

# Características locales de forma para la clasificación de objetos

## *Local shape features for object categorization*

Leonardo Chang Fernández y José Hernández Palancar

Centro de Aplicaciones de Tecnología de Avanzada (CENATAV),  
7a # 21406 e/ 214 y 216, Rpto. Siboney, Playa, C.P. 12200, La Habana, Cuba.  
{lchang, jpalancar}@cenatav.co.cu

**Resumen:** Desde los inicios de las investigaciones en el campo del reconocimiento de objetos, la forma ha sido ampliamente utilizada. Esto ha estado motivado, fundamentalmente, por ser este tipo de características la más natural a los seres humanos en el proceso de reconocer categorías de objetos. Durante este proceso es necesario lidiar con variaciones de escala, rotación y traslación, así como con problemas de oclusión parcial y ruido en los contornos. En este trabajo se propone un método de extracción y descripción de características locales de forma robusto a estos cambios, además de un método de clasificación de objetos usando las características propuestas. El mismo fue probado en un subconjunto de la base de datos MPEG-7 CE-Shape-1 obteniéndose una efectividad promedio de 99.3 %.

**Palabras clave:** Clasificación de objetos, descriptores de forma, características locales invariantes de forma.

**Abstract:** Since the beginning of research in the field of object recognition, shape features have played an important role. This has been mainly motivated by the fact that such features are the most natural to human beings in the process of recognizing object categories. During this process it is necessary to deal with variations of scale, rotation and translation, as well as partial occlusion and noise problems. This paper proposes a method of extraction and description of local features robust to these changes, and a method of object classification using the proposed features. The method was tested on a subset of the database MPEG-7 CE-Shape-1 yielding an average effectiveness of 99.3 %.

**Keywords:** Object classification, shape descriptors, shape local invariant features.

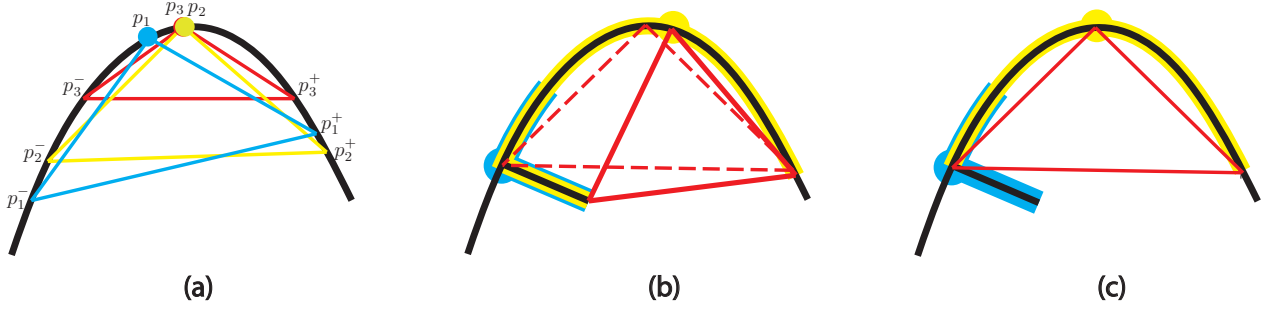
## 1. Introducción

La clasificación de objetos es uno de los problemas más antiguos dentro del campo de la visión por computadoras. Sin embargo, se mantiene como uno de los problemas más difíciles, debido a la naturaleza indefinida de la similitud entre las clases de objetos. Los descriptores de formas son una de las herramientas comúnmente utilizadas en aplicaciones de clasificación de objetos. Los descriptores de formas son funciones matemáticas que se aplican a una imagen y producen valores numéricos que son representativos de una característica particular de la imagen. Estos valores numéricos pueden entonces ser procesados a fin de proporcionar alguna información sobre la imagen. Por ejemplo, estos valores pueden servir como entrada a un clasificador con el fin de asignar una etiqueta de clase a la imagen.

La clasificación de objetos usando características de forma es un problema que involucra grandes retos. Las formas a comparar son usualmente el resultado de algún proceso de segmentación, el cual (siendo imperfecto) puede introducir una cantidad considerable de ruido, la cual necesita ser tolerada. Los cambios de escala, rotación y traslación tampoco deben afectar el proceso de clasificación. Además, otras complicaciones adicionales pueden ser ocasionadas por la oclusión parcial de los objetos. Finalmente, en aplicaciones reales, todos los factores anteriormente mencionados no aparecen de manera aislada, sino que contribuyen en conjunto a aumentar el nivel de complejidad de la tarea.

En el contexto de este trabajo se aborda el tema de la clasificación de objetos usando forma considerando los problemas mencionados con anterioridad. En el mismo se propone un método de extracción y descripción de características locales invariante a cambios de escala, rotación y traslación. Las características extraídas y sus descripción solo dependen de una vecindad, por lo que su carácter local las dota de cierta invariabilidad a oclusión parcial y ruido en el proceso de segmentación.

Varios esfuerzos han sido dedicados por la comunidad científica al problema de la representación de forma. Según [15][13], estos métodos pueden ser clasificados en métodos basados en el contorno y métodos basados en la región. Esta clasificación se basa en si las características son extraídas solamente del contorno de la



**Figura 1.** Detección de fragmentos de contorno: (a) son fragmentos candidatos todos aquellos segmentos de contornos donde se pueda inscribir un triángulo con apertura menor que  $\alpha_{max}$  y lados incidentes en el contorno con longitudes entre  $d_{min}$  y  $d_{max}$ . Varios triángulos pueden ser encontrados sobre un mismo punto o sobre puntos cercanos, se selecciona el más agudo en una vecindad. (b) El ruido en el contorno puede introducir falsos fragmentos (fragmento de contorno en amarillo). (c) Para contrarrestar este fenómeno se agrega como restricción que los triángulos candidatos solo crecerán hasta encontrar otra esquina.

forma (e.g. [3][14][11][2]) o si son extraídas de toda la región de la forma (e.g. [6][9][10][16]) respectivamente. A su vez, estos métodos se dividen en enfoques globales y enfoques estructurales. Esta sub-clase se basa en si la forma será representada en su totalidad o representada por segmentos o secciones. Esta es la clasificación más general y de más amplio uso, no obstante, según [15], también se pueden clasificar en métodos basados en dominio espacial y dominio de transformaciones. Esta clasificación se basa en si las características son extraídas en el dominio espacial o en el dominio de transformaciones.

## 2. Método de extracción de características locales de forma propuesto

Estudios psicológicos [5][8] muestran que los humanos somos capaces de reconocer categorías de objetos a partir de fragmentos de contornos y bordes (Ver Figura 3). De ahí que si se seleccionan los fragmentos adecuados del contorno<sup>1</sup> de un objeto estos son representativos del mismo.

Las líneas rectas son poco discriminativas ya que estas solo están definidas por su longitud (lo que es poco útil cuando se busca invariabilidad a la escala). Sin embargo las curvas ofrecen una descripción más rica del objeto ya que estas están definidas, además de por su longitud, por su ángulo de curvatura (una línea puede verse como un caso específico de una curva, i.e. una curva con curvatura nula). Además, frente a distintas variaciones como cambios de escala, rotación, traslación, rotaciones afines y cambios de iluminación y textura, las curvas tienden a permanecer presentes. En este trabajo proponemos usar como características fragmentos de contornos definidos por fragmentos de curvas.

Para detectar fragmentos del contorno de alta curvatura o esquinas (como también le llamaremos de manera indistinta) nos basamos en el método propuesto por Chetverikov [7]. El método de Chetverikov inscribe triángulos en un segmento de puntos del contorno y evalúa el ángulo de su vértice del medio el cual debe ser menor que  $\alpha_{max}$ . Los lados del triángulo que inciden en el vértice del medio se requieren que tengan longitud mayor a  $d_{min}$  y menor a  $d_{max}$ . Varios triángulos pueden ser encontrados sobre un mismo punto o sobre puntos adyacentes en una misma curva, por lo tanto se selecciona el punto de mayor curvatura. Cada fragmento de contorno  $i$  seleccionado está definido entonces por un triángulo  $(p_i^-, p_i, p_i^+)$ , donde  $p_i$  es el vértice del medio y los puntos  $p_i^-$  y  $p_i^+$  definen los límites del fragmento del contorno. Ver Figura 1 (a).

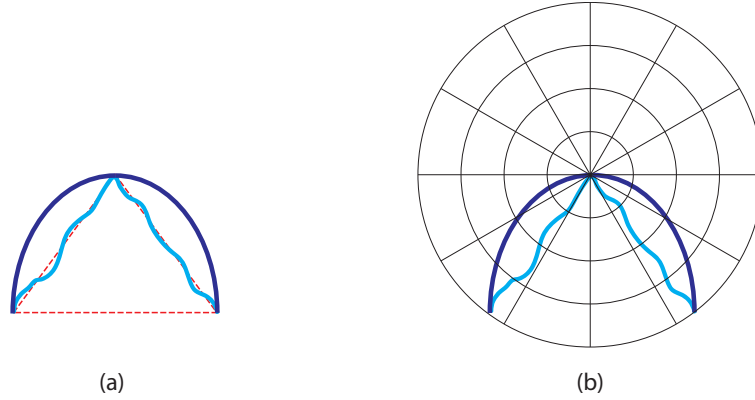
El detector de esquinas en contornos de Chetverikov tiene la desventaja de ser poco estable ante ruido en el contorno o en presencia de contornos muy ramificados, lo que puede ocasionar que se seleccionen falsas esquinas debido a la introducción de triángulos más agudos por el ruido. Véase como ejemplo la Figura 1 (b). Para eliminar este problema, introducimos además como restricción en el proceso de extracción de Chetverikov que el triángulo candidato  $(p_k^-, p_k, p_k^+)$  crecerá mientras que los puntos  $p_k^-$  y  $p_k^+$  no coincidan con el punto  $p_j$  de otra esquina. Esto evitará la introducción de falsas esquinas debido al ruido o contornos muy ramificados. En la Figura 1 (c) se muestra como esta restricción resuelve el error en el ejemplo de la Figura 1 (b).

<sup>1</sup> Definimos como contorno a las líneas que conforman la silueta y los bordes internos del objeto

Entonces, cada característica  $\varsigma_i$  extraída del contorno está definida por  $\langle P_i, T_i \rangle$ , donde  $T_i = (p_i^-, p_i, p_i^+)$  es el triángulo inscrito en el fragmento de curva y  $P_i = \{p_1, \dots, p_n\}, p_j \in \mathbb{R}^2$  es el conjunto de  $n$  puntos que forman el fragmento de contorno  $\varsigma_i$ , ordenado de modo tal que el punto  $p_j$  es adyacente al punto  $p_{j-1}$  y  $p_{j+1}$  y los puntos  $p_1, p_n \in P_i$  coinciden con los puntos  $p_i^-, p_i^+ \in T_i$  respectivamente.

### 3. Método de descripción propuesto

La propia definición de esquina brindada por el proceso de extracción (específicamente el triángulo  $(p_i^-, p_i, p_i^+)$ ) provee una descripción de la misma ya que brinda indicios de amplitud, orientación y longitud, no obstante consta de poco poder discriminativo ya que curvas distintas pueden compartir el mismo triángulo. En la Figura 2 (a) se muestra un ejemplo de este fenómeno.



**Figura 2.** En (a) se muestra como dos curvas distintas (azul claro y azul oscuro) pueden compartir un mismo triángulo (rojo en líneas discontinuas), por lo que estos triángulos son una descripción poco discriminativa de los fragmentos de contorno. En (b) se puede apreciar como representar cada fragmento de contorno en un espacio polar discretizado de origen de coordenadas  $p_i$  es una representación más discriminativa. En este ejemplo  $n_r = 4$  y  $n_\theta = 12$ .

Con el objetivo de brindarle un mayor poder discriminativo a las características extraídas, proponemos representar cada fragmento de contorno en un espacio polar de origen de coordenadas  $p_i$ , donde la longitud  $r$  y orientación  $\theta$  de cada vector son discretizados para formar un histograma bidimensional de  $n_r \times n_\theta$  bins y que será el que describa a cada fragmento de contorno:

$$H_i(b) = |\{w \in P_i : (w - p_i) \in \text{bin}(b)\}| \quad .$$

La idea gráfica de este descriptor se muestra en la Figura 2 (b). Nótese que para un número suficientemente grande de  $n_r$  y  $n_\theta$  esta es una representación exacta del fragmento de contorno.

El uso de este tipo de descriptor está motivado, además de por su poder descriptivo, por su invariabilidad a la rotación y la posibilidad de hacerlo invariante a la escala mediante el muestreo de los fragmentos de contorno a un número fijado  $M$  de puntos.

Un método de representación de formas muy conocido que usa este tipo de descripción es el *shape context* de Belongie *et. al.* [4], donde se usa para obtener una descripción global del objeto a partir de su contorno expresando las relaciones espaciales de cada punto con el resto. En [4] prefieren usar bins en un espacio log-polar para hacerlo más sensible a puntos más cercanos.

### 4. Invariabilidad y robustez del método propuesto

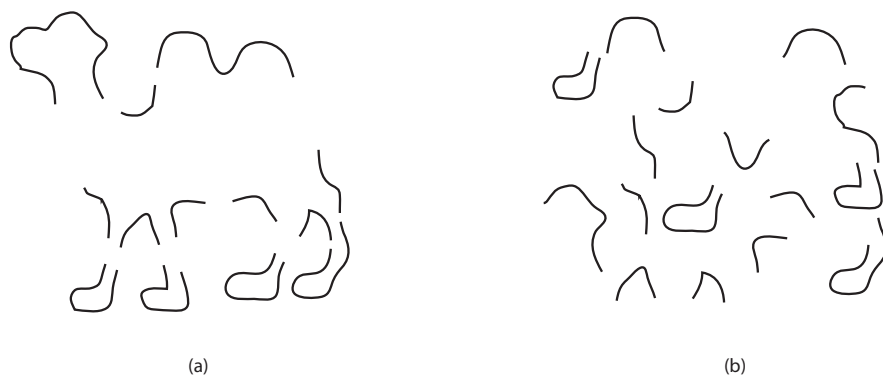
Como se planteó el descriptor de forma propuesto debía cumplir las propiedades de localidad, invariabilidad a traslación, rotación y escala. Las características de forma propuestas cumplen con estas propiedades:

- **Local:** la propiedad de localidad se cumple de manera directa a partir de las definiciones de fragmento de contorno y de su descriptor dadas en las Secciones 2 y 3. Una esquina y su descriptor solo dependen de un punto y el conjunto de puntos en una vecindad mucho menor que el área de la imagen, por ende, tanto en el proceso de extracción como de descripción, un cambio o variación en una porción del contorno (producido, por ejemplo, por ruido en el contorno, oclusión parcial u otra deformación del objeto), solo afectará a las características extraídas de esa porción.
- **Invariable a traslación:** tanto el proceso de extracción como de descripción de las características son inherentemente invariantes a la traslación ya que se basan en coordenadas relativas a los puntos de interés.
- **Invariable a rotación:** el proceso de extracción de esquinas es invariable a la rotación ya que una esquina es un triángulo inscrito en un segmento de contorno, lo que solo depende de la forma del contorno y no de su orientación.  
En los descriptores, se logra invariabilidad a la rotación mediante la rotación de sus sistemas de coordenadas de modo tal que queden alineados por la bisectriz del vértice  $p_i$  de cada triángulo.
- **Invariable a escala:** la invariabilidad a la escala se logra, en el proceso de extracción, detectando características en el espacio de la escala, o simplemente detectando esquinas a distintos valores de  $d_{min}$  y  $d_{max}$ .  
En el proceso de descripción esta invariabilidad se logra mediante el muestreo de los fragmentos de contorno (i.e.  $P_i$ ) a un número fijado  $M$  de puntos o normalizando los histogramas.

## 5. Representación de categorías de objetos usando el descriptor de forma propuesto

En las secciones anteriores se presentó un descriptor de forma local invariante a traslación, rotación y escala. En esta sección presentamos un método para describir categorías de objetos a partir de nuestras características de forma propuestas.

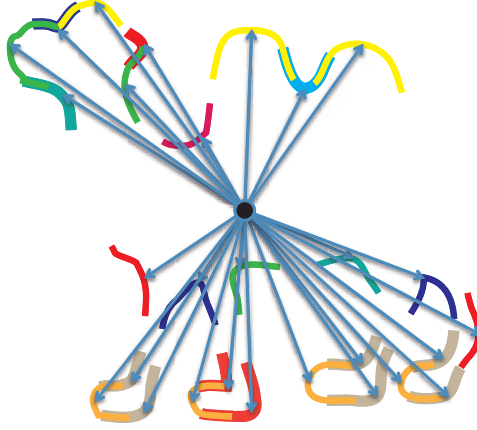
En la Sección 2 se mencionó que estudios psicológicos [5][8] muestran que los humanos somos capaces de reconocer categorías de objetos a partir de fragmentos de contornos y bordes, pero además, estos estudios señalan que dichos fragmentos de contorno deben estar espacialmente organizados. Véase la Figura 3 (a) en contraposición a la Figura 3 (b).



**Figura 3.** Los humanos somos capaces de reconocer objetos a partir de (a) fragmentos de su contorno, pero estos deben estar organizados espacialmente. Nótese en (b) cómo es más difícil el reconocimiento si los contornos no están espacialmente organizados.

Motivado por este hecho, usamos un modelo donde cada fragmento de contorno es ordenado espacialmente alrededor de un único punto, el centro del objeto. Combinado, además, con un enfoque de bolsas de palabras visuales con el objetivo de obtener prototipos y una representación más compacta de las características, y lograr abstraer una clase de objetos y no objetos en específico. Entonces, el modelo para cada categoría de objeto, expresará la probabilidad de que exista una característica de tipo  $k$  en la posición  $(r, \theta)$  alrededor del centro del objeto. En la Figura 4 se muestra gráficamente esta idea.

Para obtener el vocabulario visual de formas, se realiza un agrupamiento sobre todas las características extraídas de todas las imágenes de todas las clases en el conjunto de entrenamiento. El centroide de cada grupo constituirá una característica prototipo.



**Figura 4.** El modelo para una clase de objetos propuesto expresa la probabilidad de que exista una característica de tipo  $k$  (las características correspondientes a una misma palabra visual son representadas con el mismo color en la figura) en la posición  $(r, \theta)$  (representada por los vectores) alrededor del centro del objeto (círculo negro).

En el conjunto de entrenamiento, cada imagen es etiquetada con un rectángulo  $\rho = (\rho_a, \rho_b)$  que delimita al objeto, donde  $\rho_a$  y  $\rho_b$  representan los puntos superior izquierdo e inferior derecho, respectivamente, del rectángulo delimitador del objeto. El rectángulo  $\rho$  define de manera implícita el centro del objeto  $c_\rho = \frac{\rho_a + \rho_b}{2}$ , así como la escala del objeto definida por la distancia entre el centro del rectángulo y una de sus esquinas. De cada muestra en el conjunto de entrenamiento son extraídas sus características y ordenadas espacialmente alrededor del centro del objeto mediante los vectores que unen al centro del objeto  $c_\rho$  (que pasará a ser el origen de coordenadas) y al punto  $p_i$  de cada fragmento de contorno. También se determina a qué característica prototipo (o palabra visual) pertenece cada característica extraída.

Luego, para construir el modelo de cada clase, las coordenadas  $(r_i, \theta_i)$  de las características extraídas de todas las muestras de la clase en el conjunto de entrenamiento son discretizadas para formar, junto a la palabra visual a la que pertenece cada característica, un histograma tridimensional de  $m_r \times m_\theta \times K$  bins, donde  $m_r$  y  $m_\theta$  es el número de intervalos al cual fue discretizada la longitud y orientación del espacio polar y  $K$  es la cantidad de palabras en el vocabulario visual. Todos los modelos son llevados a escala  $s = 1$  y los histogramas normalizados por la cantidad de imágenes en el conjunto de entrenamiento para esa clase.

## 6. Resultados experimentales

Para probar la efectividad del método de extracción y descripción de características de forma y la representación de clases de objetos basados en estas características propuestos en este trabajo, realizamos un conjunto de experimentos en la base de datos MPEG-7 CE-Shape-1[12][1]. Esta es una base de datos diseñada para evaluar algoritmos de descripción de formas. En cada imagen aparece solamente un objeto, centrado en la imagen y definido por su silueta. Consta de 70 clases de objetos y 20 instancias por cada clase.








































































Para nuestros experimentos, seleccionamos aleatoriamente 10 de las 70 clases de la base de datos. Las clases seleccionadas usadas, así como ejemplos de imágenes de cada clase son mostradas en la Tabla 1. De las 20 imágenes por clase, seleccionamos aleatoriamente 8 para el entrenamiento y las restantes 12 para prueba.

Para formar el vocabulario de palabras usamos  $K$ -means con  $K = 100$ , como clasificador utilizamos un 1-NN, donde usamos la distancia chi-cuadrado ( $\chi^2 = \sum_i \frac{(x_i - y_i)^2}{x_i + y_i}$ ) para obtener la matriz de semejanza entre los histogramas de los modelos y los extraídos de cada imagen.

La Tabla 2 muestra la matriz de confusión para los experimentos realizados, la cual resume los resultados de clasificación. Como se puede apreciar, los mayores problemas estuvieron al confundir dos camellos con ganado. Esto se debió a que estas clases comparten varias palabras visuales en las regiones de las patas y la cabeza.

Para medir el desempeño se usaron las medidas de recuerdo, precisión, tasa de verdaderos negativos y exactitud. La Tabla 3 muestra los valores de cada una de estas medidas para cada una de las clases.

**Tabla 1.** Clases de la base de datos MPEG-7 CE-Shape-1 utilizadas en nuestros experimentos e imágenes de ejemplo por cada clase.

Clase	Imágenes de ejemplos							
(1) Murciélago								
(2) Escarabajo								
(3) Campana								
(4) Ave								
(5) Hueso								
(6) Camello								
(7) Ganado								
(8) Carro clásico								
(9) Taza								
(10) Venado								

**Tabla 2.** Matriz de confusión para los experimentos realizados.

[illegible]

**Tabla 3.** Medidas de desempeño por clase para los experimentos realizados. Los altos valores de estas medidas muestran la repetitividad y discriminatividad de las características propuestas, así como poder de generalización de cada modelo de clase.

	Recuerdo (%)	Precisión (%)	TNR (%)	Efectividad (%)
Murciélago	100	100	100	100
Escarabajo	100	100	100	100
Campana	100	85.7	98.1	98.3
Ave	100	100	100	100
Hueso	91.6	100	100	99.2
Camello	75	100	100	97.5
Ganado	100	85.7	98.1	98.3
Carro clásico	100	100	100	100
Taza	100	100	100	100
Venado	100	100	100	100
Promedio	96.7	97.1	99.6	99.3

El recuerdo mide la proporción de muestras positivas que son clasificadas como tal:

$$recuerdo = \frac{tp}{tp + fn}.$$

La precisión se define como la relación entre los resultados verdaderos positivos y todos los positivos:

$$precisión = \frac{tp}{tp + fp}.$$

La tasa de verdaderos negativos (TNR) expresa la proporción de muestras negativas que son clasificadas como tal:

$$TNR = \frac{tn}{fp + tn}.$$

La exactitud es la razón de resultados positivos (tanto en muestras positivas como negativas) dentro de la población.

$$exactitud = \frac{tp + tn}{tp + tn + fp + fn}.$$

Los altos valores de estas medidas muestran la repetitividad y discriminatividad de las características de forma y su descriptor propuestos en este trabajo. Además muestran la capacidad del modelo propuesto de generalizar una clase de objetos a partir de pocas muestras de entrenamiento.

## 7. Conclusiones y trabajo futuro

En este trabajo se presentó un método para la extracción y descripción de características de forma. El método propuesto tiene como ventajas ser invariante a la rotación, escala y rotación; además de ser local, lo que le brinda cierta invariabilidad a oclusión parcial y ruido. Se propuso también un método de clasificación de objetos usando dichas características, el mismo se probó en la base de datos MPEG-7 CE-Shape-1 obteniéndose altas tasas de desempeño.

Como trabajo futuro se recomienda adaptar y extender el uso de estas características a imágenes naturales (no a siluetas), partiendo de bordes o de los gradientes de la imagen.

## Referencias

1. Mpeg-7 ce-shape-1 dataset. available to download in <http://knight.cis.temple.edu/~shape/mpeg7/mpeg7dataset.zip>.
2. H Asada and M Brady. The curvature primal sketch. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(1):2–14, 1986.

3. S Belongie, J Malik, and J Puzicha. Matching shapes. In *Proceedings Eighth IEEE International Conference on Computer Vision ICCV 2001*, volume 1, pages 454–461. IEEE, IEEE Comput. Soc, 2001.
4. S Belongie, J Malik, and J Puzicha. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):509–522, 2002.
5. I Biederman and G Ju. Surface versus edge-based determinants of visual recognition. *Cognitive Psychology*, 20(1):38–64, 1988.
6. H Blum. A transformation for extracting new descriptors of shape. *Models for the perception of speech and visual form*, 19(5):362–380, 1967.
7. Dmitry Chetverikov. A Simple and Efficient Algorithm for Detection of High Curvature Points in Planar Curves. *Proceedings of the 23rd Workshop of the Austrian Pattern Recognition Group*, pages 746–753, 2003.
8. Joeri De Winter and Johan Wagemans. Contour-based object identification and segmentation: stimuli, norms and data, and software tools. *Behavior research methods instruments computers a journal of the Psychonomic Society Inc*, 36(4):604–624, 2004.
9. Ming-Kuei Hu. Visual pattern recognition by moment invariants. *Information Theory IRE Transactions on*, 8(2):179–187, 1962.
10. W Y Kim and Y S Kim. A region-based shape descriptor using Zernike moments. *Signal Processing Image Communication*, 16(1-2):95–102, 2000.
11. M Kliot and E Rivlin. Invariant-Based Shape Retrieval in Pictorial Databases. *Computer Vision and Image Understanding*, 71(2):182–197, 1998.
12. L J Latecki, R Lakamper, and T Eckhardt. Shape descriptors for non-rigid shapes with a single closed contour. *Proceedings IEEE Conference on Computer Vision and Pattern Recognition CVPR 2000 Cat NoPR00662*, 1:424–429, 2000.
13. José M Martínez. MPEG-7 Overview, 2004.
14. Peter J. Van Otterloo. *A contour-oriented approach to shape analysis*. Prentice Hall International (UK) Ltd., Hertfordshire, UK, UK, 1991.
15. D Zhang. Review of shape representation and description techniques. *Pattern Recognition*, 37(1):1–19, 2004.
16. Dengsheng Zhang and Guojun Lu. Generic Fourier descriptor for shape-based image retrieval. *Proceedings IEEE International Conference on Multimedia and Expo*, 1(1):425–428, 2002.