Proceedings of the International Multiconference on
Computer Science and Information Technology pp. 129–133
ISBN 978-83-60810-27-9
ISSN 1896-7094

Multiscale Segmentation Based On Mode-Shift Clustering

Wojciech Tarnawski
Chair of Systems and
Computer Networks
Wroclaw University
of Technology, Poland
Email: wojciech.tarnawski@pwr.wroc.pl

Lukasz Miroslaw
Institute of Informatics
Wroclaw University
of Technology, Poland
Email: lukasz.miroslaw@pwr.wroc.pl

Roman Pawlikowski
and Krzysztof Ociepa
Institute of Informatics
Wroclaw University of Technology,
Poland

***Abstract*—We present a novel segmentation technique that effectively segments natural images. The method is designed for the purpose of image retrieval and follows the principle of clustering the regions visible in the image. The concept is based on the multiscale approach where the image undergoes a number of diffusions. The algorithm has been visually compared with a reference segmentation.**

I. Introduction

KNOWLEDGE-BASED society is taking advantage of communication services and IT-tools that manage, store and retrieve the information more and more often. There is a continuing demand to enhance the services that are based on visual information such as movies or images.

The problem of content-based image retrieval in large databases has been a research topic for many years. There exist semi-automatic tools to accomplish this task but still there is no standard method of general applicability. The difficulty lies in understanding what actually the image presents (image understanding) and how the contents can be described (image annotation problem). Additional limitations arise when the goal is to search for similar images in large databases. Since the analysis will require enormous demand for computer resources, the framework allowing for automated analysis will be of great use. In this paper we present a novel segmentation method based on the multiscale approach that was designed especially for image retrieval.

The concept treats the image as a set of disjoint regions that can be described by a set of features such as color, texture or shape:

$$\textbf{Image} \rightarrow \textbf{Image regions} \rightarrow \textbf{Region features} \rightarrow \rightarrow \textbf{Distance between feature vectors} \quad (1)$$

With this assumption an image can be represented in multidimensional feature space as a set of points which number is consistent with the number of significant regions identified during segmentation. Regions characterized by a set of features are located in different locations in the feature space and the position depends on their visual properties. Following the Leibniz's principle called *Identity of indiscernibles* "two things are identical if and only if they share the same and only the same properties" we assumed that similar images share objects/regions of similar properties. Which means that for a subset of images containing objects from the same category, we will observe a set of closely located points, as the regions in that subset will be similar.

Such an approach has the following consequences. Images of little complexity, for example with a few objects and an uniform background will contain only a small number of clusters separated in the feature space [1]. For more complex scenes the number of clusters will be higher and their separation will be probably difficult.

In image understanding it is difficult to find a compromise between interpretation of all image details and the interpretation where certain details could be omitted. Without *a priori* knowledge there is no automated way to determine which details can be disregarded and which objects are large enough to be treated as significant. Therefore, a scale should be considered as a parameter that changes dynamically and generates images with different level of details. With changing scale the degree of precision also changes. Generated images form a so-called multiscale representation.

The concept of multiscale representation was first introduced by Rosenfeld and Thurston [2]. They observed the influence of linear operators of different scale on edge detection. Also, Klinger [3], [4] and Tanimoto [5] used the multiscale approach to describe an image, similarly to Burt and Adelson [6] who proposed a popular, pyramidal representation of an image.

An important aspect in all these attempts is that the images at a larger scale are simplified version of images at smaller scales. Therefore, increasing the scale is equivalent with eliminating the details from images at lower scales. Following such a definition, the scale-space filtering was firstly introduced by Witkin [7] and then further developed by Koenderink [8].

Our concept to image retrieval differs from methods named *Query By Image Content* that tries to extract the information from the whole image [9]. In contrary, the method considers the image as a set of regions represented by a cluster of points in the feature space where the similarity between images is equivalent to a degree of similarity between feature vectors. Therefore, correct segmentation of the regions is a prerequisite

in image retrieval and machine vision tasks where objects play the necessary role. Determination of the criterion for object homogeneity is equally important. Mostly features based on color are used but also other features are often employed, i.e. low-level features such as SIFT, visual descriptors in MPEG-7 standard [10]–[12].

Multiscale approach to segmentation has been already proposed. Wang used a multiscale approach based on high frequency wavelet coefficients and their statistics to perform context-dependent classification of individual blocks of the image. Unlike other edge-based approaches, his algorithm does not rely on the process of connecting object boundaries [13], [14].

The next section describes the segmentation method. The method takes into account the color as the most discriminating feature. Since it is based on concepts such as mean-shift clustering and multiscale approach based on anisotropic diffusion, also these concepts are presented. The last section presents the results and conclusions.

II. Segmentation method

The aim of segmentation is to partition the image into non-overlapping regions that share common features. In case of images of human nature the significant information are derived by color, therefore the features of interest are taken from color model and position. We have decided to take the multiscale approach because such a concept is natural for human perception. When we see the picture, first, we focus on the core objects and analyze regions of strong contrast and different color, next we analyze their details, such as texture. By running a number of diffusions on the image, such concept can be imitated, as with the number of diffusions the details in the image get blurred and only the objects with highest contrast remain. The information on the image is, therefore, simplified.

The segmentation method is depicted in Fig. 1 and can be described as follows. Original image (1) undergoes the multiscale operation and a set of images with different degree of diffusion is generated (2). Next, mean-shift segmentation is produced for each of the image. The results are accumulated in special storage system called accumulator (4). The mode-shift clustering together with a certain metric (6) and a threshold value assigned to it (7) is used in order to label disjoint region of interest and partition them into two layers. The principle layer stores labeled objects that are clearly visible at all the scales and the vague layer contain the regions less distinguishable (8).

A. Anistropic diffusion in multiscale approach

This is the initial step of the algorithm. Here, a number of images are generated in the process of convolving the original image I_0 with the Gaussian kernel with the t variance:

$$I(x, y, t) = I_0(x, y) * G(x, y; t) \tag{2}$$

The variance controls the degree of details visible in the image. Higher values correspond to the image with fewer details that can be clearly distinguishable. The set of derivative

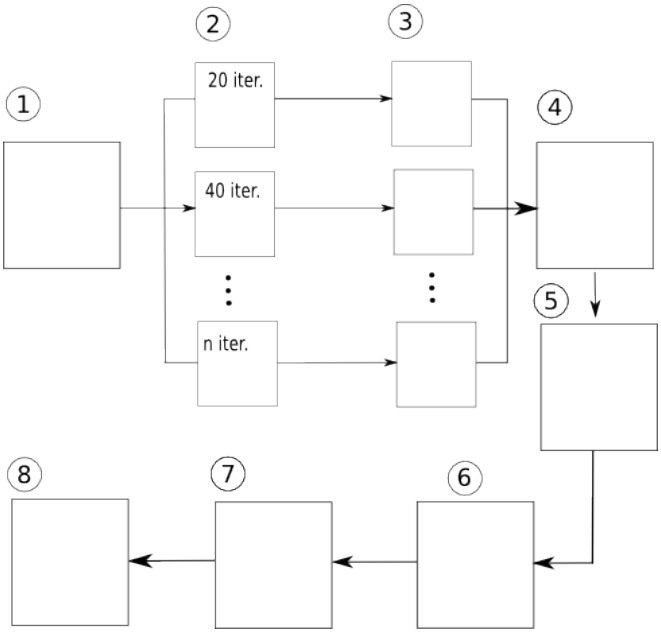

Fig. 1. Segmentation Algorithm. 1. Input image. 2. Multiscale approach, anisotropic diffusion. 3. Meanshift Segmentation. 4.Accumulator. 5. "Modeshift" clustering. 6. Calculation of the metric. 7. Adaptive thresholding. 8. Output image.

images $I(x, y, t)$ is equivalent to concurrent solutions of the heat transport problem or diffusion on the plane [8], namely:

$$I_t = \nabla^2 I = I_{xx} + I_{yy} \tag{3}$$

with initial conditions defined as $I(x, y, 0) = I_0(x, y)$.

Since, the convolution operation smoothes the whole image together with boundaries between objects, we decided to use edge-preserving anisotropic diffusion. The importance of this approach lies in the fact that the diffusion coefficient is not the same for all the pixels. The method is define as follows [15]:

$$I_t = div(c(x, y, t) \cdot \nabla I) = c(x, y, t)\nabla^2 I + \nabla c \cdot I \tag{4}$$

where $c(x, y, t) = g(||\nabla I(x, y, t)||)$ is monotonically decreasing so that within homogeneous regions the diffusion is stronger than in the vicinity of region edges.

In the case of approximation of ∇I with 4-directional neighbourhood we can describe this process as

$$I_{x,y}^{(t+1)} = I_{x,y}^{(t)} + \lambda[D_N \cdot \Delta_N I + D_S \cdot \Delta_S I + D_E \cdot \Delta_E I + D_W \cdot \Delta_W I]_{x,y}^{(t)} \tag{5}$$

where $\lambda = 1/4$, symbols N, S, E, W correspond to directions North, South, West, East, respectively, and Δ is the difference between pixel values in the directions for each iteration t:

$$\Delta_N I \equiv I_{x,y-1} - I_{x,y},\ \Delta_S I \equiv I_{x,y+1} - I_{x,y}, \quad \Delta_W I \equiv I_{x-1,y} - I_{x,y},\ \Delta_E I \equiv I_{x+1,y} - I_{x,y}, \tag{6}$$

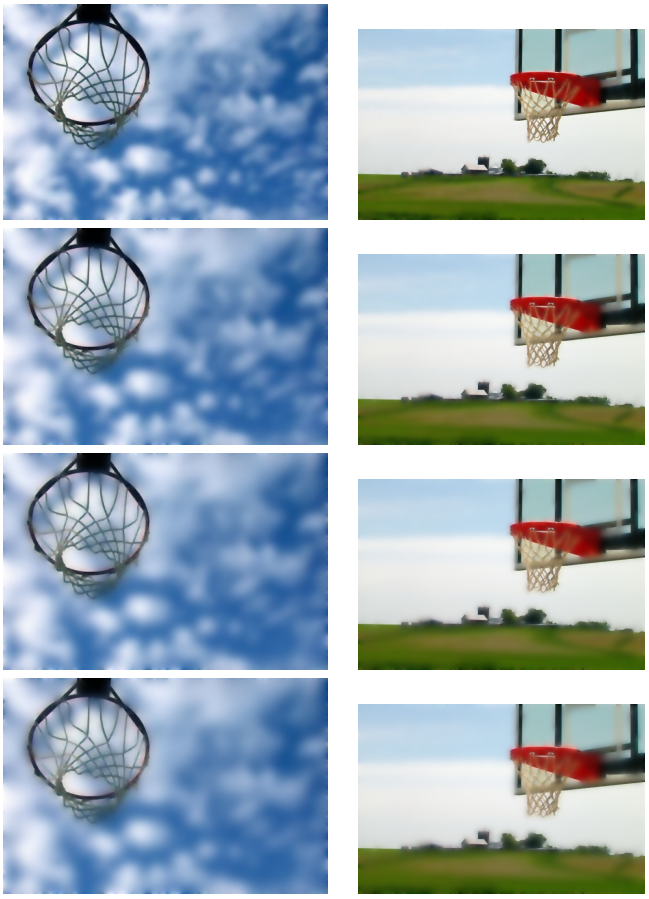

Fig. 2. Filtration results with anisotropic diffusion for $t = 50, 100, 150, 200$ iterations and . $K = 25$.

Similarly, diffusion coefficients are modified for each iteration t as follows:

$$\begin{aligned} D_N &= g\left(||(\nabla I)_{x,y-\frac{1}{2}}||\right) \; D_S \\ &= g\left(||(\nabla I)_{x,y+\frac{1}{2}}||\right) \\ D_W &= g\left(||(\nabla I)_{x-\frac{1}{2},y}||\right) \; D_N \\ &= g\left(||(\nabla I)_{x+\frac{1}{2},y}||\right) \end{aligned} \quad (7)$$

Gradient values $||(\nabla I)||$ were calculated by the Canny filtering [16] that guarantees that the edges are one pixel thick. Similarly to [17] the $g(\cdot)$ was defined as :

$$g\left(||(\nabla I)||\right) = 1 - \exp(-\frac{\tau}{[\Psi(||\nabla I||)/K]^m}). \quad (8)$$

where $m = 4$, $\tau = 3.31488$ a K is a diffusion parameter.

Fig. II-A describes the results of anisotropic diffusion for $K = 25$ and $t = 50, 100, 150, 200$ iterations.

B. Mean Shift Segmentation

During this step, each of the image at different scale goes through so called meanshift segmentation. In contrary to low-level segmentation methods that depend on parameters or apriori knowledge about image contents, clustering methods have ability to be independent on these limitations. The "mean-shift" algorithm is a non-parametric clustering method widely used for segmentation of images of human nature. The method does not require to know the number and the shape of clusters. For each pixel of an image, the set of adjacent pixels is established according to a certain kernel function. For these adjacent pixels the new spatial and color mean values are calculated and used the new center for the next iteration. These steps are repeated until the means do not change. At the end of the iteration, the final mean color will be stored.

The method can be described as follows. Let us define an image as a set of n points $x_i, i = 1, \cdot, n$ in d-dimensional space R^d, and estimator for the density kernel $K(x)$ with radius h defined as

$$f(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \quad (9)$$

For radially symmetric kernels it is sufficient that $K(x)$ follows

$$K(x) = c_{k,d} k(||x||^2) \quad (10)$$

where $c_{k,d}$ is normalization constant for which $\int_x K(x) = 1$. Modal values of density function are located in intercepts of the $\nabla f(x) = 0$. Density gradient is defined as follows:

$$\begin{aligned} \nabla f(x) = \frac{2c_{k,d}}{nh^{d+2}} \sum_{i=1}^{n} (x_i - x) g\left(||\frac{x - x_i}{h}||^2\right) = \\ \frac{2c_{k,d}}{nh^{d+2}} \left[\sum_{i=1}^{n} g\left(||\frac{x - x_i}{h}||^2\right)\right] \left[\frac{\sum_{i=1}^{n} x_i g\left(||\frac{x-x_i}{h}||^2\right)}{\sum_{i=1}^{n} g\left(||\frac{x-x_i}{h}||^2\right)} - x\right]. \end{aligned} \quad (11)$$

where $g(s) = -k'(s)$. The first term is propotional to the estimator of the density kernel of x vector calculated with the the kernel function $G(x) = c_{g,d} g(||x||^2)$ and the second term is defined as

$$\frac{\sum_{i=1}^{n} x_i g\left(||\frac{x-x_i}{h}||^2\right)}{\sum_{i=1}^{n} g\left(||\frac{x-x_i}{h}||^2\right)} - x \quad (12)$$

as is called a "mean shift" vector. This vector is responsible for traversals between the given point to the prototype of a given cluster (attractor). Direction of the vector is oriented to the highest ascend of density. The range of the kernel h is described by two components: a planar one h_s and the feature-driven one h_r that describes the range of features. This procedure is an iterative process made of two steps:

- calculation of the mean-shift vector

$$m_h(x^{(t)}) \quad (13)$$

- translation within the window with the vector

$$x^{(t+1)} = x^{(t)} + m_h(x^{(t)}) \quad (14)$$

and converges to the point where the density gradient is equal to zero. The more detailed description of the method can be found in [18].

C. Definition of Accumulator

Here all the segmentation results at various scale t_q are accumulated and the final segmentation is derived. This procedure has the following steps:

Step 1 Define L-dimensional matrix A which we will call *accumulator*, where L corresponds to the dimension of the feature space that describe the regions. A will store regions defined by a prototype $v_i(t_q)$ composed of features $v_i^{(\cdot)}(t_q)$. In the current stage the features will be defined in the following LAB color model. Obviously, each feature space will determine different prototype distributions in the feature set $v_i(t_q)$. Let us assume, that a single prototype will be described by a $L+2$–dimensional vector:

$$v_i(t_q) = [v_i^{(x)}(t_q), v_i^{(y)}(t_q), v_i^{(1)}(t_q), \ldots, v_i^{(L)}(t_q)]^T \quad (15)$$

where T is a symbol of transposition. In the accumulator space the features $v_i^{(1)}(t_q), \ldots, v_i^{(L)}(t_q)$, will be added without planar features of the prototypes. Matrix A is initialized with zeros.

Step 2 Scan the pixels of $I(x, y, t_q)$ image for each scales t_q and define the feature vector for corresponding prototypes $[v_i^{(1)}(t_q), \ldots, v_i^{(L)}(t_q)]$. Store the mapping:

$$I(x, y, t_q) \to [v_i^{(x)}(t_q), v_i^{(y)}(t_q), v_i^{(1)}(t_q), \ldots, \ldots v_i^{(L)}(t_q)] \to [v_i^{(1)}(t_q), \ldots, v_i^{(L)}(t_q)] \quad (16)$$

Step 3 For each pixel in each scale increment the accumulator:

$$A[v_i^{(1)}(t_q), \ldots, v_i^{(L)}(t_q)] = A[v_i^{(1)}(t_q), \ldots, v_i^{(L)}(t_q)] + 1 \quad (17)$$

D. Mode Shift clustering

This procedure aims at finding significant *modal values* in the accumulator space (4-dimensions: three color values and one corresponding to the accumulator range) where each region is represented by a single point. Mode shift clustering groups points in the vicinity of modal values. The resulting clusters are restrained by spatial and range limit values: σ_S and σ_R. The first parameter defines the searching window of the cubic shape, the latter one compares the accumulator range with local extrema. During this searching procedure, each point in the accumulator moves along the direction of the closest modal value with some finite number of steps.

The points that are grouped form a cluster and define the next mapping:

$$[v_i^{(1)}(t_q), \ldots, v_i^{(L)}(t_q)] \to [V_j^{(1)}, \ldots, V_j^{(L)}];\ j = 1, 2, \ldots, C \quad (18)$$

where C is the number of detected modal values. Note, that the mapping (18) does not have to be defined for all prototypes from the set $v_i(t_q)$.

E. Calculation of the metric

The clusters formed in the previous step are restrained according to a certain metric that defines the *type of distance*. Different metrics (for the sake of simplicity we named them Epsilon) were considered to calculate the distance and number of steps between a given point in the accumulator and the modal value (the cluster prototype) established during "mode-shift" clustering. They are defined as follows:

- Epsilon type 0—Sum of consecutive steps between a given point and the modal value
- Epsilon type 1—Manhattan distance
- Epsilon type 2—Euclidean distance
- Epsilon type 3—Sum of consecutive steps in a given time
- Epsilon type 4—Euclidean distance weighted by the number of steps needed to achieve a cluster prototype
- Epsilon type 5—Sum of consecutive steps in a given time weighted by the number of steps needed to achieve a cluster prototype

F. Adaptive threshold

With each metric a certain threshold value is associated. This value controls whether the point belongs to the *principle layer* or to the *vague layer*. The regions of higher contrast that survived diffusions and have higher values in the accumulator constitute the principle layer. Objects with lower contrast have smaller values in accumulator and, therefore, belong to the vague layer.

During this step the labeling of the pixels in the image with the mappings: (18) and (16) is performed according to the mentioned threshold. Pixels with undefined mapping (16) are labeled as -1.

The output image contains disjoint regions and the algorithm is terminated. The image will contain two subsets with pixels:

- with a label equal to -1, which form the vague layer.
- with a label different than -1, which will create the principle layer.

III. Results

The algorithm has been tested for different parameter settings on Segmented and Annotated Benchmark set [19]. Especially metric Epsilon has been thoroughly tested and the segmentation results were compared visually with a reference segmentation done by the expert. The analysis of results indicated that the best results were achieved for Epsilon type 4 and 5. Example images 3 were analyzed with two different threshold values and presented in consecutive rows in Figs. 4–5. As one can see, the results are comparable to each other, alas, more work is needed to perform a detailed analysis of parameter influence.

Nevertheless, it may be easily noticed that the segmented regions correspond to the reference segmentation with a satisfactory accuracy, i.e. the significant objects on principle layer are detected (the right column) and their periphery are close to the boundaries defined by the expert. Such region

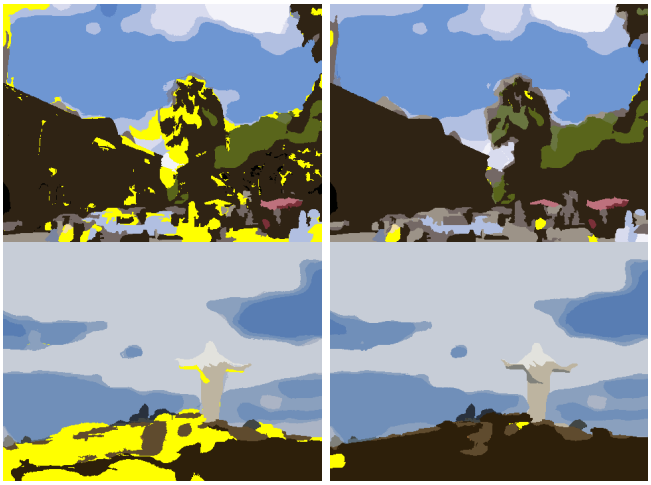

Fig. 5. Example results for epsilon type 5 and different threshold values. The vague layer is indicated by yellow pixels.

Fig. 3. Original images.

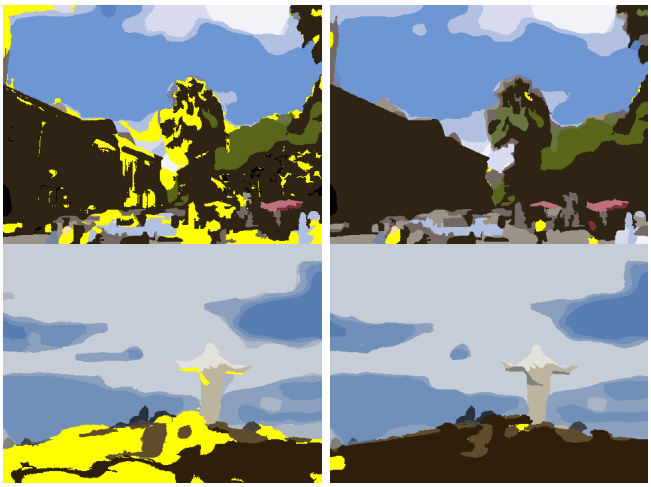

Fig. 4. Example results for epsilon type 4 and different threshold values. The vague layer is defined by yellow pixels.

representation as a feature set are a requirement for image retrieval based on clustering principle.

IV. Conclusions

The paper presents the segmentation method based on multiscale approach and mean-shift clustering. A novel algorithm partitions the image into disjoint sections that form the layers with significant (principal layer) and insubstantial regions (vague layer). Detected regions The results were assessed visually for a number of image classes method. Although, the results were satisfactory for a number of image classes and the method can serve as a potential tool in image retrieval task, more work is needed to fully customise the method.

Acknowledgment

The financial support of the Polish-Singapur Research Fund (project No. 65/N- SINGAPORE/2007/0) has made this research possible and is kindly acknowledged.

References

[1] W. Pedrycz, A. Amato, V. Di Lecce, and V. Piuri, "Fuzzy clustering with partial supervision in organization and classification of digital images," *Fuzzy Systems, IEEE Transactions on*, vol. 16, no. 4, pp. 1008–1026, Aug. 2008.

[2] A. Rosenfeld and M. Thurston, "Edge and curve detection for visual scene analysis," *IEEE Trans. Comput.*, vol. 20, no. 5, pp. 562–569, 1971.

[3] M. Klinger, "Pattern and search statistic," *Optimizing Methods in Statistics*, 1971.

[4] L. Uhr, "Layered "recognition cone" networks that preprocess, classify, and describe," *Computers, IEEE Transactions on*, vol. C-21, no. 7, pp. 758–768, July 1972.

[5] S. Tanimoto and T. Pavlidis, "A hierarchical data structure for picture processing," *Computer Graphics and Image Processing*, vol. 4, no. 2, pp. 104 – 119, 1975.

[6] P. J. Burt and E. H. Adelson, "The laplacian pyramid as a compact image code," *IEEE Transactions on Communications*, vol. 31,4, pp. 532–540, 1983.

[7] A. Witkin, "Scale-space filtering: A new approach to multi-scale description," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '84.*, vol. 9, Mar 1984, pp. 150–153.

[8] J. J. Koenderink, "The structure of images," *Biological Cybernetics*, vol. 50, pp. 363–370, 1984.

[9] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker, "Query by image and video content: the qbic system," *Computer*, vol. 28, no. 9, pp. 23–32, Sep 1995.

[10] M. Bober, "Mpeg-7 visual shape descriptors," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 11, no. 6, pp. 716–719, Jun 2001.

[11] M. Bober and P. Brasnett, "Mpeg-7 visual signature tools," in *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*, 28 2009-July 3 2009, pp. 1540–1543.

[12] M. Bober, W. Price, and J. Atkinson, "The contour shape descriptor for mpeg-7 and its applications," in *International Conference on Consumer Electronics, Digest of Technical Papers.*, 2000, pp. 286–287.

[13] Z. Chi and H. Yan, "Image segmentation using fuzzy rules derived from k-means clusters," *Journal of Electronic Imaging*, vol. 4, no. 2, pp. 199–206, 1995.

[14] J. Z. Wang, J. Li, R. M. Gray, and G. Wiederhold, "Unsupervised multiresolution segmentation for images with low depth of field," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, pp. 85–90, 1999.

[15] P. Perona and J. Malik, "Scale-space and edge detection using anisotropic diffusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 7, pp. 629–639, 1990.

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

[17] J. Weickert, *Anisotropic Diffusion in Image Processing*, ser. ECMI Series. Stuttgart, Germany: Teubner-Verlag, 1996.

[18] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 603–619, 2002.

[19] H. J. Escalante, C. Hernández, J. Gonzalez, A. López, M. Montes, E. Morales, E. L. Sucar, and M. Grubinger, "The segmented and annotated IAPR TC-12 benchmark. computer vision and image understanding," *Computer Vision and Image Understanding*, vol. 114, no. 4, pp. 419–428, 2009.