Recuperación Eficiente de Información Visual Utilizando Momentos

Gloria Elena Jaramillo, John Willian Branch

Universidad Nacional de Colombia, Escuela de Sistemas, Medellín, Colombia
{gejarami,jwbranch}@unalmed.edu.co

Abstract. Currently, multimedia repositories have presented a fast growing because of the Internet massification and the development of a huge amount of devices that capture visual information in digital format. Traditional research of images based on keywords may be vague and sometimes tedious. Content-based image retrieval (CBIR) comes up as a solution to the problem of visualization and access to visual information, which is supported in digital image processing and patter recognition. In this article, we present the results of an implementation of a CBIR system. For the image retrieval, we use shape descriptors, specifically statistics moments. A database of mechanical and automotive parts is used as case of study. The results show an average effectiveness of the system of around 90%.

Keywords: Content-based image retrieval, shape descriptor, moments.

1 Introducción

La recuperación de información visual surge como una necesidad de acceso rápido y eficiente en grandes colecciones de datos. Dicha necesidad ha ido incrementando debido al internet y la posibilidad que brindan las redes teleinformáticas para acceder a grandes repositorios de información. Las búsquedas tradicionales de recuperación de imágenes se han basado en palabras claves; tal enfoque puede presentar varios escenarios, principalmente se han reconocido dos: búsqueda basada en palabras circundantes, las cuales no reflejan necesariamente el contenido de la imagen, y búsqueda textual soportada en anotación manual de los metadatos de las imágenes. En ambos casos, la cantidad de información visual, el alto valor de subjetividad asociado, las imprecisiones y el costo humano confluyen en un enfoque que si bien está soportado en los grandes desarrollados de búsqueda de información en bases de datos textuales, se convierte en una técnica imprecisa para recuperar imágenes.

La recuperación de información visual basada en contenido o CBIR (Content-Based Image Retrieval) surge como una respuesta a esta problemática, centrando su atención en nuevas formas de recuperación semi o completamente automática de imágenes basándose en sus características propias como son color, forma y textura.

Estas características son expresadas en términos matemáticos, obteniéndose una descripción no subjetiva de su contenido. Dichos valores matemáticos se almacenan en los denominados vectores característicos, los cuales deben ser capaces de describir por completo la imagen. Adicionalmente, el método de descripción empleado debe ser lo suficientemente robusto para considerar traslaciones, rotaciones y cambios en el tamaño de los objetos presentes en la imagen. Otro aspecto que debe ser considerado y que influye directamente en la recuperación, y por tanto, en la efectividad del sistema, es el tipo de descriptor que debe ser utilizado. Este aspecto es una de las principales causas de que, hasta el momento, no haya sido posible la estandarización de una técnica para la descripción de las imágenes. Por ello, es necesario conocer cuáles son las características más relevantes del tipo de imagen que pretende ser recuperado para emplear el descriptor que mejor describa el contenido de la imagen.

En el presente artículo, se presenta un sistema para la recuperación automática de imágenes por contenido basándose en descriptores de forma. Para comprobar la efectividad del sistema, se toma como caso de estudio una base de datos de imágenes mecánicas y de automoción y se utiliza la precisión como medida de desempeño. El artículo se estructura como sigue: en la sección 2 se presenta la revisión de la literatura. Posteriormente, se describen las técnicas utilizadas para el reconocimiento de los patrones relevantes en las imágenes. En la sección 4 se presentan los experimentos y resultados. Finalmente, se presenta el trabajo futuro y las conclusiones.

2 Revisión de la literatura

El presente artículo se concentra en la recuperación de información visual basada en rasgos característicos de la forma, por lo que una revisión exhaustiva de técnicas utilizadas para describir el color o la textura de los objetos, está fuera de este alcance. Sin embargo, algunas referencias pueden ser encontradas en [5][26][29].

Históricamente, la forma ha sido un descriptor natural para la identificación y reconocimiento de objetos. Los primeros enfoques de percepción orientados hacia la identificación a partir de ésta característica han sido abordados por las teorías neuropsicológicas. En este campo de estudio, las doctrinas clásicas presentaban teorías principalmente orientadas hacia las imágenes bidimensionales. Los mayores representantes de estas doctrinas pertenecen a la escuela Gestalt, los cuales en 1923 publican una serie de leyes de percepción visual [34] entre las que se destaca la propiedad de invariabilidad, la cual plantea la preservación del contenido semántico de un objeto frente a cambios en la posición, tamaño o color. Uno de los trabajos que ha cobrado creciente interés por sus relaciones con las redes neuronales ha sido la teoría del reconocimiento de objetos basado en un proceso de aprendizaje mediante la asociación de sus partes constituyentes [9]. Otros representantes de este mismo periodo plantean que el reconocimiento de los objetos debe estar asociado a las propiedades de la superficie 3D en el espacio y no a sus proyecciones bidimensionales [6][7]. Estos trabajos basados en la percepción sirvieron de base para teorías

posteriores de reconocimiento, las cuales tradicionalmente han perseguido como objetivo imitar el sistema de visión humano. Otros trabajos más recientes de percepción visual pueden ser encontrados en [8][33][3][10][19].

Las primeras teorías modernas orientadas hacia enfoques computacionales comenzaron a estudiar la forma a partir de características como la textura, el contorno o los fractales, tomando como base las doctrinas de percepción. Entre los primeros desarrollos de esta época se encuentra una implementación que parte de la teoría de percepción de los puntos de alta curvatura [2] denominada primal sketch, la cual es propuesta por Marr y Hildreth [22] en 1980, basado en un filtro de detección de bordes a partir de puntos de inflexión. Estos métodos plantearon nuevas posibilidades de abordar la forma desde un punto de vista dinámico, considerando múltiples escalas de resolución. Entre éstos métodos se encuentran técnicas basadas en polígonos [24][25][4], los cuales aproximan la forma de un objeto a partir de puntos de alta curvatura.

Actualmente, existen una gran cantidad de técnicas para la descripción de un objeto según su forma, las cuales pueden ser clasificadas utilizando distintos criterios. En [20] se proponen tres clasificaciones: la primera asociada a si el resultado de la descripción es numérico o no. La segunda asociada al grado de preservación de información que permita reconstruir la imagen a partir del descriptor. La tercera, y más ampliamente utilizada para la tipificación de un método consiste en agrupar las técnicas en basadas en contorno o basadas en región. Entre los descriptores basados en regiones se distinguen el código cadena, los momentos (geométricos, Zernike, pseudo-Zernike, Legendre), Grid, ART y métodos de descomposición en primitivas. Las descripciones basadas en código cadena, CSS, perímetro, excentricidad y Fourier pertenecen a las técnicas basadas en contorno. En la literatura es posible encontrar estudios comparativos entre las técnicas basadas en contorno y en regiones [35][36].

Algunas de las técnicas que han llamado más fuertemente la atención, probablemente motivado por la cantidad de estudios comparativos que surgieron en el marco del desarrollo del estándar MPEG-7 [27] para la asignación de metadatos a la información visual, ha sido el uso de Contour Scale Space (CSS) y momentos para la descripción basada en contorno y regiones, respectivamente. La principal diferencia entre estas dos técnicas consiste en que la técnica basada en contorno utiliza solamente información de la silueta o el contorno cerrado que define el objeto. Por otra parte, la técnica basada en regiones calcula el descriptor a partir de todos los píxeles que conforman la imagen, soportando definiciones para regiones aisladas.

Particularmente, CSS utiliza un enfoque basado en puntos de inflexión, el cual va suavizando la imagen por medio de un filtro gaussiano, para cada iteración se hallan los puntos de inflexión y se detiene el proceso iterativo cuando la curva que representa el objeto se vuelve completamente convexa. Esta aproximación aunque es muy robusta para imágenes simples y complejas y es invariante a rotación, traslación y escalado, presenta como gran debilidad el tiempo de computo necesario para realizar el proceso, lo que adicionalmente dificulta la recuperación si la aplicación se ejecuta en tiempo real. Para resolver esto, se han creado aproximaciones que eliminan el ruido en la imagen y permiten calcular la evolución de la curvatura directamente utilizando el filtro Gaussiano [37]. En [1] se plantea un algoritmo para mejorar la

correspondencia de los puntos en el contorno de los objetos en secuencias de video. Adicionalmente, a la carga computacional asociada a esta técnica, existe una problemática asociada a la baja definición que ofrece el descriptor para objetos que no poseen muchas secciones cóncavas. Este aspecto ha sido abordado por medio de técnicas de mapeo de inversión del contorno [31]. Sin embargo, este enfoque ha sido sólo validado para reconocimiento de caracteres. En general, los métodos basados en contorno son robustos para imágenes tanto simples como complejas. Sin embargo, no ofrecen una definición para formas que constan de regiones no conexas.

Por otra parte, los métodos basados en regiones ofrecen una solución para la descripción de objetos compuestos de varias regiones no unidas al utilizar todos los píxeles de la imagen. En particular, los momentos estadísticos ha sido un enfoque ampliamente estudiado [15][28][11]. Entre las mayores ventajas de los momentos se encuentran la preservación de información, robustez frente a cambios en tamaño, escala y rotación. Sin embargo, la mayoría de ellos incorporan algún nivel de ruido, principalmente asociado a los momentos de alto orden, y redundancia en información. Entre las comparaciones que se encuentra en la literatura de los distintos momentos aplicados a la recuperación de información visual [36] es posible concluir que, en general, los momentos centrales y cartesianos son menos precisos que los momentos de Zernike, debido principalmente a su sensibilidad asociada frente a transformaciones geométricas [30]. En particular, los momentos de Zernike han sido una técnica ampliamente implementada debido a su robustez para describir formas más genéricas, esto se debe principalmente a que captura información en el dominio espectral, mientras que otras aproximaciones como los momentos geométricos, los cuales son más fáciles de calcular en comparación con los momentos de Zernike, capturan información en el dominio espacial.

Para nuestro caso de estudio, trabajos previos [14] han mostrado que los métodos basados en contorno, particularmente CSS, los momentos geométricos basados en regiones e incluso algunos descriptores basados en textura, si bien poseen una buena descripción robusta a invariabilidad, no ofrecen una descripción con un alto nivel de precisión.

3 Momentos de Zernike

Los momentos permiten representar una imagen bidimensional como una función de densidad de probabilidad de variables aleatorias en 2D.Particularmente, los momentos de Zernike forman un conjunto de funciones complejas definidas en un círculo unitario, las cuales aplicadas al tratamiento de imágenes digitales representan la longitud de la proyección ortogonal de las funciones base a la imagen.

Las primeras referencias encontradas del uso de momentos de Zernike para recuperación de información visual se atribuyen a Teague [32], el cual plantea que éstos son más efectivos y eficientes en la recuperación frente a métodos como la transformada de Fourier-Mellin y los momentos invariantes de Hu [12], presentando una menor redundancia de información.

Los momentos de Zernike se forman a partir de los polinomios de Zernike definidos en (1), sujeto a p-|q| un número par y |q|≤p; con q un entro positivo, p un entero positivo o cero y (ρ, θ) las coordenadas polares de (x, y) ϵ R^2.

$$R_{p\,q}(\rho) = \sum_{s=0}^{(p-|q|)/2} \frac{(-1)^s (p-s)!}{s!\left(\frac{p+|q|}{2}-s\right)!\left(\frac{p-|q|}{2}-s\right)!}\rho^{p-2s} \quad (1)$$

Los momentos de Zernike de orden p quedan definidos por (2).

$$Z_{p\,q} = \frac{p+1}{\pi}\sum_{x2+y2\le 1}\sum V^*_{p\,q}(x,y)f(x,y) \quad (2)$$

Donde $V^*_{p\,q}$ representa la conjugada de $V_{p\,q}$, la cual forma las funciones base calculadas mediante la expresión (3)

$$V(x,y) = V_{p\,q}(\rho\cos\theta, \rho\sin\theta) = R_{p\,q}(\rho)\exp(iq\theta) \quad (3)$$

Debido a que los momentos de Zernike están definidos sobre el círculo unitario, es necesario realizar una transformación de la imagen a coordenadas polares mediante las ecuaciones 4 y 5 antes del cálculo de los momentos.

$$\rho = \sqrt{x^2+y^2} \quad (4)$$

$$\theta = \tan^{-1}\left(\frac{y}{x}\right) \quad (5)$$

Donde ρ representa el radio en el punto (x, y) y θ representa el ángulo formado entre el eje de la abscisa y ρ.

Para garantizar la invariabilidad frente a cambios de traslación y escalado es necesario mover el centro de masa de la imagen (Z_{00}) al centro del círculo unitario. Adicionalmente, todas las imágenes deben ser normalizadas a un radio de tamaño fijo. La invariabilidad a rotación se obtiene utilizando sólo las magnitudes de los momentos, ya que una rotación de la imagen original solo produce un cambio de fase. Una vez se han obtenido los momentos, éstos deben ser normalizados dividiendo por la masa de la imagen; el vector obtenido es utilizado como índice de la misma.

Entre las ventajas de los momentos de Zernike, se encuentra que cada momento brinda una contribución única en la descripción de la imagen, adicionalmente, es capaz de describir imágenes muy complejas. Para la implementación del algoritmo se hallaron 36 momentos, lo cual requiere una gran carga computacional para efectuar los cálculos. Para lograr una recuperación eficiente, se utilizó una LUT (Look up Table) para almacenar los polinomios de Zernike, ya que éstos son independientes del contenido de la imagen. Para mejorar el desempeño en la etapa de matching, se realizó un filtrado consistente en recuperar como imágenes relevantes sólo aquellas

cuya masa y excentricidad estén dentro de una rango de tolerancia predefinido. Se realizaron varias pruebas considerando diversos umbrales, asignando un rango de tolerancia de 30% para la masa y 70% para la excentricidad.

4 Experimentos y resultados

4.1 Configuración del experimento

Las pruebas de la efectividad del sistema se realizaron en un PC con procesador Core Duo de 3.0 Ghz, memoria RAM de 3 Gb, bajo el sistema operativo Microsoft Windows XP. Los algoritmos propuestos se implementaron en Java. Adicionalmente, el procesamiento de las imágenes, incluyendo transformación a escala de grises, binarización y normalización de la imagen, se realizó utilizando el API de procesamiento de imágenes JAI (Java Advenced Imaging).

Las pruebas se realizaron utilizando una base de datos de 985 imágenes de piezas mecánicas y de automoción, cada una de las cuales fue ingresada al sistema como imagen petición. Para cada recuperación se calculó la precisión, la cual fue utilizada como parámetro para comprobar la efectividad del sistema.

4.2 Resultados obtenidos

La Fig. 1 muestra los resultados de la precisión del sistema antes de realizar el filtrado de las imágenes relevantes; el eje x corresponde a cada una de las categorías que agrupan semánticamente a las imágenes y el eje y representa la precisión.

Para este caso, se obtienen resultados que en su mayoría pertenecen a la misma categoría de la imagen petición. Sin embargo, debido a que la efectividad total se calcula como un promedio, las recuperaciones negativas afectan en gran medida el desempeño global del sistema. Adicionalmente, es posible observar del análisis de los resultados que algunas de las imágenes recuperadas para una imagen en particular presentan similitud geométrica mas no semántica. La efectividad total del sistema llega a 74% sin realizar filtrado.

La masa fue calculada inicialmente como filtro que restringiera los resultados a aquellas imágenes que poseen cierto grado de similitud respecto a la imagen petición. Sin embargo no se observaron buenos resultados debido a que ésta no es una medida que pueda describir eficientemente el contenido de una imagen. Por ello, junto con los momentos de Zernike se implementó otro descriptor de forma que sirve de filtro: excentricidad. Se utilizó un umbral de tolerancia de 30% para la masa y 70% para la excentricidad, ambos asignados empíricamente. La Fig. 2 presenta los resultados de precisión del sistema con este método.

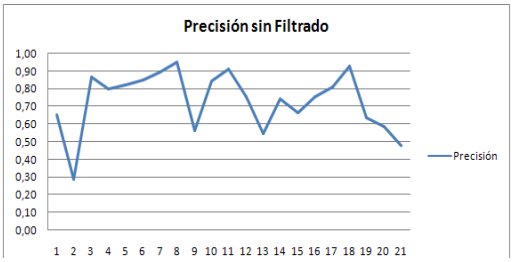

Fig. 1. Precisión obtenida por el sistema antes de realizar el filtrado

Del análisis de los resultados, se observa una mejora significativa en la precisión, llegando a un promedio aproximado de 90%, presentándose un incremento del 16% de recuperación positiva. Lo cual indica que en promedio se recuperan 18 imágenes de 20. Las Fig. 3 y 4 presentan algunas de las recuperaciones obtenidas con sistema tanto para imágenes simples como complejas, respectivamente.

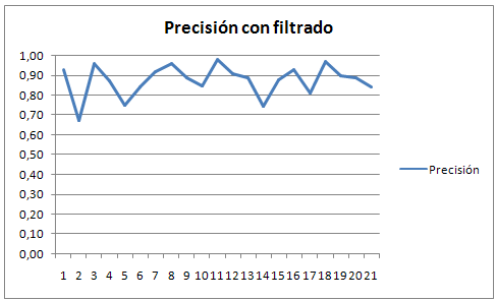

Fig. 2. Precisión obtenida por el sistema después de implementar el filtro

5 Conclusiones y trabajo futuro

En el presente artículo se presentó los resultados de la implementación de un sistema de recuperación de información visual por contenido. A partir de los resultados, es posible concluir la efectividad de los descriptores de forma para la descripción de las imágenes. Específicamente, los momentos de Zernike lograron una descripción efectiva tanto de imágenes simples como complejas.

Observando los resultados arrojados por el sistema es posible observar que todavía algunas categorías presentan una baja recuperación, para estas categorías en específico, se observa que la tasa de recuperación negativa podía atribuirse a rasgos en la textura que identifican la imagen y la diferencias de las demás. Dichos rasgos se pierden al binarizar la imagen. Como trabajo futuro, se plantea la implementación de un descriptor que extraiga tanto rasgos de forma como de textura. Se espera que la

implementación de un descriptor con tales características sea capaz de describir con un mayor nivel de detalle los rasgos representativos de las imágenes.

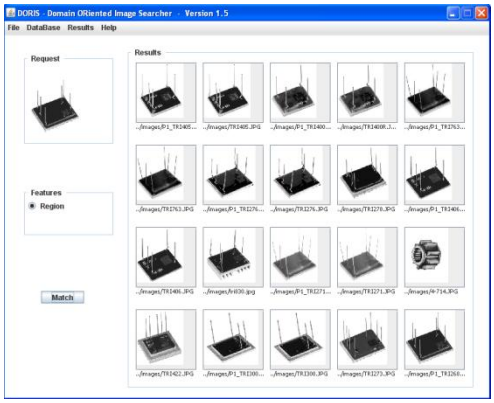

Fig. 3. Resultados arrojados por el sistema

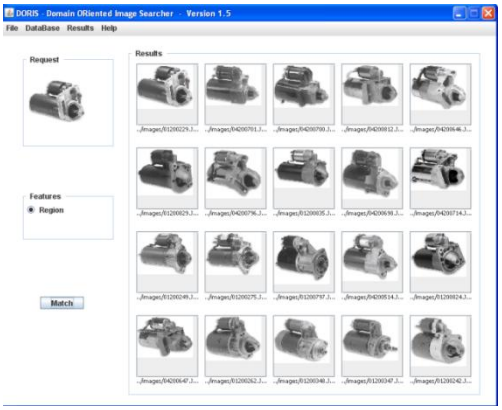

Fig. 4. Resultados arrojados por el sistema

Referencias

1 Adamek, T. and O'Connor, N. (2003). Efficient Contour-based Shape Representation and Matching. Multimedia Image Retrieval (MIR'03), pp. 138-143

2 Attneave, F. (1954). Some Information Aspects of Visual Perception. Psycological Review, vol. 61, pp. 183-193

3 Bertolo, H. (2005) Visual Imagery without Visual Perception? Psicológica, Vol. 26, No. , pp. 173-188

4 Chen, D. Z.; Daescu, O. (1998) Space-efficient Algorithms for Approximating Polygonal curves in two-dimensional space.

Proceedings 4th International Computing and Combinatorics Conference, Vol. 1449, pp. 55-64

5 Datta, R.; Joshi, D.; Li, J. and Wang, J. (2008). Image Retrieval: Ideas, Influences and Trends of the New Age. ACM Computing Surveys, Vol. 40, No. 2. Artículo 5.

6 Gibson, J. J (1950). The Perception of the Visual World. Greenwood Pub Group

7 Gibson. J.J (1950). The Perception of Visual Surfaces. American Journal of Psychology, Vol. 63, pp. 367-384

8 Han, S. and Fang, F. (2008). Linking Neural Activity to Mental Processes. Brain Imaging and Behavior, Springer, No. 2, pp. 242-248

9 Hebb, D. O. (1949). The Organization of Behavoir, John Willey.

10 Henriques, D. Y. P.; Flanders, M and Soechting, J. F. (2004) Distortions in the Visual Perception of Shape. Experimental Brain Research, Springer, Vol. 160, No. 3, pp. 384-393

11 Hu, X.; Kong, B.; Zheng, F. and Wang, S. (2007). Image Recognition Based on Wavelet Invariant Moments Neural Networks. Proceedings of the 2007 International Conference on Information Acquisition, pp. 275-279

12 Hu, M. K. (1962). Visual Pattern Recognition by Moments Invariants. IEEE Trans. on Information Theory, Vol. 8. No. 2, pp. 179-187

13 Jaramillo, G. E y Branch, J. W (2008). DORIS: Sistema para la Recuperación de Imágenes de Piezas Mecánicas y de Automoción Utilizando Descriptores de Textura. Revista de Avances en Sistemas e Informática. Vol. 5, no. 2, pp. 131-137

14 Jaramillo, G. E. y Branch, J. W. (2008) Sistema para la Recuperación de Imágenes Usando Descriptores de Forma. XIII Simposio de Tratamiento de Señales, Imágenes y Visión Artificial. STSIVA 2008, pp. 183-187

15 Kim, W-Y. and Kim, Y-S (2000). A Region-based Shape Descriptor Using Zernike Moments. Signal Processing Image Communication. Vol. 16, pp. 95-102

16 W. Y. Kim and P. Yuan. (1992). A Practical Pattern Recognition System for Translation, Scale and Rotation Invariance. IEEE Conference on Computer Vision and Pattern Recognition, pp. 391-396

17 J. Koenderink and A. Van Doorn (1986). Dynamic Shape. Biological Cybertenics, Vol. 53, pp. 383-396.

18 Kopf, S.; Haenselmann, T. and Effelsberg, W. (2005). Enhancing Curvature Scale Space Features for Robust Shape Classification. IEEE International Conference on Multimedia and Expo. Vol. 4, pp. 478-481

19 Lamouret, I.; Cornilleau-Pérès, V. and Droulez, J. (1997) Top-down Processes and the Visual Perceptions of Shape from Motion. Trends in Cognitive Science, Vol. 1, No. 2, pp. 43-44

20 Loncaric, S. (1998) A Survey of Shape Analysis Techniques. Journal on Pattern Recognition. Vol. 31, pp. 983-1001

21 Marr, D. and Poggio, T. (1979) A Computational Theory of Human Stereo Vision. Proceedings of the Royal Society of London, pp. 301-328

22 Marr, D and Hildreth, E. (1980). Theory of Edge Detection. Proceedings of the Royal Society of London, Series B, Biological Sciences, Vol. 207, No. 1167, pp. 187-217

23 Mokhtarian, F.; Abbasi, S. and Kittler, J. (1997) Efficient and Robust Retrieval by Shape Content through Curvature Scale Space. Image Databases and Multi-Media Search. A.W.M Smeulders and R. Jain eds, pp. 51–58

24 Pavlidis, T. and Iri, M. (1986). Computational-Geometric Methods for Polygonal Approximations of a Curve. Computer Vision, Graphics, Image Processing, Vol. 36, pp. 31-34

25 Rosin, P.L. (1996). Techniques for Assessing Polygonal Approximations of Curves. IEEE Transaction on Pattern Analysis and Machine Intelligence. Vol. 19, pp. 659-666

26 Rui, Y. and Huang, T.S. (1999). Image Retrieval: Current Techniques, Promising Directions, and Open Issues. Journal of visual Communication and Image Representation, Vol. 10, pp. 39-625

27 Sikora, T. (2001) The MPEG-7 Visual Standard for Content Description – An Overview. IEEE Transactions on Circuits and Systems for Video Technology. Vol. 11, No. 6, pp. 696 - 702

28 Sim, D-G.; Kim, H-K. and Park, R-H. (2004). Invariant Texture Retrieval Using Modified Zernike Moments. Image and Computing. No. 22, pp. 331-342

29 Smeulders, A. W. M.; Worring, M; Santini, S.; Gupta, A. and Jain, R. (2000). Content-Based Image Retrieval at the End of the Early Years. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 12, pp. 1349 - 1380

30 Sochacki, S.; Richard, N. and Bouyer, P. (2007) A Comparative Study of Different Statistical Moments Using an Accuracy Criterion. 7th IAPR International Workshop on graphics recognition – GREC 2007.

31 Kopf, S.; Haenselman, T. and Effelsberg, W. (2005). Enhancing Curvature Scale Space Features for Robust Shape Classification. IEEE International Conference on Multimedia and Expo.

32 Teague, M. (1980) Image Analysis via the General Theory of Moments. Journal Opt. Society American, Vol. 70, No. 8, pp 920-930.

33 Verstraelen, L (2005). A Geometrical Description of Visual Perception. Kragujevac Journal of Mathematics. Vol. 28, pp. 7-17

34 Wertheimer, M. Laws of Organization in Perceptual Forms (1923). W. D. Ellis. A Source Book of Gestalt Psychology. Harcourt Brace Jovanovic.

35 Zhang, D. and Lu, G. (2001). Content-Based Shape Retrieval Using Different Shape Descriptors – A comparative Study. IEEE International Conference on Multimedia and Expo.

36 Zhang, D. and Lu, G. (2003). Evaluation of MPEG-7 Shape Descriptors against Other Shape Descriptors. Multimedia Systems, Springer-Verlag, Vol. 9, pp. 15-30.

37 Zhong, B. and Liao, W. (2004). A Hybrid Method for Fast Computing the Curvature Scale Space Image. Proceedings of the geometric modeling and Processing.