de consulta basadas en contenido de las imágenes es encontrar de forma eficiente las imágenes en una base de datos de imágenes que son similares una indicada. A diferencia de las típicas consultas en bases de datos, en este caso se utiliza un criterio de similitud que no tiene por que ofrecer una coincidencia exacta.

2.1. Ejemplos reales de Sistemas de recuperación basados en Contenido.

A continuación se expone, de forma resumida, las realizaciones más notables relativas a esta temática, exponiendo los descriptores utilizados por las mismas. En la Tabla 1 se muestran, de forma resumida, los sistemas analizados y los conjuntos de descriptores separados por su nivel de información asociado.

En contextos cerrados (en los que las imágenes tienen un menor grado de variabilidad en cuanto a sus condiciones de partida y de consulta posterior al sistema) otros sistemas desarrollados ex-profeso han alcanzado cotas de acierto mayores. Pero sus funcionalidades no son generalizables a otras colecciones de imágenes. Por ejemplo en ámbitos como la catalogación de mapas [5], planos y reconocimiento de caras [1]. En estos, un proceso de vectorización de las imágenes permite introducir una fase de análisis sintáctico [18].

Sistema	Descriptores	
	Bajo	**Medio**
Netra [6]	color, textura	forma, regiones, localización espacial
PickToSee [1]	Color, contornos	invariantes geométricos
CANDID	color, textura	forma, firmas e histogramas globales
Chabot	color	textuales (fechas, localización geográfica), vectores de longitud variable
WebSeek	color	texturales, configuraciones espaciales, regiones, histogramas globales
QBIC	color, textura	textuales, forma, rectángulos de color orientados, regiones, posiciones espaciales
PhotoBook	color, textura	forma, regiones

Tabla 1. Sistemas de Recuperación de imágenes por Contenido

El sistema de *CANDID*[2] se parece a otros que utilizan histogramas de color como base para la comparación. A diferencia de éstos, calcula una función de densidad continua lo que le permite (en principio) diferenciar o considerar dos valores (de cualquier característica que utilice) muy próximos. Pero debido al elevado número de las distribuciones de probabilidad continuas que obtiene como representación normalizada de las signaturas, optan por utilizar un algoritmo de agrupamiento general: *k-means*. Con lo cual se desvirtúa esta idea inicial de sensibilidad en la apreciación de diferencias.

En el diseño del sistema *Chabot* [9] se presupone que la mayor parte de las consultas serán construidas en modo texto. Y sólo implementa un método de análisis de color como aproximación al análisis del contenido de las imágenes.

Las técnicas básicas (histogramas y disposición espacial de regiones de color y textura) de *WebSEEk* [16], *VisualSEEk* y *SaFe* son características globales de una imagen, aunque la descripción de las regiones introduce una componente de análisis local. Es de lamentar que en los métodos de análisis no se haya considerado la búsqueda de motivos atendiendo a variaciones de los mismos (transformaciones afines) puesto que en la aproximación indicada sólo es posible hablar en términos de contenido global de las imágenes y no acerca de si contienen un determinado motivo por ejemplo.

[1] http://www.wins.uva.nl/research/isis/zomax/

[2] http://www.c3.lanl.gov/kelly/CANDID/main.shmtl

En el caso de *QBIC (Query by Image Content,* [2], [8]*)*, es básicamente un sistema mejorado de búsquedas por palabras claves, al que se unen las medidas que es capaz de calcular, pero sin incluir información de carácter semántico de las imágenes de forma automática.

Photobook ([10, 11, 12, 13]) permite utilizar una técnica de aprendizaje supervisado conocido como *relevance feedback* (que requiere la presencia de un observador humano que refine las consultas realizadas en esta fase) para ajustar las etapas de segmentación y elección de parámetros para la clasificación. Pero, esta aproximación no permite el establecimiento de características acerca de la situación espacial que les impide la construcción de consultas efectiva basadas en objetos.

Para construir un sistema de consultas orientado a objetos es necesario alcanzar niveles más abstractos en la jerarquía de representación de la imagen y codificar la disposición espacial utilizando esta característica de agrupación de alto nivel. Puesto que las aproximaciones examinadas son inapropiadas para bases de datos heterogéneas si realizan la búsqueda basándose en una única característica básica como es el color o en una simple combinación de ellas.

3. Audio y Vídeo

Respecto a los restantes medios (audio y vídeo). Fuhr [3] realiza se puede encontrar una aproximación similar a la comentada al respecto de las imágenes.

3.1. El caso del audio

En el caso de elementos de tipo sonoro contenidos en una base de datos multimedia, se puede plantear el tratamiento de este medio a diferentes niveles:

- Nivel 1: primitivas básicas. Establece correspondencias de forma exacta a partir de muestras de sonido;
- Nivel 2: sintáctico. Permite las correspondencias inexactas de muestras de sonido. En este caso hay que considerar que puede darse por dos causas. En primer lugar por que existe una discrepancia en cuanto a propiedades de la captura del sonido: frecuencia de muestreo o cuantización; que podrían extenderse también a detalles de la compresión utilizada. Y por otro lado puede radicar en las características acústicas o perceptuales del sonido;
- Nivel 3: semántico. En función del contenido del audio analizado. distingue cuestiones como, por ejemplo, si se trata de una conversación o de una pieza de música gregoriana.

En este último caso, de nuevo, hay que introducir una componente de inexactitud en la búsqueda de similitudes que no existe en los otros niveles, como ya se ha expuesto en el caso de las imágenes.

Esta división se suele utilizar como fundamento para describir la estructura jerárquica que compone un documento de tipo sonoro en el que en una primera aproximación se puede describir en términos de lo expuesto en el tercer nivel (nivel semántico). Estos, se pueden descomponer en unidades dadas por el segundo de los criterios (nivel sintáctico) y que en su formulación más elemental se descomponen en otras que se ajustan al primero de los niveles (primitivas básicas).

3.2. El caso del vídeo

En este particular se habla de una secuencia de imágenes, generalmente acompañada de audio, en la que se pueden distinguir diferentes enfoques en función de la representación que se tenga de esa secuencia y que generalmente se conceptualiza como una secuencia de:

- Nivel 1: primitivas básicas. Se ocupa de detectar planos, pudiendo demandar la detección de encuadres/escenas;
- Nivel 2: sintáctico. Considerando que los cuadros o imágenes estáticas se suceden con una carencia temporal;
- Nivel 3: semántico. Se distingue entre secuencias de imágenes en la que los puntos de inicio y final los marcan los objetos y/o estructuras en movimiento en las mismas.

Esta división se suele utilizar como fundamento para describir la estructura jerárquica que compone un documento de tipo “vídeo” en el que en una primera etapa se pueden contextualizar las secuencias descritas en función del tercer nivel. Que a su vez se puede descomponer en unidades dadas por el segundo de los criterios y que en su formulación más elemental se descomponen en otras que se ajustan al primero de los criterios.

Aunque debido a la naturaleza compleja de este medio, también se puede combinar la estructuración en términos de las imágenes individuales que lo componen y las diferentes formas de elementos sonoros que las acompañan. Que en cada caso serían descritas como ya se ha expuesto en los puntos en que se han tratado estos media por separado.

4. Conclusión

El manejo de bases de datos multimedia de gran tamaño y contenidos diversos precisa de técnicas adecuadas de implementación de sistemas de recuperación y exploración ([14]). Centrándonos en el campo de las bases de datos de imágenes, la razón por la que el problema de recuperación es tan complejo es que el usuario espera del sistema que encuentre elementos relevantes basado en semánticas personales o culturales ([7]). La representación de informaciones de carácter semántico es muy compleja y requiere soluciones a problemas como la detección automática de características, segmentación y reconocimiento. Estos problemas todavía permanecen abiertos.

Las aproximaciones más sencillas a estos sistemas pasan por disponer de información textual asociada a cada imagen y utilizar, sobre estas representaciones, las técnicas clásicas de los sistemas de bases de datos tradicionales. Aunque esta sea una solución rápida, no es efectiva para grandes volúmenes de imágenes. La variabilidad y riqueza de interpretaciones son tan grandes como el esfuerzo requerido para realizar estas anotaciones. Los aspectos clave a la hora de construir un sistema de recuperación de información son la elección de los atributos o descriptores, su representación, los métodos de especificación de las consultas, las métricas que definen las correspondencias y las estrategias de indexación.

Las observaciones realizadas apuntan hacia métodos basados en el contenido que, sin necesitar la extracción de características o realizar segmentaciones sobre la imagen, permitan la recuperación de imágenes en el sentido de lo que indique la similitud o apariencia de la consulta inicial. Estas consultas se construyen a partir de imágenes en bruto. Sobre estas, las regiones de interés junto a su disposición relativa se combinan para construir la consulta a la base de datos. Las imágenes que son escogidas de la base de datos se han de ordenar en función de su similitud a la apariencia de la consulta.

Estas mismas ideas tienen su aplicación en el campo del audio y del vídeo como se ha expuesto en el análisis particular de estos media.

Referencias.

[1] D. Beymer, Vectorizing Face Images by Interleaving Shape and Texture Computations, A. I. Memo nº 1537, A.I. Lab. MIT, 1995.

[2] M. D. Flickner, H. Sawhney, W. Niblack, R. Barber, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, y P Yanker , Query by Image and Video Content: The QBIC system, IEEE Computer Magazine, vol. 28, nº 9, pp. 22-32, 1995.

[3] Norbert Fuhr, Multimedia Information Retrieval, SIGIR'98, 1998.

[4] Etemad Kamran, David S. Doermann, y Rama Chellapa, Multiscale Document Page Segmentation Using Soft Decision Integration, IEEE Transactions on PAMI, 1997.

[5] C. Leung, Image and Vision Computing, pp. 463-464, 1999.

[6] W. Y. Ma y B. S. Manjunath, NeTra: A Toolbox for Navigating Large Image Databases, Proc. of the Int. Conf. on Image Processing, 1997.

[7]M. Mandal, F. Idris, y S. Panchaanathan, A critical evaluation of image and video indexing techniques in the compressed domain, Image an Vision Computing, vol. 17, págs. 513-529, 1999.

[8] W. Niblack, R. Barber, W. Equitz, M. D. Flickner, E. M Glasman, D. P. Petkovic., y P. Yanker, The QBIC project: querying images by content using color, texture and shape IS and T/SPIE, Storage and Retrieval for Image and Video Databases, Intern. Symp. Electr. Imaging: Science and Technology. Conference núm. 1908, 1993.

[9] V. E. Ogle y M. Stonebraker, Chabot: Retrieval from a Relational Database of Images, IEEE Computer, vol. 28, nº 9, pp. 40-48, 1995.

[10] A. Pentland, R. W. Picard, y S. Sclaroff, Photobook: Content-Based Manipulation of Image Databases, The Media Laboratory. Massachusetts Institute of Technology, Perceptual Computing Section, TR. nº 55, 1993.

[11] A. Pentland, R. W. Picard, y S. Sclaroff, Photobook: Tools for Content-Based Manipulation of Image Databases, 1994.

[12] A. Pentland, R. W. Picard, y S. Sclaroff, IEEE Multimedia, Summer, págs. 73-75, 1994.

[13] A. Pentland, R. W. Picard, y S. Sclaroff, Photobook: Content-Based Manipulation of Image Databases, The Media Laboratory. Massachusetts Institute of Technology , Perceptual Computing Section, TR. nº 255, 1995 .
[14] S. Ravela, R. Manmatha, y E. M. Riseman, Image Retrieval Using Scale-Space Matching, Computer Vision - ECCV'96, págs. 273, 282, 1996.
[15] Arnold W.M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, y Ramesh Jain, Content-Based Image Retrieval at the End of the Early Years, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, nº 12, 2000.
[16] R. K. Srihari, Automatic indexing and content-based retrieval of captioned images, IEEE Computer, vol. 28, nº 9, pp. 49-56, 1995.
[17] H. Stone, Content-based imagen retrieval - Research issues, Multimedia Technology for Applications, capítulo 9, 1998.
[18] P. Vaxivière y K. Tombre, Celestin: CAD Conversion of Mechanical Drawings, IEEE Computer, vol. 18, págs. 64-54, 1992.