A Novel Shape Descriptor over Multi-Scale Edge Field: 2D Walking Ant Histogram

M. Ferreira, S. Kiranyaz and M. Gabbouj
Tampere University of Technology, Finland
{miguel.ferreira, serkan.kiranyaz, moncef.gabbouj}@tut.fi

Abstract - **A generic shape descriptor, which can be extracted from the major object edges in an automatic way and used for the content-based retrieval in multimedia databases, is presented. Via adopting a multi-scale approach over the edge field where the scale represents the amount of simplification, the most relevant edge segments, as we refer to *sub-segments*, which eventually represent the major object boundaries, are extracted from a *scale-map*. Adopted from an imaginary concept of a walking ant with a limited line of sight over the boundary of a particular object, we traverse through each *sub-segment* and describe a certain line of sight, whether it is a continuous branch or a corner, using individual 2D histograms. As a result integrating the whole process as feature extraction module into MUVIS framework allows us to test the mutual performance of the proposed shape descriptor in the context of multimedia indexing and retrieval.**

1. INTRODUCTION

As long as the object (shape) information extracted properly and semantically intact, it becomes one of the most powerful descriptor in the context of CBIR. Several descriptors such as Fourier transform [15], wavelet transform [11], Hough transform [8], curvature scale space [2], Zernike moments [12], etc., can conveniently be extracted either over the shape boundaries or the entire area (the region of the object shape). Most of these methods achieve a significant performance in terms of retrieval efficiency and accuracy in binary shape databases; however, especially in large multimedia databases containing ordinary images or video clips, extraction of the true shape information of the natural objects first requires an automatic and highly accurate segmentation, which is still an open and ill-posed problem because the semantic objects in natural images do not usually correspond to homogenous spatial regions in color or texture. Henceforth, there is a lack of a generic and robust shape descriptor and the existing methods are merely applicable on such databases where object shapes are extracted manually (e.g. binary shape databases). Alternatively, the efforts are mainly focused on edge-based approaches since the edge field in an image usually represents both object boundaries and texture. MPEG-7 Edge Histogram (EHD) [9] generates a histogram of the main edge directions (vertical, horizontal and two diagonals) within fixed size blocks. It is an efficient texture descriptor for the images with heavy textural presence. It can also work as a shape descriptor as long as the edge field contain the true object boundaries and not saturated by the background texture. In this case the method is particularly efficient on describing geometric objects due to its block-based edge representation only with four directions. A similar but pixel-based method applied directly over Canny edge field [3] is called Histogram of Edge Directions (HED) [1]. Another approach, so called Angular Radial Partitioning (ARP), is presented in [5]. ARP basically works over radial blocks (angular slices from quantized radial steps from the center of mass of a re-scaled image). Although rotation invariance can be obtained within this method, the shape outlines are degraded due to the loss of aspect ratio during re-scaling of the image into square dimensions to fit a surrounding circle. A promising method, Edge Pixel Neighborhood Histogram (EPNH) [4], creates a 240-bin histogram from the direction of the neighbor edge pixels. Although it can describe one-pixel neighborhood over the entire edge field, it exhibits a comparable performance to MPEG-7 EHD. Nevertheless, all these methods turn out to be texture descriptors since they cannot discriminate the true object boundaries that are usually suppressed from the surrounding texture edges.

In order to extract the true object boundaries among a crowded edge field due to such textural dominance within arbitrary images, recently a multi-scale approach is proposed [6] where the scale represents the amount of simplification using Bilateral Filter [13]. It relies on a basic assumption, that is, the lowest-scale (Canny) edge detector captures most edges in the image, including the boundary edges of the objects of interest. Henceforth, erroneous results are thus encountered whenever this assumption fails, i.e. textural predominance, insignificant intensity difference between object and background or some illumination effects such as strong shades are typical cases where only the human visual perception can exhibit the usage of "knowledge" or "intelligence" to interpolate missing object edges. Note that most of the object edges can still be extracted from the scenery but the method cannot link a major gap or gaps whenever they are beyond the allowed (maximum) tolerance inherited from the Gestalt law [ref], *proximity* and hence the object (boundary) extraction cannot be performed accurately on such circumstances.

In this paper we address this problem by proposing a shape descriptor, which works as long as the majority of object edges are available yet the full object (boundary) extraction may or may not be possible. So the main advantage of the proposed method is that it can still describe a shape from its rough boundary sketch with some missing parts. If extracted directly from a natural image, the edge field might also be incomplete, missing, noisy or saturated as well as the background or the edges of object texture, which are not related with the shape information at all and they significantly disturb the "true" object edges (boundaries). In brief, natural images are usually too "detailed" to achieve an accurate shape extraction over the edge field. Therefore, we adopt a similar approach as in [6] in order to extract the relevant *sub-segment*s from the scale-map as the first step and then a novel shape description, as referred to 2D Walking Ant Histogram (WAH), is applied over them. It is basically motivated from the following imaginary scenario; suppose an ant walking over a solid object and every once in a while, say in a few steps, it "describes" its "Line of Sight (LoS)" in a convenient way. So cumulating all the intermediate LoS descriptions in a (2D) histogram, particularly focusing on continuous branches and major corners, yield an efficient cue about the shape. Such a description is still feasible if some portion of the object boundary is missing and this is essentially the major advantage of this method. The description frequency (i.e. how often the ant makes a new –intermediate-description) and the length of LoS will obviously be the two major parameters of this scheme. The third one is the amount (number) of

relevant *sub-segment*s that are taken into consideration (description). Keeping this number sufficiently low yield the method to describe only the major object(s') boundaries whilst discarding the rest.

The proposed method is fully automatic (i.e. without any supervision, feedback or training involved). Forming the whole process as a *FeX* module into MUVIS framework [7], [10], allows us to test the overall performance in the context of multimedia indexing and retrieval. This paper is organized as follows: Section 2 presents an overview about the proposed method, focusing on the extraction of (most) relevant *sub-segment*s. We discuss the formation of the 2D WAH from the relevant *sub-segment*s and implementation of the proposed method as a *FeX* module used for indexing and retrieval of MUVIS multimedia databases in Section 3. Experimental results are given in 4 and Section 5 concludes the paper.

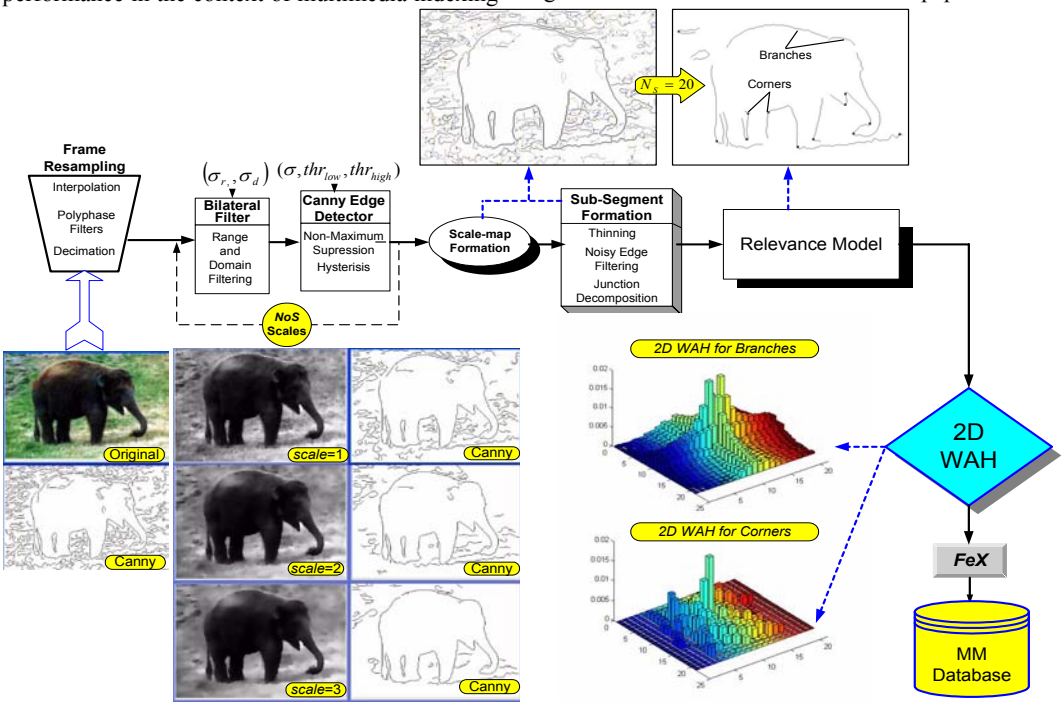

Figure 1: Overview of 2D WAH algorihm

2. MULTI SCALE *SUB-SEGMENT* FORMATION

As Figure 1 illustrates the general overview, the proposed method is primarily designed for the shape-based retrieval from general purpose multimedia databases containing indefinite number of arbitrary image and video clips possibly in a compressed format, with varying dimensions and resolutions. All the analytical steps are performed using the luminance component of the frames extracted (decoded) either from the images or (key-) frames of the video clips. In order to extract (most) relevant *sub-segment*s over which the 2D WAH description is applied, a certain preliminary phase is performed, which mainly consists of four major parts: Frame resampling (size transformation), Bilateral filtering and *scale-map* formation over Canny edge field, *sub-segment* formation and analysis and finally the selection of the relevant *sub-segment*s using a relevance model. Once the required number of relevant *sub-segment*s are selected for both branches and corners, 2D WAH is formed distinctively. Feature eXtraction (*FeX*) process transforms the 2D WAH in a suitable descriptor (feature vector) for MUVIS indexation operation.

The first and natural step is resampling (frame size transformation with respect to the original aspect ratio) into a practical dimension range, which is sufficiently big for perceivable shapes but small enough not to cause infeasible analysis time. The typical figures can be in between 200 to 400 pixels for both dimensions. By this way a standard approach can be applied since it then becomes invariant to any resolution variations. Bilateral filter is especially efficient for removing the details when used iteratively. Each iteration further removes the details from the image and eventually reaches a stable solution. Since the strong edges remain with good localization, this information can be gathered in a *scale-map* and the *scale* can then be used as the amount of simplification of an image from which the edges are extracted. Canny edge detection is applied for each scale (after Bilateral filtering), however its low pass filter (*Gaussian*) is only applied once to the input image in order not to cause excessive blurring on the higher scales. Once Canny edge detection is performed for each scale, the scale-map can then be formed over the edge field of the lowest scale where the edge field is complete. As mentioned earlier, the *scale-map* assigns the maximum scale factor to each edge pixel as long as it prevails and therefore, both (edge) localization and scale information can be obtained from the *scale-map*. From the scale-map, the sub-segments are formed. A *sub-segment* can basically be defined as the series of connected edge pixels with two end-points and they are formed after the following pre-processing steps directly applied onto edge field of the *scale-map*: *One-Pixel Thinning, "Noisy Edge" Removal, Junction Decomposition.* More detailed information about *sub-segment* formation can be obtained in [6].

As mentioned earlier the relevant *sub-segments*, which bear the major object edges are usually longer with higher scale weights. On contrary the irrelevant ones, which belong to details (texture, noise, illumination varitations, etc.) are usually shorter with a lower scale weights since Bilateral filter is likely to remove them in earlier iterations (e.g. see Figure 1). Therefore, the relevancy, *R,* of a *sub-segment SS*, can then be expressed as follows:

$$R(SS) = W(SS) \times L(SS) \qquad (1)$$

where $W(SS)$ is the scale weight and $L(SS)$ is the length (total number of edge pixels) of *SS*. This model obviously favours long and highly relevant (having higher *scale* weights) *sub-segments* which

can either be closed loop (CL) or non-closed-loop (NCL). As mentioned earlier, in order to tune the proposed method as a shape descriptor N_S can initially be set to a sufficiently big number (i.e. $N_S \geq 10$) in order not to loose relevant *sub-segments;* however not too big to avoid the significant disturbance of the irrelevant segments. As a result, only N_S *sub-segments* with the highest $R(SS)$ values are selected for the extraction of 2D WAH, as explained in the following section.

3. FORMATION OF 2D WALKING ANT HISTOGRAM

The relevant N_S *sub-segments,* whether in CL or NCL form and all in one-pixel thick, are used for the formation of 2D WAH, which is the union of two 2D histograms, each with equal dimensions. For instance for a NCL *sub-segment*, the ant with a certain LoS (say L_S pixels long) and a walking pace Δ (i.e. $\Delta < L_S$), traverses from one end-point to the other, whilst describing each LoS section into the appropriate WAH, via incrementing the matching 2D bins, which actually represent the pixel grid centred from the first and last pixels of LoS. For a CL *sub-segment*, the traversing can start from any point and terminates whenever it looped-back to the start.

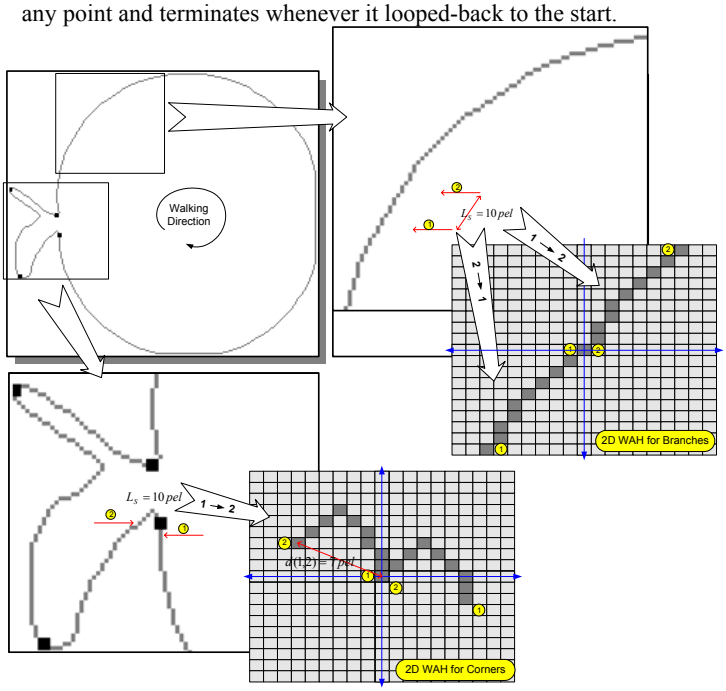

Figure 2: Localization of a branch and corner (LoS) sections into their associated 2D WAH.

Let p_1 and p_2 be the first and the last pixels of LoS and the number of pixels in between is the pixel length of LoS (i.e. $L(p_1 \rightarrow p_2) = L(p_2 \rightarrow p_1) = L_S$). In order to achieve an invariant representation with respect to traversing (ant's walking) direction, both of the LoS sections (i.e. $\{\forall p : p \in (p_1 \rightarrow p_2)\} \cup \{\forall p : p \in (p_2 \rightarrow p_1)\}$) are equally represented in 2D WAH, and therefore it has a dimension of $2L_S - 1 \times 2L_S - 1$. Figure 2 illustrates a sample CL segment and two sample LoS representations over two distinct 2D WAHs, one for branches and the other for corners. The walking (traversing) direction is clock-wise and note that it does not make any difference on 2D WAHs since both possible localisations are considered.

Two fundamental characteristics of an arbitrary shape for visual perception are its branches (long and smooth sections) and corners where a radical change occurs on the direction of shape boundary. The co-occurrence of corners can be quite minimal when compared to branches and thus, the bins on branch 2D WAH dominate the few from the corners if both are mixed (represented in the same WAH). This is the particular reason for having two distinct 2D WAHs describing branches and corners separately. So all LoS sections, which are encountered within a *sub-segment* (i.e. $\{\forall p : p \in (p_1 + n\Delta \rightarrow p_2 + n\Delta)\}$, $n = 0,1,2,...$) can be described by appropriate 2D WAHs. As shown in the right side of Figure 2, the corresponding bins of branch WAH (for both LoS localization) are simply incremented by one. The process continues with the next LoS section (proceeds Δ pixels towards the walking direction) and keeps moving until a corner is detected by an inbuilt corner detector and represented on corner WAH. Due to space limitations, the algorithmic details of corner detector is skipped in this paper.

In order to test the retrieval efficiency of the proposed descriptor over multimedia databases, we embed the algorithm into a dynamic Feature eXtraction (*FeX*) module to be used for indexing and retrieval processes in MUVIS framework [10]. Any *FeX* module should perform both feature extraction operation over frames, and a similarity distance metric for retrieval. Before starting an indexing process, Canny edge detection and Bilateral filter parameters and three parameters for 2D WAH descriptor, N_S, L_S and Δ, can be set *a priori* depending on the database content. As a mandatory requirement of the *FeX* framework, the descriptors extracted during the indexing process should be (unit) normalized, the details of which are also skipped due to space limitation.

The retrieval process in MUVIS is based on the traditional query by example (QBE) operation. The features of the query item are used for (dis-) similarity measurement among all the features of the visual items in the database. Ranking the database items according to their similarity distances yields the retrieval result. Let $\overline{WAH}^{B,q}_{i,j}$, $\overline{WAH}^{B,x}_{i,j}$ and $\overline{WAH}^{C,q}_{i,j}$, $\overline{WAH}^{C,x}_{i,j}$ be the (normalized) branch and corner WAHs of the query (q) and a particular database item (x). The total (dis-) similarity distance, $D(q,x)$, calculated from its branch, $D^B(q,x)$, and corner, $D^C(q,x)$, components can be expressed as follows:

$$D^B(q,x) = \frac{1}{2}\sum_{i=-L_S+1}^{L_S-1}\sum_{j=-L_S+1}^{L_S-1}\left|\overline{WAH}^{B,q}_{i,j} - \overline{WAH}^{B,x}_{i,j}\right|$$

$$D^C(q,x) = \begin{cases} 1 & if\ \overline{WAH}^{C,q}_{i,j} = 0\ or\ \overline{WAH}^{C,x}_{i,j} = 0,\ \forall i,j \\ \frac{1}{2}\sum_{i=-L_S+1}^{L_S-1}\sum_{j=-L_S+1}^{L_S-1}\left|\overline{WAH}^{B,q}_{i,j} - \overline{WAH}^{B,x}_{i,j}\right| & else \end{cases} \quad (2)$$

$$D(q,x) = \alpha_C D^C(q,x) + (1-\alpha_C) D^B(q,x)$$

where α_C is the weight for corners in the $D(q,x)$ calculation. In such a weighted linear interpolation, we can set $\alpha_C \sim TCW$ for CL segment(s) (e.g. in a binary shape or natural image database where the CL segments for objects are already extracted as in [6]) since the corner information is complete and reliable, otherwise it should be set to a small empirical value (i.e. $\alpha_C < 0.25$) in order to give more weight to branches since some major corners might be missing on the NCL *sub-segment*s present.

4. EXPERIMENTAL RESULTS

First set of experiments are performed to evaluate the accuracy of the proposed corner detector with respect to HVS perceptive criteria (subjective test). Afterwards the retrieval performance of the

proposed *FeX* module via query-by-example (QBE) scheme within a set of image databases is examined. The similarity based retrievals are evaluated using the *ground-truth* methodology In the experiments performed in this section, we used 3 sample (image) databases:

1) **Corel_1K** Image Database: There are 1000 medium resolution (384x256 pixels) images from diverse contents such as wild life, city, buses, horses, mountains, beach, food, African natives, etc.

2) **Corel_10K** Image Database: There are 10000 images from Corel database with similar contents in **Corel_1K**.

3) **Shape** Image Database: There are 1500 black and white (binary) images that mainly represent the shapes of different objects such as animals, cars, accessories, geometric objects, etc.

All experiments are carried out on a Pentium-5 1.8 GHz computer with 1024 MB memory. The following Canny and Bilateral filter parameters are used for all the experiments performed throughout this section: $\sigma = 1.5, thr_{low} = 0.4, thr_{high} = 0.8$ and $\sigma_d = 2, \sigma_r = 20$.

For comparative evaluation, we implemented three edge-based and generic shape descriptors: MPEG-7 Edge Histogram (MPEG7 EHD) [9], Edge Pixel Neighborhood Histogram (EPNH) [4] and Angular Radial Partitioning (ARP) [5], in addition to the proposed method, 2D WAH. Each algorithm is developed as a distinct FeX module under MUVIS framework. The sample databases are indexed using each of these FeX modules and the (sub-) features extracted are used for retrieval. For MPEG7 EHD we set block size as 8 and 201x201 is set as the normalized frame size, 12 and 8 are set for angular and radial partitions of ARP respectively. For 2D WAH, we set $L_S = 10, \Delta = 1$ and $N_S = 10$ for three Corel databases; however $N_S = 1$ is a natural setting since there is only one CL segment extracted per shape. Due to the reasoning given earlier, we set $\alpha_C = 0.5$ only for **Shape** database but $\alpha_C = 0.1$ for **Corel** databases.

In order to measure the retrieval performance, we used an unbiased and limited formulation of *Average Modified Normalized Retrieval Rank* (A*NMRR)* which is defined in MPEG-7 as the retrieval performance criteria. Accordingly, the corresponding *ANMRR* values presented in Table I approves that 2D WAH performs the best shape-based retrieval performance over all the sample databases. In addition to this, the results approve that the proposed method scales much better than the other three as the database size increases and thus it achieves an improved discrimination factor as a generic shape descriptor.

Table I: ***ANMRR*****s of 4 methods for 4 sample databases.**

Databases	Methods			
	2D WAH	MPEG7 EHD	EPNH	ARP
Shape	0.402	0.565	0.46	0.48
Corel_1K	0.252	0.367	0.28	0.742
Corel_10K	0.429	0.71	0.629	0.782

5. CONCLUSIONS

2D WAH method presented in this paper enables shape retrievals over general-purpose multimedia databases. It is based on multi-scale *sub-segment* analysis over (Canny) edge field and therefore, it can conveniently work over arbitrary images, which may encapsulates one or more objects in an inhomogeneous background possibly with strong textural structure. The proposed method basically achieves the following major innovative properties:

- The overall algorithm is unsupervised (no training), with minimal parameter dependency. In our knowledge it is one of the first attempts to index a natural image database according to object shape descriptor in a fully automatic way.
- A multi-scale approach is employed to simplify the image iteratively by removing "details" and texture at each scale and such a pre-processing stage prevents the significant degradations in the following shape analysis as the other methods usually suffer. The cost function used in the relevance model increases the likelihood of obtaining most of the true object boundaries by favoring the longest and the most relevant *sub-segment*s.
- It has a simple, yet efficient corner detector, which improves the description power especially over CL segments. Accordingly, it achieves an improved retrieval performance and most important of all, it can still provide reasonable results for large databases.
- The major advantage of 2D WAH is that as a shape descriptor it deals directly with the arbitrary (natural) images without any unreliable segmentation or object extraction pre-processing stage and hence it achieves generality and robustness.

6. REFERENCES

[1] M. Abdel-Mottaleb, "Image retrieval based onedge representation", in *Proc. 2000 International Conference on Image Processing*, Piscatway, NJ, USA, 2000, vol. 3, pp. 734–737.

[2] M. Bober, "MPEG-7 visual shape descriptors", *IEEE Trans. on Circuits and Systems for Video Technology,* 11 (6), pp. 716-719, Jun 2001.

[3] J. Canny, "A Computational Approach to Edge Detection", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 8, no. 6, pp. 679-698, November 1986.

[4] A. Chalechale and A. Mertins, "An abstract image representation based on edge pixel neighbor-hood information (EPNI)," *Lect. Notes Comput. Science,* 2510, pp. 67-74, 2002.

[5] A. Chalechale, A. Mertins, G. Naghdy ,"Edge Image Description Using Angular Radial Partitioning " *IEE Proc. Vis. Image Signal Proc.,* Vol 151, No. 2, April 2005.

[6] M. Ferreira, S. Kiranyaz and M. Gabbouj, "Multi-scale Edge Detection and Object Extraction for Image Retrieval", ICASSP 2006, Toulouse, France, May 15-19, 2006.

[7] S. Kiranyaz, K. Caglar, O. Guldogan, and E. Karaoglu, "MUVIS: A Multimedia Browsing, Indexing and Retrieval Framework", in *Proc. of Third International Workshop on Content Based Multimedia Indexing*, CBMI 2003, Rennes, France, 22-24 September 2003.

[8] H. Koshimizu, K. Murakami, M. Numada, "Global feature extraction using efficient Hough transform", *Int.l Workshop on Industrial Applications of Machine Intelligence and Vision*, pp. 298-303, Tokyo, Japan, 1989.

[9] B. S. Manjunath, J. R. Ohm, V. V. Vasudevan, A. Yamada, "Color and Texture Descriptors", *IEEE Transactions on Circuits and Systems for Video Technology,* Vol 11, No. 6, pp. 716-719, 2001.

[10] MUVIS. http://muvis.cs.tut.fi

[11] D. Shen, and H. S. Ip, "Discriminative Wavelet Shape Descriptors for Recognition of 2-D patterns", *Pattern Recognition*, 32(2):pp. 151-165, 1998.

[12] C.-H. Teh, and R.T. Chin "On image analysis by the method of moments", *IEEE Trans. Pattern Anal. Mach. Intell.,* 10, (4), pp. 496-513. 1998.

[13] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images*", In Proc. of the Sixth International Conference on Computer Vision,* Bombay, India, January 1998.

[14] M. Wertheimer, "Laws of organization in perceptual forms", Partial translation in W.B. Ellis, editor, "A Sourcebook of Gestalt Psychology", pages 71-88, NY, Harcourt, Brace and Company, 1938.

[15] D. S. Zhang, G. Lu, "A comparative study of Fourier descriptors for shape representation and retrieval", *In Proc. of the Fifth Asian Conference on Computer Vision (ACCV02),* Melbourne, Australia, pp. 646-651, 2002.