2006 International Joint Conference on Neural Networks
Sheraton Vancouver Wall Centre Hotel, Vancouver, BC, Canada
July 16-21, 2006

Environmental Sound Classification Using Hybrid SVM/KNN Classifier and MPEG-7 Audio Low-Level Descriptor

Jia-Ching Wang, Jhing-Fa Wang, Kuok Wai He, and Cheng-Shu Hsu

Department of Electrical Engineering, National Cheng Kung University
1 University Road, Tainan, Taiwan, R.O.C.
Tel: 886-6-2757575 ext. 62341 Fax: 886-6-2080478

ABSTRACT

In this paper, we present a new environmental sound classification architecture. The proposed sound classifier is performed in frame level and fuses the support vector machine (SVM) and the k nearest neighbor rule (KNN). In feature selection, three MPEG-7 audio low-level descriptors, spectrum centroid, spectrum spread, and spectrum flatness are used as the sound features to exploit their ability in sound classification. Experiments carried out on a 12-class sound database can achieve an 85.1% accuracy rate. The performance comparison between the HMM sound classifier using audio spectrum projection features demonstrates the superiority of the proposed scheme.

1. Introduction

The rapid increase in speed and capacity of computers has allowed the inclusion of sound as a type of data in many modern computer applications. Users accustomed to searching, scanning and retrieving text data can be frustrated by the inability to look inside the sound objects. Multimedia databases or file systems, for example, can easily have thousands of sound recordings. Such libraries are often poorly indexed or named to begin with. Searching for a particular sound or class of sound (such as doorbell ringing, laughing, or dog barks) can be a daunting task. The generalized sound classification technique thus plays an important role in multimedia content retrieval.

Also, in our life environment, there are different kinds of sound due to the difference of location. Through the different sound properties, we can easily determine the environment situation. For example, when we hear the fire alarm sound, we can judge there must be fire happening, sound of car horn and judge that a car is waiting behind us. If we can classify and identify in accordance with the sound information, it will be a great help to us for monitoring or understanding surrounding environment. This application is especially useful for the deaf people.

Comparing with speech recognition, a closely related area, research on sound classification is relatively new. Wold *et al*. present a sound retrieval system named Music Fish based on sound classification. This work is a milestone about sound retrieval because of the content based analysis which distinguishes it from previous works [1]. In this system, pitch, harmonicity, loudness, brightness and bandwidth are used as the sound features. The nearest neighbor rule is adopted to classify the query sound into one the defined sound classes. Foote [2] proposes the use of mel frequency cepstral coefficients (MFCCs) plus energy as sound features. The classification procedure is also done by the nearest neighbor rule. More recently, Casey [3] uses decorrelated, dimension-reduced log-spectral features to represent a sound. The generalized sound classification is achieved by train each sound class as a hidden Markov model (HMM). The MPEG-7 sound recognition tools [4], [5] also adopt his proposal. Kin *et al*. [6] present the detail analysis results about the sound classification based on HMM and audio spectrum projection which is one of the MPEG-7 low level audio descriptors.

In this paper, new techniques for sound classification are presented. The system block diagram is depicted in Fig. 1. The proposed classifier makes use of the support vector machine (SVM) and *k* nearest neighbor rule (KNN). The classifier is performed in frame level. For a frame, this classifier converts both the outputs of SVM and KNN into probabilistic scores. These two score are then fused to a frame score. The total score accumulating all the frame score gives the classification result. In the feature selection, we exploit sound classification abilities of three MPEG-7 audio low-level descriptors, spectrum centroid, spectrum spread, and spectrum flatness.

2. Classifier Design

2.1. Standard Support Vector Machine

The SVM theory is a new statistical technique and has drawn much attention on this topic in recent years. An SVM is a binary classifier that makes its decisions by constructing an optimal hyperplane that separates the two classes with the largest margin [7]. It is based on the idea of structural risk minimization (SRM) induction principle [8] that aims at minimizing a bound on the generalization error, rather than minimizing the

0-7803-9490-9/06/$20.00/©2006 IEEE

mean square error. For the optimal hyperplane $\mathbf{w}\cdot\mathbf{x}+b=0$, $\mathbf{w}\in R^N$ and $b\in R$, the decision function of classifying a unknown point $\mathbf{x}$ is defined as:

$$f(\mathbf{x})=\text{sign}(\mathbf{w}\mathbf{x}+b)=\text{sign}(\sum_{i=1}^{N_S}\alpha_i m_i \mathbf{x}_i\cdot\mathbf{x}), \quad (1)$$

where N_S is the support vector number, $\mathbf{x}_i$ is the support vector, α_i is the Lagrange multiplier and $m_i\in\{-1,+1\}$ describes which class $\mathbf{x}$ belongs to.

In most cases, searching suitable hyperplane in input space is too restrictive to be of practical use. The solution to this situation is mapping the input space into a higher dimension feature space and searching the optimal hyperplane in this feature space. Let $\mathbf{z}=\varphi(\mathbf{x})$ denote the corresponding feature space vector with a mapping φ from R^N to a feature space Z. It is not necessary to know about φ. We just provide a function $K(*,*)$ called kernel which uses the points in input space to compute the dot product in feature space Z, that is

$$\mathbf{z}_i\cdot\mathbf{z}_j=\varphi(\mathbf{x}_i)\cdot\varphi(\mathbf{x}_j)=K(\mathbf{x}_i,\mathbf{x}_j). \quad (2)$$

Finally, the decision function becomes

$$f(\mathbf{x})=\text{sign}(\sum_{i=1}^{N_S}\alpha_i m_i K(\mathbf{x}_i,\mathbf{x})+b). \quad (3)$$

Functions that satisfy Mercer's theorem [9] can be used as kernels. Typical kernel functions are the following:

Linear Kernel: $K(\mathbf{x},\mathbf{y})=\mathbf{x}\cdot\mathbf{y}$, (4)

Polynomial: $K(\mathbf{x},\mathbf{y})=(\gamma\cdot\mathbf{x}\cdot\mathbf{y}+c)^d$, (5)

Radial Basis Kernel: $K(\mathbf{x},\mathbf{y})=\exp(-\gamma\cdot\|\mathbf{x}-\mathbf{y}\|^2)$. (6)

2.2. Frame-Based Sound Classification Using Standard SVM

The standard SVM can be used in sound classification. The process of our frame-based classification is illustrated in Fig. 2. The top of Fig. 2 is the waveform of a sound file. The waveform is first segmented into separate frames. Passing through the procedure of feature extraction, each frame will be transformed into a feature vector. After sending these feature vectors to the SVM classifier, each frame will be tagged as +1 or -1 according to (3). Finally, we calculate the sum of all tags. If the sum of all tags in a sound file is greater than zero, this sound file is classified to +1 class. Otherwise, it is classified to -1 class.

2.3. Hybrid SVM/KNN Classifier

The frame-based classifier described in previous subsection gives a solution to sound classification. However, we found a weakness in standard SVM. To clarify this weakness, an example of two-class classification is given in Fig. 3. The cross and the circle are used to represent feature vectors belonging to the -1 class and +1 class, respectively. They are separated by optimal separating hyperplane.

Assume there is one testing vector which belongs to -1 class. If it falls on the location of x, it will be determined as +1 class. But we can find that this feature vector is more like -1 class. Moreover, assume there is one testing vector which belongs to +1 class and falls on location y. Although we can observe that it close the nodes of +1 class more than -1 class, it is still assigned to -1 class according to the hyperplane. These observations motivate us to also consider the nearest neighbor rule to improve a standard SVM classifier.

The proposed method starts from a 2-class classifier. First, this study converts both the outputs of SVM and KNN to probabilistic scores, respectively. Assume a N_F-frame sound file is to be classified into class C_m, $m\in\{-1,+1\}$ and $\mathbf{x}_j$, j = 1, …, N_F is the corresponding feature vector. For class C_m, the distance ratio of the distance between $\mathbf{x}_j$ and optimal hyperplane to the margin distance is defined by

$$R(\mathbf{x}^{(j)})=\frac{\mathbf{w}\mathbf{x}^{(j)}+b}{\|\mathbf{w}\|}\Big/\frac{1}{\|\mathbf{w}\|}=\mathbf{w}\mathbf{x}^{(j)}+b. \quad (7)$$

This study then converts the distance ratio to a value between 0 and +1 through a sigmoid function

$$score_{SVM}(C_m|\mathbf{x}^{(j)})=\frac{1}{1+e^{-R(\mathbf{x}^{(j)})}}. \quad (8)$$

This score denotes a kind of possibility that $\mathbf{x}_j$ is belonged to C_m.

For the k nearest neighbor rule, the k nearest neighbors to the unknown feature vector $\mathbf{x}_j$, irrespectively of class label, are selected from all the training vectors in accordance with the Euclidean distance measure. We then identify the number of vectors, $k_m^{(j)}$, that belong to class C_m, out of these k training vectors. Instead of assigning $\mathbf{x}_j$ to the class C_m with the maximum number $k_m^{(j)}$, this study gives a probabilistic score by

$$score_{KNN}(C_m|\mathbf{x}^{(j)})=\frac{k_m^{(j)}}{k}. \quad (9)$$

With the probabilistic scores of SVM and KNN, the combined frame score is defined as

$$score(C_m|\mathbf{x}^{(j)})=\beta\cdot score_{SVM}+(1-\beta)\cdot score_{KNN}, \quad (10)$$

where β is a weighting factor between 0 and 1.

Therefore, the total score for the N_F-frame testing sound file is given by

$$score(C_m|\mathbf{x}) = \sum_{i=1}^{N_F}\left[\beta\left(\frac{2}{1+e^{-R(\mathbf{x}^{(j)})}}-\frac{1}{2}\right)+(1-\beta)\frac{k_m^{(j)}}{k}\right]. \quad (11)$$

A multi-class sound classification system can be obtained from the two-class hybrid SVM/KNN classifier. Assume there are M sound classes, each pair of the classes are used to train a SVM classifier, i.e. there are totally $M(M\text{-}1)/2$ SVM models. As for the KNN, there is no need to perform any training.

Considering the classification phase, because the hybrid SVM/KNN classifier takes more computational load than standard SVM classifier, the Directed Acyclic Graph [10] strategy is adopted to implement the multi-class classification.

3. Feature Extraction

In the MPEG-7 part 4, the audio framework contains low-level tools designed to provide a basis for construction of higher-level audio applications. Low-level audio descriptors [11] consist of a collection of simple, low complexity features. This study intends to evaluate the sound classification performance of three important MPEG-7 audio low-level audio descriptors: audio spectrum centroid, audio spectrum spread, and audio spectrum flatness. The derivation of these three sound features follows.

(a) Audio Spectrum Centroid

Spectrum centroid is an economical description of the shape of the power spectrum. It indicates whether the power spectrum is dominated by low or high frequencies and, additionally, it is correlated with a major perceptual dimension of timbre; i.e. sharpness. Denote p_i as the power associated with frequency f_i, the spectrum centroid is calculated as:

$$ASC = \sum_i \log_2(f_i/1000)\,p_i \Big/ \sum_i p_i \,. \quad (12)$$

(b) Audio Spectrum Spread

Spectrum spread is an economical descriptor of the shape of the power spectrum that indicates whether it is concentrated in the vicinity of its centroid, or else spread out over the spectrum. It allows differentiating between tone-like and noise-like sounds. The spectrum spread is defined as the RMS deviation with respect to the centroid, on an octave scale:

$$ASS = \sqrt{\sum_i ((\log_2(f_i/1000) - ASC)^2\, p_i) \Big/ \sum_i p_i} \,. \quad (13)$$

(c) Audio Spectrum Flatness

This descriptor describes the flatness properties of the spectrum of an audio signal within a given number of frequency bands. The spectrum of the windowed signal is divided into quarter-octave resolution, logarithmically spaced, overlapping frequency bands. The flatness of a band is defined as the ratio of the geometric mean and the arithmetic mean of the spectral power coefficients within the band.

$$ASF = \sqrt[N]{\prod_{n=1}^{N} c_n} \Big/ \frac{1}{N}\sum_{n=1}^{N} c_n \,, \quad (14)$$

where N is the number of coefficient within a sub-band and c_n is the n-th spectral power coefficient of the sub-band.

This descriptor expresses the deviation of the signal's power spectrum over frequency from a flat shape (corresponding to a noise-like or an impulse-like signal). A high deviation from a flat shape may indicate the presence of tonal components.

4. Experimental Results

Our database contains 12 common kinds of home environmental sounds. These sounds are male speech (50), female speech (50), cough (50), laughing (49), screaming (26), dog barking (50), cat mewing (45), frog wailing (50), piano (40), glass breaking (34), gun shooting (33), and knocking (50). There are totally 527 sound files in our database. The sampling rate is 16 kHz and each sample is 16 bits. For each sound class, half of the sound files are utilized for training and the others are used for testing.

The experimental results of the proposed sound classification system are listed from Table I to Table IV. Among the three MPEG-7 audio features, spectrum flatness (19-dimsension) has much better performance than spectrum centroid (1-dimsension) and spectrum spread (1-dimsension). The classification rate can be much improved by using these three features together. With the radial base kernel function and the combined three features, the classification rate is about 85.1% and is the highest one in our experiments. To compare the performance with MPEG-7 sound recognition tools, Table V lists the experimental results using HMM classifier. The audio spectrum projection feature and the combined feature (spectrum centroid, spectrum spread, spectrum flatness) are respectively adopted to represented a sound wave. This experiment demonstrates the superiority of the proposed system.

There is one interesting observation in the HMM classifier experiments. While the audio spectrum projection was adopted, we found 20 of 25 dog's sound files were misclassified as laugh. This phenomenon can be explained by Figs. 4 and 5. In the view of spectrogram, some dog's sound files to be classified are very similar to the laugh's sound files in the database. This caused audio spectrum projection fail to classify those dog's sound files correctly. However, this drawback was lessened drastically by using the

proposed combined feature set. The misclassified dog's sound files were reduced from 20 to four in Table V.

5. Conclusions

This study has implemented an environmental sound classification system based on the hybrid SVM/KNN classifier and MPEG-7 audio low-level descriptors. In our experimental results, the proposed hybrid SVM/KNN classifier outperforms the HMM classifier in MPEG-7 sound recognition tool. As for the feature selection, although audio spectrum projection is recommended in MPEG-7, the feature set combining audio spectrum centroid, spread and flatness are also proven to provide good classification performance. Future works include the expansion of sound classes and the study of sound segmentation which is a preprocess for sound classification.

References

[1] E. Wold, T. Blum, D. Keislar, and J. Wheaton, "Content-based classification search and retrieval of audio," *IEEE* Multimedia Magazine, vol. 3, pp. 27–36, July 1996.

[2] T. Foote, "Content-based retrieval of music and audio," *Multimedia Storage and Archiving Systems II, Proc. of SPIE*, vol. 3229, pp. 138–147, 1997.

[3] M. A. Casey, "Reduced-rank spectra and minimum entropy priors for generalized sound recognition," *Proceedings of the Workshop on Consistent and Reliable Cues for Sound Analysis*, September 2001.

[4] M. A. Casey, "Sound classification and similarity," *Introduction to MPEG-7: Multimedia Content Description Interface*, New York: Wiley, June 2002.

[5] M. A. Casey, "MPEG-7 sound recognition tools," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 6, pp. 737–747, June 2001.

[6] H. G. Kim, N. Moreau, and T. Sikora, "Audio classification based on MPEG-7 spectral basis representations," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 5, pp. 716–725, May 2004.

[7] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[8] V. Vapnik, *Statistical Learning Theory*, New York: Wiley, 1998.

[9] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, Interscience Publishers, 1953.

[10] J. C. Platt, N. Cristianini, and J. Shawe-Taylor, "Large margin DAG's for multiclass classification," *Advances in Neural Information Processing Systems*, Cambridge, MA: MIT Press, 2000, vol. 12, pp. 547–553.

[11] *Information Technology—Multimedia Content Description Interface—Part 4: Audio*, ISO/IEC CD 15938-4, 2001.

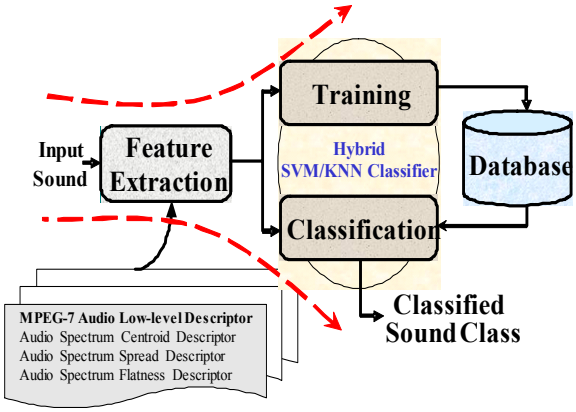

Fig. 1. Block diagram of the proposed sound classification system.

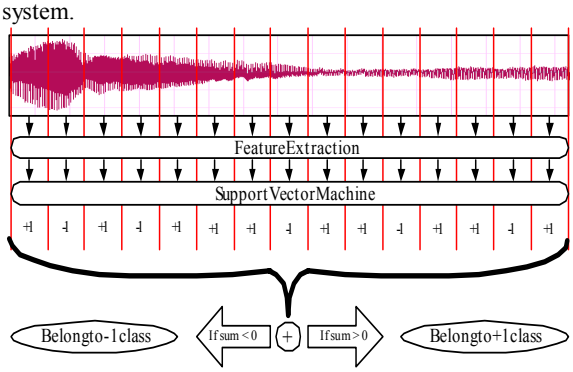

Fig. 2. Concept illustration of frame-based sound classification using standard SVM.

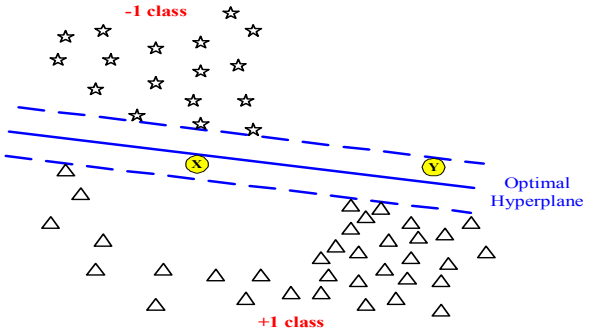

Fig. 3. An example illustrates the weakness of standard SVM.

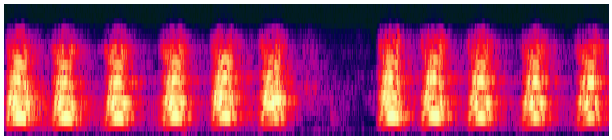

Fig. 4. Spectrogram of a dog's sound file.

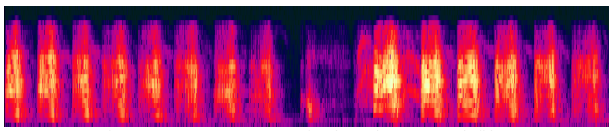

Fig. 5. Spectrogram of a laugh's sound file.

Table I. Classification Result Using Hybrid SVM/KNN Classifier and Audio Spectrum Centroid

Feature	Audio Spectrum Centroid					
Kernel function	Linear		Polynomial		Radial Basis	
hit/miss	hit	miss	hit	miss	hit	miss
cat	10	12	10	12	10	12
cough	0	25	0	25	0	25
dog	0	25	0	25	0	25
female	25	0	25	0	25	0
frog	0	25	0	25	0	25
glass	13	4	13	4	13	4
gun	0	16	0	16	0	16
knock	16	9	18	7	18	7
laugh	11	13	11	13	11	13
male	0	25	0	25	0	25
piano	0	20	0	20	0	20
scream	0	13	0	13	0	13
Accuracy rate	28.6%		29.4%		29.4%	

Table II. Classification Result Using Hybrid SVM/KNN Classifier and Audio Spectrum Spread

Feature	Audio Spectrum Spread					
Kernel function	Linear		Polynomial		Radial Basis	
hit/miss	hit	miss	hit	miss	hit	miss
cat	1	21	1	21	1	21
cough	0	25	0	25	0	25
dog	0	25	0	25	0	25
female	25	0	25	0	25	0
frog	0	25	0	25	0	25
glass	0	17	1	16	1	16
gun	4	12	4	12	4	12
knock	0	25	1	24	1	24
laugh	0	24	3	21	3	21
male	0	25	0	25	0	25
piano	0	20	0	20	0	20
scream	0	13	0	13	0	13
Accuracy rate	11.5%		14.5%		14.5%	

Table III. Classification Result Using Hybrid SVM/KNN Classifier and Audio Spectrum Flatness

Feature	Audio Spectrum Flatness					
Kernel function	Linear		Polynomial		Radial Basis	
hit/miss	hit	miss	hit	miss	hit	miss
cat	19	3	19	3	19	3
cough	2	23	7	18	6	19
dog	10	15	16	9	9	16
female	25	0	25	0	25	0
frog	16	7	20	5	20	5
glass	5	12	6	11	6	11
gun	4	12	4	12	4	12
knock	16	9	18	7	16	9
laugh	24	0	24	0	24	0
male	18	7	22	3	22	3
piano	9	11	4	16	4	16
scream	4	9	4	9	4	9
Accuracy rate	58.0%		66.4%		60.7%	

Table IV. Classification Result Using Hybrid SVM/KNN with Audio Spectrum Centroid, Spread, and Flatness

Feature	ASC+ASS+ASF					
Kernel function	Linear		Polynomial		Radial Basis	
hit/miss	hit	miss	hit	miss	hit	miss
Cat	19	3	19	3	19	3
Cough	5	20	13	12	19	6
Dog	11	14	24	1	24	1
female	25	0	25	0	25	0
Frog	17	8	22	3	24	1
Glass	13	4	13	4	13	4
Gun	11	5	9	7	12	4
Knock	17	8	22	3	22	3
Laugh	24	0	24	0	24	0
Male	18	7	22	3	23	2
Piano	8	12	8	12	8	12
Scream	4	9	7	6	10	3
Accuracy rate	65.6%		80.2%		85.1%	

Table V. Classification Result Using HMM Classifier

Feature	Audio Spectrum Projection		ASC+ASS+ASF	
hit/miss	hit	miss	hit	miss
Cat	22	0	22	0
Cough	23	2	21	3
Dog	5	20	21	4
female	25	0	25	0
Frog	22	3	18	7
Glass	14	3	11	6
Gun	3	13	11	5
knock	18	7	16	9
laugh	22	2	24	0
male	23	2	23	2
piano	17	3	14	6
scream	11	2	12	1
Accuracy rate	78.2%		83.2%	