THE JOURNAL OF THE ACOUSTICAL SOCIETY OF KOREA VOL.23, NO.3E 2004. 9 pp. 91~97

Classification of TV Program Scenes Based on Audio Information

Kang-Kyu Lee*, Won-Jung Yoon*, Kyu-Sik Park*
*Division of Information and Computer Science, Dankook University
(Received August 2 2004; revised September 8 2004; accepted October 28 2004)

Abstract

In this paper, we propose a classification system of TV program scenes based on audio information. The system classifies the video scene into six categories of commercials, basketball games, football games, news reports, weather forecasts and music videos. Two type of audio feature set are extracted from each audio frame- timbral features and coefficient domain features which result in 58-dimensional feature vector. In order to reduce the computational complexity of the system, 58-dimensional feature set is further optimized to yield 10-dimensional features through Sequential Forward Selection (SFS) method. This down-sized feature set is finally used to train and classify the given TV program scenes using k-NN, Gaussian pattern matching algorithm. The classification result of 91.6% reported here shows the promising performance of the video scene classification based on the audio information. Finally, the system stability problem corresponding to different query length is investigated.

Keywords: *Video scene classification, TV program scene classification, Feature extraction, SFS, k-NN, Gaussian model, GMM*

I. Introduction

As more and more video data, for example, TV broadcasting programs of several tens channels, is stored in multimedia database, user requires more efficient ways of indexing and retrieval for given digital video contents. Recently, video scene classification based on video content has been a growing area of research. As part of a video indexing and retrieval system, it can automatically classify the incoming video file according to its content, thus providing semantic content discrimination capability and enabling fast video browsing and retrieval.

Video data is a representative multimodal information media, containing text, speech, audio, image, motion, etc. Because of this multimodal characteristic, there have been many different ways of approach to the automatic video scene classification system. Most of these works are based on the search of low-level visual features such as color, texture, shape of objects and images, etc[1-3]. However, as pointed out in Ref.[4], visual information alone cannot achieve satisfactory result and audio track in a video can provide very useful and complementary semantic cues to aid scene detection. For this reason, some research works have been done on integrating visual and audio information in video structure and content analysis. In Ref.[5], Saraceno et al. classify audio according to silence, speech, music, or noise and use this information to verify video scene boundaries hypothesized by image-based features. Boreczky[6] used HMM (Hidden Markov Model) framework for video segmentation using both audio features based on cepstrum coefficient and image features of color difference and motions. In[7], Zhang et al. developed audio classification scheme based on heuristic rules and it was used to assist video segmentation. Another interesting work by Liu et al[8] was solely based on the audio information.

Corresponding author: Kyu-Sik Park (kspark@dankook.ac.kr)
Division of Information and Computer Science, Dankook University, San 8, Hannam-Dong, Seoul, Korea, 140-714

There, TV broadcast programs are classified using HMM as one of commercials, basketball games, football games, news reports, and weather forecasts.

Audio as a counterpart of visual information in video sequence got more attention recently for its semantic content discrimination capability and system implementation with low complexity. Up to present, the work in[8] seems to be only one that considers possible use of audio information alone for video scene classification. In this paper, we propose automatic system to classify broad TV program scenes using audio information alone to verify the usefulness audio cues to aid scene detection. The proposed system automatically classifies the TV program scenes into six categories such as commercial, basketball games, football games, news reports, weather forecasts and music videos. At first, TV program sequence is manually segmented such that each sequence is distinguished as one of the six categories. Then 58-dimensional audio feature sets are extracted from each video scene and then these features are further optimized to yield 10-dimensions through SFS (Sequential Forward Selection) method[9]. This down-sized feature set is finally used to train and classify the given TV program using k -NN (Nearest Neighbor), Gaussian model pattern matching algorithm.

This paper is organized as follows. Section II describes the proposed system in brief. Section III explains a method for audio feature extraction and optimization. Extensive experimental results and analyses are given in Section IV. Finally, Section V gives the conclusion.

II. Proposed System

Figure 1 shows the overall structure of the proposed classification system for TV program scenes. The system accepts short TV program clip as an input query and it responses the category of that video scene. The proposed system consists of three principle stages: feature extraction, training of the classifier based on the sample database set, and classification.

First, suitable sample video clips must be accumulated into database. Then the audio data corresponding to each sample video clip is segmented and hamming windowed, and parameterized into so called audio feature vectors. Note that these features must contain unique audio characteristics to be discriminated. These works can be explained by steps from 3 to 6 in figure 1. In this paper, we construct 180 sample video clips from the six categories of TV programs scenes in commercial, basketball games, football games, news reports, weather forecasts and music videos. In contrast to five classes of sample video clips in paper[8], we add one more sample set of music videos.

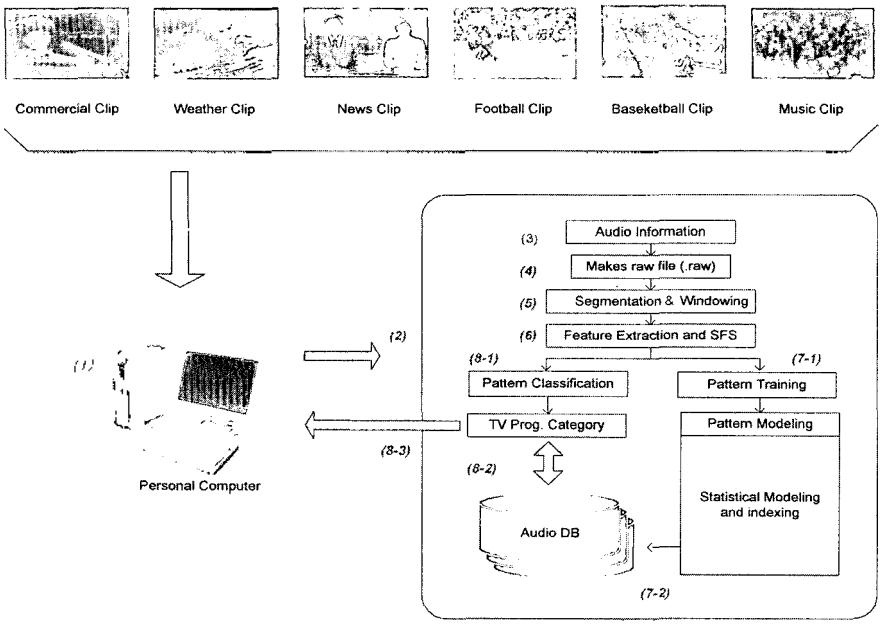

Figure 1. Overall structure of the proposed system

Furthermore a new set of audio feature is developed and used to characterize audio information and it will be further described in section III.

Second, as shown in steps 7-1 and 7-2 of figure 1, the pattern training of the classifier is performed based on the sample database set. Here pattern training, based on some statistical model, is to learn the audio characteristics of six kinds of TV program clips and later it will be used for the video scene classification using k -NN pattern matching algorithm.

For the final stages of classification, whenever the query clip is entered to the system, the system goes through the same stages of feature extraction as in steps from 3 to 6 and the pattern matching between the feature vector of the query and the sample database is performed (steps 8-1 and 8-2). Finally, based on this pattern decision, the category of the query clip is returned as a result of classification (step 8-3).

III. Feature Extraction and Classification Algorithm

3.1. Feature Extraction

Before classification, the audio signals of each TV program clip are normalized to have zero mean and unit variance in order to avoid numerical problems caused by small variances of the feature values as in [8, 9]. At the sampling rate of 22000 Hz, the signals are divided into 23ms frames with 25% overlapped hamming window at the two adjacent frames. Two types of features are computed from each frame: One is the timbral features such as spectral centroid, spectral rolloff, spectral flux and zero crossing rates. The other is coefficient domain features such as mel-frequency cepstral coefficients (MFCC) and linear predictive coefficients (LPC). The means and standard deviations of these six original features are computed over each frame for each music file to form a total of 58-dimensional feature vector. This 58-dimensional feature vector is consist of 8-dimensions in means and variances of spectral centroid, spectral rolloff, spectral flux and zero crossing rates, 26-dimensions over 13 MFCC, and 24-dimensions over 12 LPC.

The following summarizes the audio feature set used in this paper. These features are well-known in the literature and only the short description of definition is given.

- Spectral centroid: It is defined as the center of gravity of STFT magnitude spectrum. The centroid is a measure of spectral brightness
- Spectral rolloff: It is defined as the frequency below when 85% of the magnitude distribution is concentrated. It measures the spectral shape
- Spectral flux: It is the squared difference between the magnitudes of successive spectral distribution. It is a measure of local spectral change
- ZCR: It is the number of time-domain zero-crossings. It measures the noisiness of the signal.
- MFCC: MFCC is the most widely used feature in speech recognition. It captures short-term perceptual features of human hearing system. Thirteen coefficients are used for class classification
- LPC: LPC is a short-time measure of the speech signal with describes the signal as the output of all-pole filter. Twelve coefficients are used for class classification.

3.2. Feature Optimization

Not all the 58-dimensional features are used for classification purpose. Some features are highly correlated among themselves and some feature dimension reduction can be achieved using the feature redundancy. In order to reduce the computational burden and so speed up the search process, while maintaining a system performance, an efficient feature dimension reduction and selection method is needed. In Ref.[9], a sequential forward selection (SFS) method is used to meet these needs. In this paper, we adopt the same SFS method for feature selection to reduce dimensionality of the features and to enhance the classification accuracy. Firstly, the best single feature is selected and then one feature is added at a time which in combination with the previously selected features to maximize the classification success rate. This process continues until all 58-dimensional features are selected. After completing the process, we pick up the best feature lines that maximize the classification success rate. This allows choosing the sub-optimum features for video scene classification. Further details about this procedure are demonstrated in section IV.

IV. Experimental Results and Analysis

4.1. Experimental Setup

The proposed algorithm has been implemented and used to classify TV program scenes from a database of 180 video files. 30 sample clips were collected for each of the six classes in commercial, basketball games, football games, news reports, weather forecasts and music videos, resulting in 180 clips in database. These excerpts of the dataset were taken from internet VOD services provided from KBS, SBS, and MBC TV broadcasting Corp. In order to separate audio from each video clips, we use the CoolEdit digital audio S/W to have 30 second audio clip.

The 180 audio samples are partitioned randomly into a training set of 120 (67%) clips and a test set of 60 (33%) clips. In order to avoid the biased classification results with fixed training and test set, this random division was iterated one hundred times. The overall classification accuracy was obtained as the arithmetic mean of the success rate of the individual iterations.

For classification purposes, a 5sec query is classified using following two pattern matching algorithms; k-NN, Gaussian model [10]. In k-NN, a query to be classified is compared to training data vectors from different classes and classification is performed according to the distance to the k nearest feature points. The implementation of k-NN classifier is quite straightforward but the computational load is high with a large set of training data. The Gaussian classifier models each class as a multidimensional Gaussian distribution. A Gaussian classifier can be fully characterized by its mean vector and covariance matrix in each class. Classification of query is done by finding the class with Gaussian distribution that most likely would produce this query vector.

4.2. Results and Analysis

Three sets of experiment have been conducted in this paper. First, we test the classification performance of the k-NN, Gaussian model with all 58 dimensional audio feature vectors described in section 3.1. Second, the effectiveness of SFS feature optimization method is demonstrated with the same set of classifiers. This result is compared to the previous experiment. Finally, the classification experiment with different query length is performed to point out the system dependency problem on query length.

Figure 2 shows overall classification accuracy using all 58-dimensional audio features set as described in section 3.1. Here overall classification accuracy is defined as the average classification accuracy over the six classes of TV program scenes. From the figure, we see that simple ^{k}-NN classifier gives better performance of 86.3%~90.3% for our task while Gaussian model have lower accuracy 77.1%. However, in general, the classification performance with ^{k}-NN classifier is not satisfactory for the practical TV program scenes classification purpose because the system requires high computational power due to the large 58-dimensional feature set. This is the main reason that we tackle for the some feature optimization method to alleviate this computational complexity problem.

Figure 3 demonstrates SFS feature optimization method described in Section 3.2. As seen from the figure with SFS method, simple k(1)-NN shows rapid convergence speed with higher classification accuracy while Gaussian model needs more features to converge. The reason for this is that the SFS method here is not considering any statistical

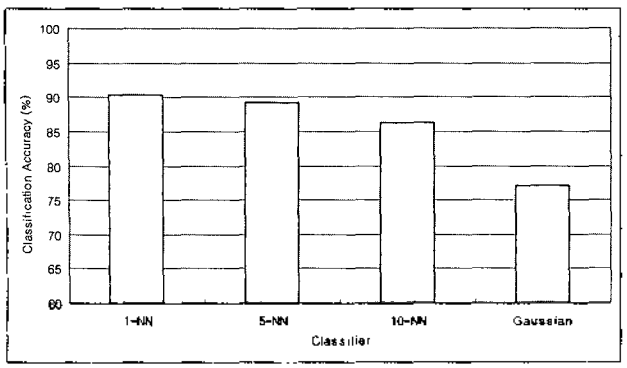

Figure 2. Classification performance using all 58 dimensional audio features

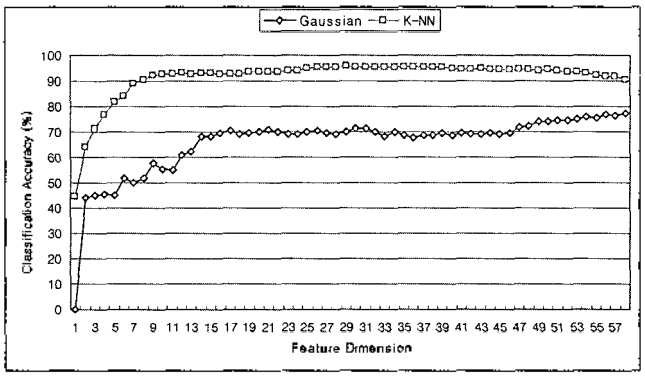

Figure 3. SFS feature selection procedure with k-NN, Gaussian model

Table 1. Class confusion matix with SFS and without SFS(unit:%)

	News Report	Whether forecast	Commercial	Football game	Basketball game	Music video
News report	78.5(75)	19.2	1.0	0	1.3	0
Whether forecast	9.6	87.4(96.2)	0.7	0	2.3	0
Commercia	1.2	1.6	91.5(83.3)	0.2	2.9	2.6
Football game	0	0	0	100(99.9)	0	0
Basketball game	2.9	0.4	0	0.5	96.2(98.7)	0
Music video	0	0	3.1	3.5	0	93.4(88)
Average success rate	91.16 (90.33)					

property of the audio samples and thus any statistical classification method such as Gaussian model will not properly working with the SFS method.

On the other hand, from the figure with k (1)-NN, we clearly see that the classification performance increases with the number of features increase up to 10 with near 90.3% of accuracy. After 10 features, it tends to show only small variation with maximum classification accuracy 95% achieved at 29 features. We note that our purpose of using SFS method is to reduce the feature dimensionality from 58 features derived in section 3.1 while maintaining the classification performance over 90%. This way allows not only reducing the computational burden, but also speeding up the search process. Therefore, we can select only first 10 features to represent each audio portion of video sequence and it will be used all throughout the experiments in this paper.

Table 1 shows detailed SFS performance in TV program scene classification in a form of a confusion matrix. As a comparison purpose, the classification results using 58-dimensional feature vector is included in the table. The percentage of correct classification results with SFS lie in the diagonal of the confusion matrix. The numbers shown in parenthesis represent statistics with 58 dimensional features.

From the table 1, we see that the system with only 10 dimensional features derived from SFS outperforms to the one with 58-dimensional features by at least 1% in average classification accuracy. The SFS method works fairly well over the class of commercials, football game, basketball game, and music videos. But the separation of the news report from weather forecast is less successful. This is not surprising because they contain primarily pure speech in common and the similar result was reported in paper [8]. On the other hand, the classifier can quite accurately distinguish between football and basketball game scene. This comes from the different environmental sound characteristics where each game played. For example, football game is one of typical outdoor sports while the basketball game is one of typical indoor sports.

As pointed out earlier, the classification results corresponding to different query lengths may be much different. It may cause some uncertainty of the system performance. Four excerpts with duration of 30sec, 20sec, 10sec and 5sec are used as a test query. Figure 4 shows the classification results with four excerpts at the prescribed query length. As we expected, the classification results somewhat depend on the query lengths and the performance is getting a little worse as query length's getting shorter. At extreme case of 5 sec query, the classification accuracy is 91.6%, as in previous SFS result, which is little less than the one with 30 sec query. It shows

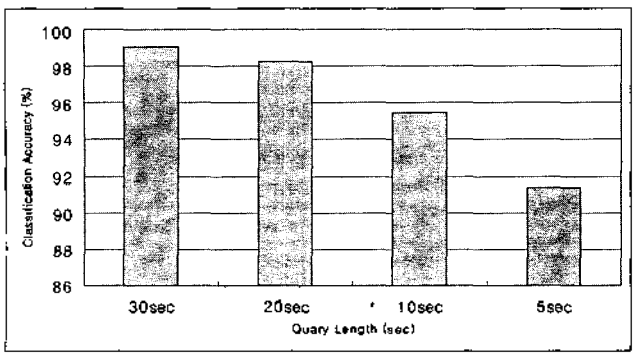

Figure 4. Classification results with different query lengths

an importance of the query length to the overall system performance and can be a starting research point to build up more practical system.

4.3 Comparison Analysis

The proposed video scene classification system was able to achieve 91.6 % average classification accuracy with only 10-dimensional audio feature set which was down-sized and optimized through SFS method. This classification accuracy was measured with a 5 sec audio queried to 180 sample clips across the six categories of TV program scene. Although no absolute comparison analysis with the previous research results is possible because of the different experimental environments, the results achieved here look quite comparative to the classification result of 84.7 % obtained in Ref. [8] where they used 14 feature set with HMM (Hidden Markov Model) pattern matching algorithm to classify the five TV program categories. In other words, in comparison with the result in Ref. [8], the proposed system was able to improve the classification accuracy at least 6% with a less set of audio feature and broader class of TV program categories.

As a final set of experiment, the system stability problem corresponding to different query length is investigated. With four excerpts of different duration as a test query, we showed an importance of the query length to the overall classification accuracy performance. Up to present, we found no previous evidences that investigate this kind of problem and this could be a starting research point to build up more practical video scene classification system.

V. Summary and Conclusion

In this paper, we propose a system that classifies the TV program scenes using audio information alone into six categories of commercial, basketball games, football games, news reports, weather forecasts and music videos. Two types of feature set are extracted from each audio frame-timbral features and coefficient domain features which result in 58-dimensional feature vector. The audio feature set is further optimized to yield 10-dimensional feature set through SFS method. This down-sized feature set is finally used to train and classify the given TV program scenes using k-NN, Gaussian model pattern matching algorithm. From the experimental result of 91.6 % classification accuracy with a 5 sec query, we show the promising performance of the video scene classification based on the audio information with low computational complexity. Also the classification performance depending on the different query length is investigated in terms of the system stability. Future research will involve the development of new audio features, integrated video scene classification system with visual information, and further analysis of the system for practical implementation.

Acknowledgement

This work was supported by grant No. R01-2004-000-10122-0 from the Basic Research Program of the Korea Science & Engineering Foundation.

References

1. A. Yoshitaka and T. Ichikawa, "A survey on content-based retrieval for multimedia databases," IEEE Trans. on Knowledge and Data Engineering, 11(1), Jan. 1999
2. H. Sundaram and S. Chang, "Efficient video sequence retrieval in large repositories," SPIE'99 Storage and Retrieval of Image and Video Databases VII, San Jose, CA, Jan. 1999
3. N. Bryan-Kinns, "A framework for video content modeling," Multimedia tools and applications, **10**, pp. 23-45, 2000
4. H. Jiang, T. Lin and H. Zhang, "Video segmentation with the support of audio segmentation and classification," ICME'2000-IEEE International Conference on Multimedia and Expo, NY, USA, July 2000
5. C. Saraceno and R. Leonardi, "Audio as a support to scene change detection and characterization of video sequences," Proc. Of ICASSP97, Munich, Germany, April 1997, pp. 2597-2600
6. J. Boreczky and L. Wilcox,"A Hidden Markov Model framework for video segmentation using audio and image features," Proc. Of ICASSP'98, pp. 3741-3744, Seattle, May 1998
7. T. Zhang and C. Kuo, "Video content parsing based on combined audio and visual information," SPIE 1999, **IV**, pp. 78-89
8. Z. Liu and J. Huang and Y. Wang, "Classification of TV programs based on audio information using Hidden Markov Model", Proc. of MMSP'98, Redonda Beach, CA, pp. 27-31, Dec 1998.
9. M. Liu and C. Wan, "A study on content-based classification retrieval of audio database," Proc. of the International Database Engineering & Applications Symposium, pp. 339 - 345, 2001.
10. R. Duda, P. Hart and D. Stork, Pattern Classification, 2nd Ed., Wiley-Interscience Publication, 2001

[Profile]

● Kang-Kyu Lee

Kang-Kyu Lee has received his B.S degree from Sangmyung University, Cheon An, Korea in 2003. Now he is attending graduate school for M.S degree in Dankook University, Seoul, Korea.
His research interests are in digital signal, speech, and acoustic processing, and system implementation.

● Won-Jung Yoon

Won-Jung Yoon has received his B.S degree from Sangmyung University, Cheon An, Korea in 2003. Now he is attending graduate school for M.S degree in Dankook University, Seoul, Korea.
His research interests are in digital signal, speech, and acoustic processing, and system implementation.

● Kyu-Sik Park

Kyu-Sik Park has received his B.S., M.S. and PhD. degrees of electrical engineering in 1986, 1988 and 1993 respectively, from Polytechnic University, Brooklyn, New York, USA. He worked as a senior researcher for Samsung electronics from 1994 to 1996. He was a Professor in the Division of Information & Communication at Sangmyung University for 2 years from 1999 to 2001 in Cheon An, Korea. Since 2001 he has been with Dankook University, where he is a professor in the Division of Information & Computer Science. His research interests are digital signal, speech, and audio processing, and digital communication.