Affine Invariant Representation and Classification of Object Contours for Image and Video Retrieval

YANNIS AVRITHIS, YIANNIS XIROUHAKIS and STEFANOS KOLLIAS
Department of Electrical and Computer Engineering
National Technical University of Athens
9 Heroon Polytechneiou str., 157 73 Zographou
GREECE
email: {iavr,jxiro}@image.ntua.gr

Abstract: - Recent literature comprises a large number of papers on the query and retrieval of visual information based on its content. At the same time, a number of prototype systems have been implemented enabling searching through on-line image databases and still image retrieval. However, it has been often pointed out that meaningful/semantic information should be extracted from visual information in order to improve the efficiency and functionality of a content-based retrieval tool. In this context, present work focuses on the extraction of objects from images and video clips and modeling of the resulting object contours using B-splines. Affine-invariant curve representation is obtained through Normalized Fourier descriptors (NFD), curve moments, as well as a novel curve normalization algorithm that leads to major preservation of object shape information. A neural network approach is employed for supervised classification of video objects into prototype object classes. Experiments on several real-life and simulated video sequences are included to evaluate the classification results for all affine-invariant representations used.

Key-Words: - affine-invariant representation, object contours, image and video retrieval and classification

1 Introduction

Due to the recent growth in interest in multimedia applications, an increasing demand has emerged for efficient storage, management and browsing in multimedia databases. The latter has been given considerable attention after the guidelines of the Moving Pictures Expert Group regarding the MPEG-4 and MPEG-7 standards. Content-based query, retrieval and indexing capabilities are of major importance in browsing digital video databases, due to the amount of information involved.

Recent literature comprises a number of prototype systems providing content-based image query and retrieval capabilities, including for example VIRAGE, QBIC, Photobook, VisualSEEk, Excalibur, CIIR and C-BIRD. A portion of these systems have already been implemented and are now in the stage of evaluation. Content information is modeled in terms of dominant colors, texture, color and texture composition, as well as shape attributes. Moreover, several works have been proposed in recent literature for the extension of the aforementioned schemes to video databases. These include video object modeling and segmentation [7] and optimal extraction of frames and shots [1]. Prototype systems have also been proposed, giving the ability of querying-by-sketch in image databases using image curvelet feature extraction and matching [5]. However, based on the functionalities of the implemented systems and the evaluation results from non-expert users, it has been often pointed out that existing systems lack the ability to extract and retrieve semantic information.

To this end, meaningful object extraction and modeling has become an emerging task in the field of content-based image and video retrieval. This task generally focuses on the extraction and representation of object shape, including object segmentation and object contour representation and modeling. In this context, it is critical that the obtained representation allows for efficient modeling and classification of objects into abstract categories. For this purpose, a number of approaches have appeared in literature including Fourier descriptors, chain codes, polygonal approximation, curve moments and B-splines among others.

In this work, a prototype system is introduced extending the use of B-spline object contour representation to video queries and allowing video object matching and classification based on object shape apart from other features (such as color, texture, motion etc.). The proposed system

emphasizes on the extraction of the object shape attribute in order to obtain image and video characterization on the basis of semantic information. In this sense, the object representation obtained in this paper is further generalized, in order to achieve video content description within a higher level of abstraction.

The proposed system consists of several modules. Each video sequence of the video database is partitioned into video shots and an optimization method is used for selecting a small number of key frames and shots. Video objects are then extracted with an unsupervised color and motion segmentation technique and a B-spline representation is used to model the curves of the derived object contours. The problem of an affine-invariant description of the curves is tackled in terms of B-spline knot-points, normalized Fourier descriptors and curve moments. A novel affine-invariant curve normalization (AICN) approach is also presented that provides an affine-invariant description of object curves while at the same time preserving all information on curve shapes. A feedforward neural network is utilized for classification of the derived representation of object contours. The resulting implementation leads to significant classification improvement compared to an earlier work [6], as well as much lower computational cost.

2 Video Object Extraction

Initially, each video sequence of the video database is partitioned into video shots, each of which corresponds to a continuous action of a single camera operation. An unsupervised color and motion segmentation technique is then applied to all frames of each video shot, and segment characteristics, such as average color/motion, location and size, which are used to construct a feature vector for each frame. Shot feature vectors are constructed, characterizing whole shots, and a set of representative shots (key shots) is extracted by means of shot clustering using the generalized Lloyd or *K*-means algorithm. Key-frames are then selected from key shots, based on an optimization method for locating a set of minimally correlated feature vectors. Optimization is performed by a genetic algorithm approach. After application of the above procedure, which is described in detail in [1], the problem of content-based retrieval from a video database is actually reduced to still image retrieval.

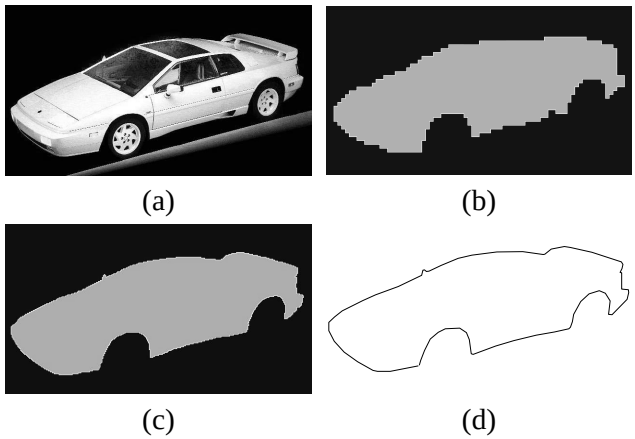


Fig. 1: Extraction of object contours through segmentation: (a) initial image, (b) first stage of segmentation, (c) final segmentation result, (d) object contour.

Video object extraction is accomplished by applying color segmentation on key frames, selecting the largest segments and obtaining the respective segment contours. The Recursive Shortest Spanning Tree (RSST) algorithm is the basis for color segmentation, as it is considered as one of the most powerful tools for image segmentation, compared to other techniques (e.g. color clustering, pyramidal region growing and morphological watershed). In order to reduce the computational complexity involved, a multiresolution RSST (M-RSST) [1] approach is used, which recursively applies the RSST algorithm on images of increasing resolution, yielding much faster execution. The results are depicted in Figure 1 for a target number of segments equal to 2 and for an initial resolution level 3 (equivalent to 8x8 blocks). It can be seen that small segments are in effect eliminated, since they are not visible at the initial (lowest) resolution. Such filtering, according to object size, is desirable since it achieves a high level video content representation. Also, since only the segment contour shapes are affected in each iteration, it is possible to acquire the exact contour shapes at the highest resolution level even without knowledge of the image at that level.

3 B-Spline Representation

Using the object contours obtained through segmentation of the key-frames, a curve modeling scheme should be applied in order to facilitate recognizing and matching object shapes. In this work B-splines are employed since they possess a number of properties which make them suitable for shape representation and analysis such as smoothness and continuity, built-in boundedness, local controlability and shape invariance under affine transformation.

3.1 Curve Modeling

Assume that we are given a dense set of m data curve points $\mathbf{s}_j$, j = 0,1,..,m-1. The initial goal is to model the input curve using closed cubic B-splines that consist of n+1 connected curve segments $\mathbf{r}_i$, i = 0,1,..,n. Each of these segments is a linear combination of four cubic polynomials in the parameter $t \in [0,1]$:

$$\mathbf{r}_i(t) = \mathbf{C}_{i-1}Q_0(t) + \mathbf{C}_i Q_1(t) + \mathbf{C}_{i+1}Q_2(t) + \mathbf{C}_{i+2}Q_3(t) \quad (1)$$

for i = 0,1,..,n, where the basis functions are $Q_k(t) = a_{k0}t^3 + a_{k1}t^2 + a_{k2}t + a_{k3}$, $k = 0,1,2,3$.

Using the continuity constraints in position, slope and curvature on the connection points between segments and the invariance property to coordinate transformations ($\sum_{k=0}^{3} Q_k(t) = 1,\ t \in [0,1]$), the polynomial factors a_k are computed and thus the basis functions Q_k (t) are defined. The B-spline used to model the input curve is given using the curve segments as:

$$\mathbf{r}(t') = \sum_{k=0}^{n} \mathbf{r}_i(t'-i) = \sum_{k=0}^{n} \mathbf{C}_{i \bmod(n+1)} N_i(t') \quad (2)$$

where $0 \le t' \le n-2$ and $N_i(t)$ denote the so-called blending functions:

$$N_i(t') = \begin{cases} Q_3(t'-i+3) & i-3 \le t' < i-2 \\ Q_2(t'-i+2) & i-2 \le t' < i-1 \\ Q_1(t'-i+1) & i-1 \le t' < i \\ Q_0(t'-i) & i \le t' < i+1 \\ 0 & \textit{otherwise} \end{cases} \quad (3)$$

In order to find the appropriate B-spline, the control points $\mathbf{C}_i$ must be determined. The approach followed in this work tries to find an approximate B-spline such that the error between the observed data and their corresponding B-spline curve is minimized. In this sense, the metric $d^2 = \sum_{j=1}^{m} \left\| \mathbf{s}_j - \mathbf{r}(t'_j) \right\|^2$ should be minimized. If appropriate parametric values of t' are allocated on the curve, then the MMSE solution for the control points is given in matrix form as $\mathbf{C}_f = (\mathbf{P}^T\mathbf{P})^{-1}\mathbf{P}^T\mathbf{f}$, where $\mathbf{f}$ and $\mathbf{C}_f$ are of size m×2 and (n+1)×2 respectively containing the given data points $\mathbf{s}_j$ and the control points $\mathbf{C}_i$ respectively. The m×(n+1) matrix $\mathbf{P}$ contains appropriate values for the blending functions, estimated on the points $\mathbf{r}(t'_j)$.

For the allocation of parametric values of t', the chord length (CL) method is employed. Specifically, for t'_1 = 0 and $t'_{\max}$ = n-2, t'_j associated with the sample point $\mathbf{s}_j$ is estimated by:

$$t'_j = t'_{j-1} + t'_{\max} \cdot \left\| \mathbf{s}_j - \mathbf{s}_{j-1} \right\| \cdot \left(\sum_{l=2}^{m} \left\| \mathbf{s}_l - \mathbf{s}_{l-1} \right\| \right)^{-1}, \quad j = 2,...m \quad (4)$$

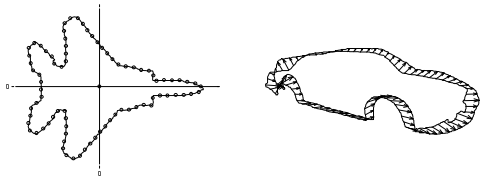
The CL is based on the fact that the chord length between any two points is a very close approximation to the arc length of the curve and under the assumption of constant speed of a particle onto the curve. The CL method is robust to uniformly distributed noise, but suffers from nonuniform noise and nonuniform sampling. Alternatively, the inverse chord length method (ICL) could be used for robust results, as reported in [3].

3.2 Curve Matching

In the sequel, the problem of comparing and matching curves using their B-spline representation is addressed. Assume that a set of M different curves, i.e. M sets of samples, are available in the database. After having modeled these sets of points with M cubic B-splines, it can be seen that their control points cannot decide shape similarity between these curves, since generally different sets of control points may describe the same curve.

For this reason, it is comfortable to derive for each curve the so-called knot points $\mathbf{p}_i$, i=0,1,...,n, using the estimated control points. For cubic B-splines, this is achieved as $\mathbf{p}_f = \mathbf{A}\mathbf{C}_f$, where $\mathbf{p}_f$ is the (n+1) ×2 matrix containing the knot-points and $\mathbf{A}$ is the (n+1)×(n+1) circulant matrix with [2/3, 1/6, 0, ..., 0, 1/6] as its first row. It must be pointed out here that the knot-points belong to the derived B-spline.

However, it can be seen that for any two curves, it is not certain that their estimated knot-points correspond, even if they are equal in number. For this reason, they must be re-allocated on each curve [2]. The first knot-point is placed on the curve point where the curve intersects the x-axis. In the sequel, we place l knot-points equally spaced w.r.t. t' onto each curve.



(a) (b)

Fig. 2: (a) Reallocated knot-points, (b) Knot-point matching between two distinct car curves.

Figure 2 illustrates how reallocated knot-points accomplish correspondence between starting points of different curves.

4 Affine-Invariant Description

In the sequel, two problems arise: (a) the comparison and classification of curves must be invariant to possible affine transformations and (b) we should indicate a way of rapid initial classification since it is impossible to compare a sample curve with all curves existing in the database. Affine-invariant comparison is addressed in literature using curve moments and Fourier descriptors. It can be seen that the former approach is computationally costly but is reported to be relatively accurate, whereas the latter reduces computational cost, however seems not to be a generic description for 2D curves. For this reason, a novel curve normalization approach is proposed that provides an affine-invariant description of object curves at a low computational cost, while at the same time preserving all information on curve shapes.

4.1 Fourier Descriptors and Curve Moments

As we mentioned above, a set of m sample points were used to describe the contour of an object. For each sample $\mathbf{s}_k$, k=0,...,m-1, the sequence $\mathbf{b}_k=\mathbf{s}_{xk}+ j\mathbf{s}_{yk}$ is obtained, where $\mathbf{s}_{xk}$, $\mathbf{s}_{yk}$ denote the x, y coordinates of $\mathbf{s}_k$ respectively. The discrete Fourier factors for this sequence are obtained by

$$F_i = \sum_{k=0}^{m-1} \mathbf{b}_k \cdot \exp\left(-\frac{j2\pi \cdot i \cdot k}{m}\right), \quad i = 0,1,\ldots,m-1 \quad (5)$$

If $\mathbf{b}'_k$ is a sequence obtained from $\mathbf{b}_k$ by scaling, translation, rotation and shift, then the discrete Fourier factors are given by

$$F'_i = a \cdot F_i \cdot \exp\left(j\frac{\vartheta - 2\pi \cdot i \cdot k_0}{m}\right) + \mathbf{b}_0 \cdot \delta(0) \quad (6)$$

and the normalized Fourier descriptors (NFD) $\mathbf{v}_i = |F'_i|/|F'_1|$, i=2,3,...,m-1, are invariant to translation, rotation and starting point.

As it will be seen in the sequel, the NFD are fed into a neural network (NN). In order to keep the inputs of the NN reasonably small, we choose to use only knot-points instead of all sample points. Thus $\mathbf{v}$ is an $l\times1$ vector.

Although the NFD possess the aforementioned desirable properties, they seem to be a poor description for the contour curve of an object. For this reason, classification based on NFD is not always reliable, as shown in the experiments, and a finer match might be necessary, accomplished through curve moments [3]. In this case, each spline is parametrized in terms of its arc lengths s as $\mathbf{R}(s)=[x(s), y(s)]$ which is a known function of its control points. The (p,q) order moments are weighted by kernels w_j, so that

$$m(p,q)^{(j)} = \int_{s=0}^{S} x^p(s) \cdot y^q(s) \cdot w_j(x,y)\, ds \quad (7)$$

By appropriate choise of the kernels, it can be seen that the affine parameters $\mathbf{L}$, $\mathbf{c}$ aligning two curves, i.e. $\mathbf{r}(t')^{(a)} = \mathbf{L} \cdot \mathbf{r}(t')^{(b)} + \mathbf{c}$, can be estimated from their moments up to order two [3].

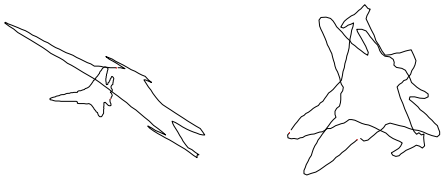
4.2 Curve Normalization Approach

As mentioned above, NFD possess the affine-invarianve property that makes them suitable for the representation of object contours. At the same time, they are preferable to curve moments, since they result to a quantitative curve description facilitating curve classification. However, rejecting phase information, through the use of the NFD, leads to a relatively poor representation of object contours. For that purpose, a novel affine-invariant curve normalization (AICN) scheme is proposed.

The basic idea of the algorithm relies on removal of all translation, scaling, rotation and starting point transformations of an object contour without discarding phase information. This is achieved applying a series of linear (affine) transformations to the obtained set of data of the B-spline curve, with transformation parameters directly estimated from first and second order statistics of curve data. Considering again a set of samples $\mathbf{s}_k$, k=0,...,m-1 on a B-spline, translation transformation is removed first by normalizing their center of gravity to the axes origin, so that $\sum_{k=0}^{m-1} \mathbf{s}_{xk} = 0$ and $\sum_{k=0}^{m-1} \mathbf{s}_{yk} = 0$. We have preserved the same symbol $\mathbf{s}_k$ for the resulting set of samples for simplicity. The removal of scale transformation is achieved by performing two successive normalization steps of the sample set (in the directions of 0^o and 45^o), so that for the resulting set $\sum_{k=0}^{m-1} \mathbf{s}_{xk}^2 = 1$, $\sum_{k=0}^{m-1} \mathbf{s}_{yk}^2 = 1$, and $\sum_{k=0}^{m-1} \mathbf{s}_{xk}\mathbf{s}_{yk} = 0$.

These properties of the transformed sample set, lead to the immediate removal of rotation and starting point transformations. Starting point transformation is removed by calculating a default starting point so that the phases of two specific elements of the Fourier transform of the curve (namely, F_1 and F_{m-1}) become symmetric. Finally, rotation is normalized in terms of a characteristic direction, such us maximum radius or mean angle direction, so that, for instance, the mean angle of curve points is normalized to zero.

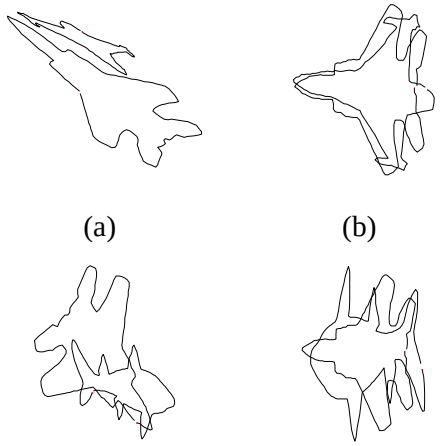
In Figure 3 the initial and transformed sets are depicted for the contours of two airplanes, having performed the translation and scale normalization steps. It can be seen that the resulting sample curves can be matched after a simple 2D rotation of the curves onto the image plane.



(a) (b)

Fig. 3: (a) Extracted object contours (sample curves), (b) Transformed sample curves after translation and scale normalization.

The resulting sample sets after all normalization steps are depicted in Figure 4. It can be seen that objects belonging to different object classes result to essentially different transformed curves. Even in the case of relatively similar objects (such as airplanes and fish), classification is remarkably improved.



(a) (b)

(c) (d)

Fig. 4: (a,c) Extracted object contours (sample curves), (b,d) Final transformed sample curves.

It should be noted that the representation obtained with the proposed normalization approach is effectively invariant to affine transformations (including starting point) without discarding phase information on the original curve shape, leading to reliable and accurate curve comparisons. Moreover, it is computationally effective and it is not based on matching, so that it can be directly used as input to any classification mechanism.

5 Neural Network Classification

Along the lines of the previous section, it is possible for a given set of curve prototypes to determine which one matches best a given curve indepedently of affine transformations. At first, using groups of curve prototypes, we define primary object classes (e.g., airplanes, cars, vases etc.), which can be further organized in an object class database. Hence, the problem of classifying a sample curve to a specific class reduces into locating the best match between this sample curve and the set of all prototypes. Note, however, that a very large amount of curve prototypes would be used in a practical system, making the procedure of direct comparison with all available prototypes extremely time consuming.

For this purpose, a neural network approach is used in order to constrain the search procedure into a small subset of object classes. In particular, the representation of curve prototypes (NFD or normalized curves) is used as an input to a feedforward NN, and a network output is assigned to each primary object class. The network attempts to implement a mapping between an input pattern $\mathbf{v}=[v_1,v_2,\ldots,v_N]^T$ and a desired output pattern $\mathbf{d}=[d_1,d_2,\ldots,d_C]^T$. A neural network with two hidden layers is used, as shown in Figure 5. Neurons of successive layers are interconnected through weights, so that each neuron input is a weighted sum of the previous layer neuron-values, transformed by the sigmoid activation function [4]. In the training stage, the *B*-spline representation $\mathbf{v}^{(p)}$, $p=1,\ldots,M$ of a set of M curve prototypes is fed as input to the NN, while the desired output $\mathbf{d}^{(p)}$, $p=1,\ldots,M$ is determined by setting the component of $\mathbf{d}^{(p)}$ that corresponds to the curve prototype class equal to one and all the other components to zero. The Levenberg-Marquardt method is used for training, attempting to minimize the sum-squared error between the desired and actual output patterns. The minimization is performed by updating the weights connecting neurons of successive layers and re-evaluating the outputs and the sum-squared error in an iterative way.

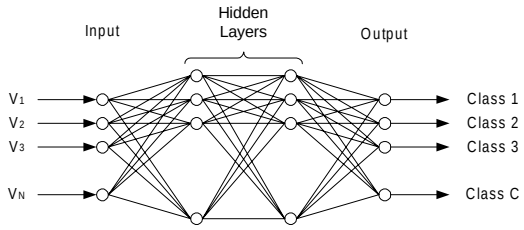


Fig. 5: Neural network architecture for object classification.

In the allocation stage, the *B*-spline representation $\mathbf{v}=[v_1,v_2,\ldots,v_N]^T$ of a test curve is used as input to the NN. Since one network output corresponds to each object class, representing the classification result of the input curve into the respective class, the input curve is typically

classified to the object class that corresponds to the maximum network output.

6 Experimental Results

The aforementioned methodology for image classification and retrieval has been tested using an MPEG video database containing video sequences of total duration 4 hours. Each sequence is partitioned into video shots and key frames and shots are extracted. Object contours are obtained through segmentation, and reallocated knot-points are derived for each curve. The Fourier descriptors as well as the proposed normalized representation of the reallocated knot-points are used as input to the neural network classifier.

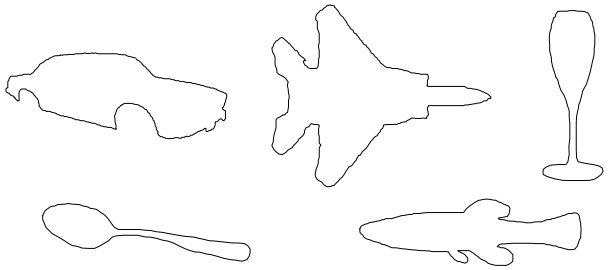


Fig.4: Sample prototype curves corresponding to distinct object classes and used for training.

Object Class	Classification Results	
	NFD	**AICN**
Cars	89.2%	98.6%
Airplanes	83.6%	99.2%
Glasses	76.1%	94.9%
Spoons	90.4%	96.3%
Fish	84.7%	97.6%
Total	**84.8%**	**97.3%**

Table 1. Classification results.

Five object classes are defined for the experiments in this paper, corresponding to cars, airplanes, glasses, spoons and fish. A sample prototype curve for each object class is illustrated in Figure 4, while 50 curves per class are approximately used for training and 50 different curves per class for classification testing. After NN training with the training set consisting of 250 curves, classification is tested using the 250 (approximately) curves of the test set. The classification results for both normalized Fourier decriptors (NFD) and affine-invariant curve normalization (AICN) are given in Table 1. Significant classification improvement is achieved with the use of AICN, which is expected since all information on curve shapes is preserved in this case. Better results could be obtained in the case of NFD by using the neural network as a pre-classification step and then employing curve matching, by means of curve moments and metric distances, for the final classification [6]. This technique, however, leads to significant increase in computational cost.

7 Conclusions – Future Work

A system for content-based image retrieval from image/video databases based on object contours has been presented in this paper, using B-splines for affine invariant contour representation, and a neural network for supervised classification of objects into prototype object classes. Normalized Fourier descriptors along with a curve normalization approach have been employed. Higher level classes can be defined by combining primary classes, providing the ability to obtain a high level of abstraction in the representation of each video sequence. This prospect is currently under investigation.

References:

[1] Y. Avrithis, A. Doulamis, N. Doulamis and S. Kollias, "A Stochastic Framework for Optimal Key Frame Extraction from MPEG Video Databases," *Computer Vision and Image Understanding*, May 1999.

[2] F.S.Cohen, Z.Huang, and Z.Yang, "Invariant Matching and Identification of Curves using B-Splines Curve Representation," *IEEE Trans. Image Proc.*, Vol.4, No.1. Jan. 95.

[3] Z.Huang and F.S.Cohen, "Affine-invariant B-Spline Moments for Curve Matching," *IEEE Trans. Image Processing*, Vol.5, No.10. Oct 96.

[4] D. R. Hush and B. G. Horne, "Progress in Supervised Neural Networks," *IEEE Signal Processing Magazine*, Jan 1993.

[5] Z.Lei, Y.Chan, and D.Lopresti, "Image Curvelet Feature Extraction and Matching," *Proc. ICIP*, Santa Barbara, USA, Oct. 97.

[6] Yiannis Xirouhakis, Yannis Avrithis and Stefanos Kollias, "Image Retrieval and Classification Using Affine Invariant B-Spline Representation and Neural Networks," *Proc. IEE Colloquium Neural Nets and Multimedia*, London, UK, Oct. 1998.

[7] D.Zhong and S.-F.Chang, "Video Object Model and Segmentation for Content Based Video Indexing," *Int. Conf. Circuits and Systems*, Hong Kong, June 97.