Video Scene Classification and Segmentation Based on Support Vector Machine

Yingying Zhu, Zhong Ming, Jun Zhang

Abstract—**Video scene classification and segmentation are fundamental steps for multimedia retrieval, indexing and browsing. In this paper, a robust scene classification and segmentation approach based on Support Vector Machine (SVM) is presented, which extracts both audio and visual features and analyzes their inter-relations to identify and classify video scenes. Our system works on content from a diverse range of genres by allowing sets of features to be combined and compared automatically without the use of thresholds. With the temporal behaviors of different scene classes, SVM classifier can effectively classify presegmented video clips into one of the predefined scene classes. After identifying scene classes, the scene change boundary can be easily detected. The experimental results show that the proposed system not only improves precision and recall, but also performs better than the other classification systems using the decision tree (DT), K Nearest Neighbor (K-NN) and Neural Network (NN).**

I. INTRODUCTION

In recent years, the demand on video data such as digital library, video on demand, long-distance education, web, mobile and etc, has increased dramatically. This requires the efficient indexing and retrieval of multimedia material. Segmentation and classification are fundamental steps for efficient accessing, retrieving, and browsing large amount of video data. Thus, it is essential to parse video structure into semantic units.

There are usually two layers of construction units in video documents: shots and scenes (also often referred as story units). In general, a shot is defined as the sequence of consecutive frames from the start to the end of recording in a camera. Traditionally, segmentation depends on the automatic detection of camera changes (hard cut) and editing effects (such as fading and dissolving). Shot separation often leads to a far too fine segmentation of a video sequence. The huge amount of digital video needs segmentation at higher levels, e.g., scene level, for efficient access. A scene is defined as one or more successive shots organized by certain semantic rules. Scene characterization should be content- and. search-dependent. And it has different scopes, i.e., scenes can be different classes such as news, music, basketball, soccer, cartoon, etc, or different advertisements within a commercial. Scene classification and segmentation is a challenging research topic and many algorithms have been presented. However, the majority of these algorithms rely on visual information only, neglecting the rich supplementary source of the accompanying audio signal and text. Thus, it is observed that to decompose video sequences into different semantic scenes, algorithms dependent on visual information alone cannot achieve satisfactory results. Audio track and text information in video data can provide very useful and complementary semantics cues to improve detection of scene transitions. Recently, some research groups have developed algorithms to detect scene changes and classify scenes by incorporating audio and visual information. At the high level, Li et al.[1] analyzed both audio and visual sources to extract high-level semantic cues. Meaningful movie events were extracted and assigned semantic labels, namely two-speaker dialogs, multiple-speaker dialogs, and hybrid events. Xu et al.[2] used PCA to reduce the dimensionality of the low-level audio and visual features, and they used Gaussian Mixture Model (GMM) based classifier models. Goela et al.[3] considered video scene as a two-classes clustering problem ("scene change and no scene change) and use SVM to classify frame. At the middle level, domain videos such as sports can be classified into different sub-classes. In [4], six main logical units for TV broadcast news were distinguished, namely, begin, end, anchor, interview, report, and weather forecast. For each frame of the video, a feature vector was generated with motion and audio features. In another approach [5], Xavier et al. classify sports into four sub categories (basketball, ice hockey, football and soccer) by using motion and color features and HMM based classifier models. At the finer level, a video sequence itself can be segmented, and each segment can then be classified according to its semantic content. Nam et al.[6] integrated cues from the visual and auditory modality and labeled violent scenes in TV drama. In [7], sports video sequences are first segmented into shots, and each shot is then classified into player, audience, player field, studio, and graphics shot categories.

Despite many initial successes, multimodal integration in scene classification and segmentation still has its problems. Just like many other data fusion problems, scene classification and segmentation depend to large extent on the weights assigned to each feature under consideration, especially when the two complementary sources of information fail to detect transition simultaneously and

This work was supported by National Natural Science Foundation of China (60673122), Guangdong Natural Science Foundation (5301029) and Shenzhen University Science Projects Foundation （200515）.

Yingying Zhu and Zhong Ming are with College of Information Engineering, Shenzhen University, Shenzhen, 518060, and Yingying Zhu is a postdoctor of Software Engineering Ltd. of Harbin Institute of Technology, Haerbin, P.R.China.

Jun Zhang is with Department of Computer Science, SUN Yet-sen University, Guangzhou, P.R.China.

Yingying Zhu is corresponding author, e-mail: zyy211@gmail.com.

978-1-4244-1821-3/08/$25.00©2008 IEEE

identify the same classes. In this paper, we present a scene classification and segmentation scheme using multimodal integration which based on SVM. The clip-based SVM classifier is trained to recognize these scene classes. According to the given training and testing data set, video scenes are classified eight classes: news, commercial, weather forecast, cartoon, MTV, tennis game, basketball game, and football game. Finally, a sophisticated comparison was conducted among the SVM-based, the DT (decision tree), K-NN (K nearest neighbor) and NN (neural network) classifiers. The results show that the proposed SVM-based scheme has outperformed the other three.

The remainder of the paper is organized as follows. In Section II, we describe the framework of scene classification and segmentation system. Section III briefly describes audio and visual features which are extracted. In Section IV, we present a clip-based SVM approach by combining audio and visual features. Experiment results are given in Section V. Finally, Section VI concludes the paper.

II. Overview of Scene Classification and Segmentation System

Support Vector Machine (SVM) [8] has been successfully used in pattern recognition, such as speaker identification, face detection, and text recognition. SVMs are built by mapping the input patterns into higher dimensional feature space using a nonlinear transformation (kernel function), and then optimal hyperplanes are built in the feature space as decision surfaces between classes. Since its inception, the SVM has gained wide attention due to its excellent performance on many real-world problems. Compared to other classifiers that separate the data in its original space, such as k Nearest Neighbor(k-NN), Neural Network (NN), and Naive Bayes (NB), SVM maps non-linear separable data to higher dimensional space and performs separation in that space. It is also reported that SVM can achieve a generalization performance that is greater than or equal to other classifiers, while requiring significantly less training data to achieve such an outcome [9].

In the classification work, it is usually assumed that video is presegmented and the input segment to the SVM classifier belongs to only one scene class. On the other hand, if each short segment of video can be classified correctly, the scene change boundary can be easily detected. Thus, we combine classification and segmentation. Figure 1 shows the block diagram of our system. We present a clip-based SVM for scene classification and segmentation. We first split a video source into audio and video streams which are segmented respectively into separate clips and extracted low/high level audio features and visual features. Passing through the procedure of feature extraction, audio features and visual features of each clip will be combined and transformed into a feature vector. The final step is to classify video scenes using the clip-based SVM and detect scene changes.

III. Feature Extraction

We use a discriminative SVM framework for classify video scene. The effectiveness of any classification scheme depends on the effectiveness of attributes in content representation. During the training phase, the classifier requires input vectors for scene classification. It is our goal to find good features for distinguishing different scene genres and input vectors to the SVM which are relatively low-dimensional. We use the audio and visual features as low-level features. The audio features are used since the audio signals are closely related with the semantic content of videos. The color features, motion features and edge features are used as the visual features since they are widely used and essential visual cues. Finally, the audio and the visual features are integrated into the same feature space

A. Audio Features

The audio signals in our proposed approach are sampled at 11kHz rate, with mono channel and 16 bits per sample. The extracted audio data are firstly segmented into non-overlapping one-second audio clips. It is further divided twenty 50-ms non-overlapping audio frames, on which a 15Hz bandwidth expansion is applied. Then, various features are extracted from each frame or clip to represent it. Currently, these audio features, listed in following subsections, are considered in this work, which are chosen due to their effectiveness in capturing the temporal and spectral structures of different video scenes.

1) Short-time Zero Crossing Rate (ZCR) is computed for every 50ms frame which coarsely measures a signal's frequency content.

2)Bandwidth (BW) is the square root of the power-weighted average of the squared difference between the spectral components and the frequency centroid. i.e..

$$BW = \sqrt{\int_0^{\bar{w}} (w - FC)^2 \left|F(w)\right|^2 dw / E} \quad (1)$$

3) Mean of the spectrum flux (SF) is defined as the average variation of the spectrum between adjacent two frames in a one-second audio clip.

4) The frequencies of audio segment are segmented into four ranges based on the relevant frequencies on speech and music. With 11kHz sampling rate, we define the four frequency sub-bands to be [0, 700Hz], [700, 1400Hz], [1400, 2800Hz], and [2800, 5512Hz]. Analyzing the frequency subbands allows us to better examine a signal's acoustic characteristics. Energy ratios of the above four frequency subbands (ERSB1/2/3/4) is the ratio of subband i's energy to the sum of the four subband energies.

5) Silence Ratio (SR) is defined as the ratio of the amount of silence frames in one audio clip to the total number of frames in the clip. SR is a useful statistical feature for audio classification, and it is usually used to differentiate music from speech [10]. Normally speech has higher SR than music. For each frame, the root mean square (RMS) is computed and compared to the RMS of the whole clip. If the frame RMS is

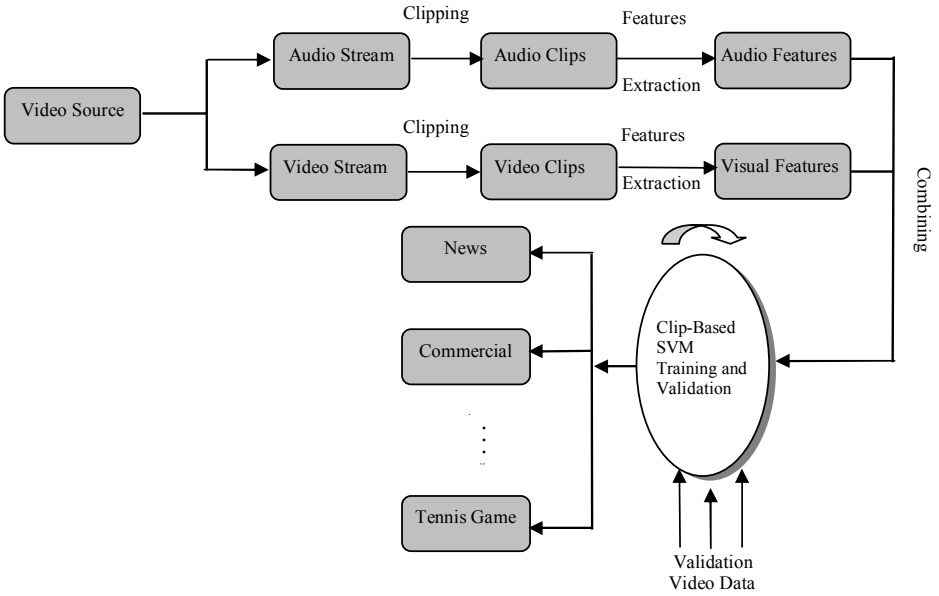

Fig. 1. The block diagram of the proposed system

less than 50% of clip RMS, we consider it as a silence frame.

6) Harmonic sound is defined as one that contains a series of frequencies which are derived from a fundamental or original frequency as a multiple of that. We compute the harmonic frequency of each frame using the algorithm in [11]. The harmonic ratio (HR) is defined as the ratio of the number of frames having a harmonic frequency to the total number of frames in the clip.

7) High ZCR Ratio (HZCRR) is defined as the ratio of the number of frames whose ZCR is above 1.5 fold average ZCR rate to the total number of frames in a one-second clip.

8) Low Shot-time Energy Ratio (LSTER) is defined as the ratio of the number of frames whose short-time energy (STE) values are less than 0.5 times of the average STE to the total number of frames in a one-second clip.

B. Visual Features

Color is the important visual cue in video indexing and retrieval. However, color changes can not be easily related to transition between two shots, the presence of continuous object motion, or camera movement. We need to combine different and more visual features to avoid such problems. Thus, three different types of visual features based on color, edge and motion are extracted. The definitions of these features and their intuitive meanings are discussed in the following subsections.

Color features can include color histogram, dominant colors, and mean and standard deviation of colors. Due to the high dimension of a color histogram, we chose the most dominant color as the color feature in this work. To determine the color feature vector, each video frame is vector-quantized into 64 colors and a color histogram is generated. The most dominant color is determined by selecting the color with the highest histogram value. Each color feature vector includes the RGB values and its appearance frequency, leading to four color features totally.

Motion is another useful visual cue. It is invariant to the changes of color and lighting, theoretically. Different scene genres present different motion patterns. In this paper, we use a simple and effective technique where motion is extracted by pixel-wise differencing of consecutive frames. Each frame is divided into 4-sub images (top-left, top-rights, bottom-left, bottom-right). Each of size equal to half the width and height of the original image. The motion information is computed by the equation:

$$M(t) = \frac{\sum_{x=1}^{w}\sum_{y=1}^{h} I_t(x,y)}{w * h} \tag{2}$$

where $I_t(x,y) = \begin{cases} 1 & if \left| p_t(x,y) - p_{t-1}(x,y) \right| > \beta \\ 0 & otherwise \end{cases}$

where $p_t(x,y)$ and $p_{t-1}(x,y)$ are the pixel values at pixel location (x,y) in t^{th} and $(t-1)^{th}$ frames, respectively. β is the threshold, and w and h are width and height of the sub-image, respectively. A 4D feature is derived from each consecutive frame pair.

Edge direction histogram derives from edge information and is one of the standard visual descriptors defined in MPEG-7 for image and video. It provides good representation of the non-homogeneous textured images. This descriptor captures the spatial distribution of edges. A given image is first segmented into 2*2 sub-images. The edge information is first segmented for each sub-image using Canny algorithm. The domain of the edge directions (0-180) is divided into 5 bins. Thus, an image partitioned into 4 sub-images results in 20 bins. For more detailed descriptions, please refer to [12].

IV. CLIP-BASED SVM FOR SCENE CLASSIFICATION

In order to apply an SVM-based classification algorithm to a video segment, audio and visual features for successive short intervals of the segment are captured and computed in an observation sequence $X = \{x_1, x_2, ..., x_T\}$, where indicates the feature vector computed for interval t. The interval t is equal to one-second, namely video clip, because it is the length of an audio clip in this work. For audio features, except HZCRR, LSTER, SR and HR, the other features are extracted from each audio frame. The means and variances in one audio clip are computed to get clip-based audio features. In addition, visual features are extracted from video frames, so the means and variances in one-second are also computed to get clip-based visual features.

As the features in the different modalities are not highly correlated, the correlation among the audio and visual features was calculated and is shown in Fig. 2. A total of 26 features are considered when deriving this figure: 10 audio features, 8 color features, 4 motion features, and 4 edge features. The first ten features are the audio features consisting of ZCR, BW, SF, ERSB1, ERSB2, ERSB3, SR, HR, HZCRR and LSTER as described in Section II-A. The following four features are edge features which include the means and variances in one video clip. The next eight features are color features which include the means of the most dominant color of each frame and its appearance frequency over the same time duration associated with a video clip, followed by their variances. The last four features are the motion features which include the means and variances in one video clip.

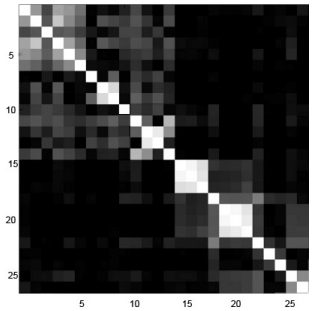

Fig. 2. Normalized correlation matrix between features from different modalities.

Traditionally, each training and testing video sequences merely generates one feature vector to train or test the SVM [13], [14]. The means and variances of the feature trajectories over all video clips are computed, and these statistics are considered as feature sets to represent a video sequence.

A video source is first split into audio and video streams which are segmented respectively into separate clips and extracted low/high level audio features and visual features. Passing through the procedure of feature extraction, audio features and visual features of each clip will be combined and transformed into a feature vector. The steps for scene classification based on clip-based SVM follows:

(1) Assume a N–clip video sequence, $\mathbf{x}_i$, i = 1, …, N , is to be classified into type T_m, $m \in \{1,2,...,M\}$. For each type T_m, and for all the types T_n ($n \neq m$), we compute

$$f_H(T_{m,n} \mid \mathbf{x}_i) = \sum_{i=1}^{N} H(\mathbf{w}\mathbf{x}_i + b) - \sum_{i=1}^{N} H(-(\mathbf{w}\mathbf{x}_i + b)) \quad (3)$$

by the $T_m - T_n$ 2-class SVM. For the optimal hyperlane , $\mathbf{w} \cdot \mathbf{x} + b = 0$, $\mathbf{w} \in R^N, b \in R$, $H(\cdot)$ is Heaviside step function.

(2) The accumulated f for each type T_m is computed by

$$f(T_m \mid \mathbf{x}_i) = \sum_n f_H(T_{m,n} \mid \mathbf{x}_i) \quad (4)$$

(3) The most possible type T_{m^*} is chosen by

$$m^* = \arg\max_m f(T_m \mid \mathbf{x}_i) \quad (5)$$

Finally, we group temporally adjoining one-second clips together if they share the same sound type. As a result, the entire video sequence will be partitioned into homogeneous segments with each having a distinct video class label.

V. EXPERIMENT RESULTS

. In our experiments, the data are collected from real TV programs and movies with eight scene classes: news, commercial, weather forecast, cartoon, MTV, tennis game, football game and basketball game, which are about 8510 seconds in total. The 4567 seconds of data are used for training, and the 3943 seconds of data are used for testing. Six sequences were collected for testing the scene classification and segmentation algorithms. Details regarding the scene content and length of each sequence are summarized in the "ground truth" column of table I. The ground truth is obtained by our manual label. These sequences include various scene class transitions, such as from news to weather forecast, basketball game to commercials, etc, as captured from TV broadcast. The audio stream is sampled at 11kHz with 16 bits per second and the video stream is at 25 frames per second with a resolution of 240*320 pixels per frame. For all the digitized sequences, we extracted the above audio features, color features, motion features and edge features. We set one-second video clip in our experiment as a test unit. If there are two scene classes in a one-second video clip, we will classify it as the time-dominant scene class.

Moreover, to find the optimal SVM parameters such as kernels, variance, margin and cost factor, we have also hand-labeled approximately 5 minutes of news, 2 minutes of weather forecast, 2 minutes of commercial, 3 minutes of cartoon, 2 minutes of MTV, 3 minutes of tennis game, 3 minutes of football game and 3 minutes of basketball game as validation data. Based on the validation results, we choose the radial basis function (RBF) as kernel:

TABLE I
SCENE CLASSIFICATION AND SEGMENTATION RESULT BY THE PROPOSED APPROACHES (UNIT: VIDEO CLIP)

Sequence No.	Ground Truth		Clip-based SVM			
	Class	Duration	Class	Duration	Scene Transition	
1	N C N	000---321 322---382 383---605	N C N	000—318 319—391 392---605	N→C C→N	-3 +9
2	N W	000---430 431---742	N W	000—450 451—742	N→W	+20
3	A C M	000---360 361---495 496---760	A C M C M	000—365 367—478 479—624 624—667 667---760	A→C C→M	+5 -17
4	F C F	000—180 181—407 408—752	F C F C F	000—208 209—415 416—459 470—492 493---752	F→C C→F	+28 +8
5	B F T	000—180 181—361 362---542	B F T	000—174 175—383 384---542	B→F F→T	-6 +22
6	W M A	000—180 181—361 362---542	W M C A	000—206 207—330 331—374 375---542	W→M M→A	+26 —

Scene Class Notation, N: News, C: Commercial, A: Cartoon, M: MTV, W: Weather Forecast, T: Tennis Game, B: Basketball Game, F: Football Game.

$$K(X_i, X_j) = e^{-\gamma D^2(x_i, x_j)} \quad (6)$$

We set the kernel bandwidth γ to 2.0, but could adjust this value for less smoothing if more training data were available. Table I presents the classification and segmentation results obtained by the proposed approaches for the six testing sequences. The results are in the form of time duration of detected segments (classes).

If each short segment of video can be classified correctly, the scene change can be easily detected. A segment is be detected correctly if the majority of that segment is classified correctly. The transition may occur earlier or later than the actual transition. The boundary accuracy will be evaluated in the fourth column of Table I.

The duration of segments detected correctly, the total duration of segments detected, and the total duration of segments in ground truth are computed and shown in Table II according to the data of Table I.

In addition, to evaluate the performance of the presented SVM, we have also carried out the scene classification task using the DT, 5-NN and NN. Figure 3 and 4 show the precision and recall of the four classifiers. Precision and recall are defined as

$$\mathrm{Re}call = \frac{N_C}{N_C + N_M}, \quad \mathrm{Pr}ecision = \frac{N_C}{N_C + N_F} \quad (7)$$

where N_C denotes the duration of correctly detected scenes. N_M indicates the duration of missed ones and N_F is the duration of falsely detected ones.

From these results, we find that commercials have a lower precision and recall than that of other classes. This may due to the fact that the definition of commercials is too board. Therefore, the characteristics displayed by different commercials may be quite different, which results in lower compared with that of the other scene classes. However, in all cases, SVM performs better than the other three classifiers. This is because SVM has a good performance on non-linear separable classes.

VI. CONCLUSION

When such tasks combine semantic information from different data sources (audio, visual) by multimodal integration, scene classification and segmentation will be more efficient. In this paper, we have proposed a clip-based SVM approach to classify and segment video scenes by combining audio and visual features. By utilizing the temporal behaviors of different scene classes, SVM classifier can effectively classify presegmented video clips into one of the predefined scene classes. After identifying scene classes, the scene change boundary can be easily detected. Our comparison experiments show that SVM-based classifier outperforms other popular classifier such as DT, K-NN and NN. However, the fuzzy definition of commercial has a negative effect on the accuracy of SVM (and other classifiers). One challenging task is to further improve the system performance by taking advantage of the integration among multimodal features.

TABLE II
THE DURATION OF SEGMENTS DETECTED CORRECTLY, THE TOTAL DURATION OF SEGMENTS DETECTED, AND THE TOTAL DURATION OF SEGMENTS IN GROUND TRUTH (UNIT: VIDEO CLIP)

Scene Classes	N_c	N_c+N_M	N_c+N_F
News	961	973	981
Commercial	369	420	497
Cartoon	527	540	532
MTV	344	444	361
Weather Forecast	471	491	497
Tennis	158	180	158
Basketball	174	180	174
Football	662	704	718

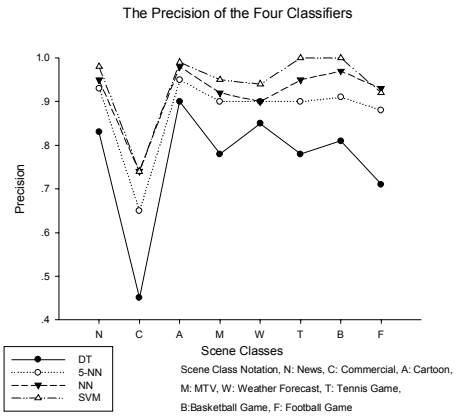

Fig. 3. The precision of the four classifiers

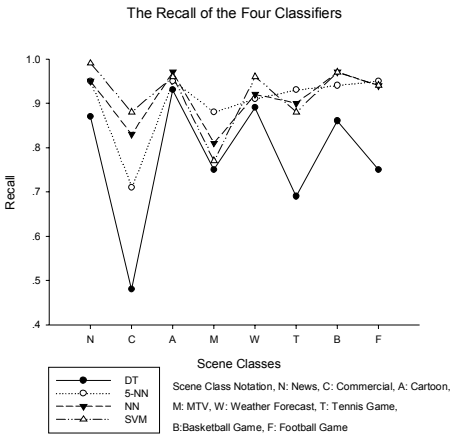

Fig. 4. The recall of the four classifiers

REFERENCES

[1] Y. Li and C.-C. J. Kuo, "Content-based movie analysis and indexing based on audio visual cues," *IEEE Transaction on Circuits and systems for Video Technology*, vol. 14, Dec, 2004, pp.1073-1085.

[2] L. Q. Xu and Y. Li, "Video classification using spatial-temporal features and PCA," in *Proc. International Conference on Multimedia and Expo*. Baltimore, MD, USA, July 6-9, 2003, pp. 345–348.

[3] N. Goela, K. Wilson, F. Niu, A. Divakaran, "An SVM framework for genre-independent scene change detection," In *Proc. IEEE ICME'2007* vol. 3, New York, USA, July.21-24, 2007, pp. 532-535.

[4] S. Eickeler and S. Muller, "Content-based video indexing of TV broadcast news using hidden markov models," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 2997-3000, 1999.

[5] X. Gibert, H. Li, and D. Doermam, "Sports video categorizing using hmm ," in *Proc. International Conference on Multimedia and Expo*. Baltimore, MD, USA, July 6-9, 2003, pp.465-469.

[6] J. Nam, M. Alghoniemy, and A.H. Tewfik, "Audio-visual content-based violent scene characterization," in *Proc. IEEE International Conference on Image Processing*, vol. 1, pp.353-357, 2005.

[7] J. Assflag, M. Bertini, C. Colombo, and A.D. Bimbo, "Sementic annotation of sports videos," *IEEE Multimedia,* vol. 9, pp.52-60, June 2006.

[8] C. J. C. Burges., "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery*, vol.2, no.2, 1998, pp.1-47.

[9] V. Wan and W. M. Campbell, "Support Vector Machines for Speaker Verification and Identification," In *Proc. of the IEEE Signal Processing Society Workshop on Neural Networks*, vol. 2, 2000, pp.775–784.

[10] A. F. Qureshi S. Kiranyaz and M. Gabbouj, " A Generic Audio Classification and Segmentation Approach for Multimedia Indexing and Retrieval," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no.3, 2006, pp.517-523.

[11] S. Pfeiffer, S. Fischer, and W. Effelsberg, "Automatic Audio Content Analysis," In *Proc. 4th ACM Int. Conf. on Multimedia*, 1996, pp.127-132.

[12] V. Suresh., C. Krishna, R. Kumaraswamy, and B. Yegnanarayana, "Combing Multiple Evidence for Video Classification," In *Proc .IEEE ICISIP'2006*, vol. 2, Aug. 15-18, 2006, pp. 187–192.

[13] G. Guo and S. Z. Li, "Content-based Audio Classification and Retrieval by Support Vector Machines," *IEEE Transactions on Neural Networks*, vol. 14, no. 1, Jan. 2003, pp.209-215.

[14] C. C. Lin, S. H. Chen, T. K. Truong, and Y. Chang, "Audio Classification and Categorization Based on Wavelets and Support Vector Machine," *IEEE Transactions on Speech and Audio Processing*, vol.13, no. 5, Sept. 2005, pp.644-651.