A survey on Sound Recognition

Suvarna Patil[1], Dhanashree Phalke[2]

[1](Student, M. E., Computer Engineering, D.Y.Patil college of Engineering, Akurdi)
[2](Professor, Computer Engineering, D.Y.Patil college of Engineering, Akurdi)

---***---

Abstract - *Sound is one of the human being most important senses which are used to gather the information. Sound information in video plays an important role in shaping the user feeling and experiences. Sound information in video provides user with an integrated audio visual experience when a sound is not obtainable in video, related test captions are provided the sound information but it not very revealing for non-verbal sound. In this survey, we will offer a qualitative and elucidatory survey on recent development. In this paper, we report on different sound recognition method. Sound recognition is formulated as a classification problem. Performance of some selected methods as compared.*

Key Words: Sound Recognition, Visualizations, Frequency Vector.

1. INTRODUCTION

Sound recognition receives more attention in the recent years due to many applications such as video surveillance [8], Health Care [9], Military and much more. In general, sound recognition scheme can directly handle the samples which are represented in a vector space.

Video is an electronic medium for the recording and display of the moving visual media. Video is collection of frames with sound information. Synchronizing sound information with video frames can give user with an integrated audio visual experience. But users are not always able to acquire the full access of sound e.g. video advertisement in public places, T.V. programs shown in crowded cafe, so the user cannot understand the video for solving this problem to add the text captions in video. But this is not effective for describing the non-verbal sounds e.g. engine noise in car racing video. So, to solving this problem the author presents a framework to automatically convert the non-verbal video sound into animated sound word and position to sound source object for visualization [1].

Sound is one of human beings most important senses. After vision it is the sense most used to gather the information about the environment [4]. Sound information in videos plays an important role in shaping the user feelings and experience. When sound is not available in videos, text captions are used to provide sound information. However, standard text captions are not very expressive and efficient for non-verbal sounds because they are specifically designed to visualize speech sounds [7]. Here author present a framework that automatically transform nonverbal sound of video into animated sound words and also position them for visualization. This provides natural representation of non-verbal sounds with more information about the sound category. It also enhances captions that use animation and a set of standard properties to express basic emotions. Understanding or recognizing context of sounds in environmental surroundings is very important in terms of making the next move based on a sound that occurred [12]. Visualization is a tool for both exploration and communication. Whereas interactive visualization is the key to insightful exploration, animation can effectively convey a complex process or structure. In particular, animation provides a powerful means for illustrating objects, evolution and interaction in a complex environment. Most visualization systems provide some animation support. However, text captions are normally static text at the bottom of the screen, which is not very effective for describing non-verbal sounds, such as those from the environment or sound effects (e.g. engine noise in car-racing video). Further, the dynamics of sounds, such as intensification or attenuation, is important for describing non-verbal sounds, but conveying such dynamics is extremely difficult. Moreover, when multiple sounds are mixed together, users have difficulty in identifying where a sound comes from with static captions [10].

In this work, we will present an updated survey on recent development and focus on the future research and development trends in sound recognition field. We will first discuss the review of literature in Section II. In Section III, we discuss the different methods for sound recognition.
In Section IV, we will present experimental comparison of selected methods. Finally, concluding remark will be given in Section V.

2. Reviews of Literature

The sound recognition and visualization can be divided into speech, music and environmental sound. Several studies have sought to identify different environmental sounds of different categories [4][5][6].

Huadong Wu, Mel Siegel, Pradeep Khosla et al [2] proposed Vehicle Sound Signature Recognition by Frequency Vector Principal Component Analysis. The strength of using adjusted frequency spectrum principal component analysis is that a sound feature is not characterized by just a few specific frequency components; rather the whole spectrum is considered. The key requirement is to build up a properly structured, correctly classified, well-featured sound library. As this would probably be too tedious do manually for a general vehicle identification system, computer-aided supervised learning as well as feasible approaches for unsupervised learning algorithms are both necessary subjects for future research.

AngelosPillos, Khalid Alghamidi, et al [3] proposed A Real-Time Environmental Sound Recognition System for the Android OS. In this paper, a viable real-time environmental sound recognition system for Android mobile devices was developed. The system uses a sound detection algorithm which helps reduce power consumption. In addition, a combination of several feature extracting algorithms was applied to accurately identify sounds. The recognition accuracy of the proposed system exceeds the results of previous efforts using the same dataset.

Ruiwei et al. [7] proposed A System for Visualizing Sound Source using Augmented Reality.
The system is made up with input, process, and output parts. In the process part, author use a PC to process the data, which is gotten from the input part in real-time, and transmit the result of the calculation to the output part. In the output part, author use AR to show the user the visualization of sound of objects in the real-world environment.

S. Chu, S. Narayanan, and C.-C. J. Kuo et al. [4] proposed Environmental sound recognition with time frequency audio features. Author proposed work for recognized the environment sound with time frequency audio features. MFCC have been worked well for the structured sounds such as speech and music. But their performance degrades in presence of noise and the environment sound contains large and diverse variety of sounds. So the Author proposed to use MP (Matching Pursuit) algorithm to analyze environmental sounds.

H. D. Tran and H. Li, et al. [5] proposed Sound event recognition with probabilistic distance SVMs. In this paper, author develops a novel classification methods based on the probabilistic distance SVM using generalized Gamma modeling of STE features.

J. Shen, L. Nie, and T.-S. Chua, et al [6] proposed Smart Ambient Sound Analysis via Structured Statistical Modeling. In this paper, author introduces a framework called SASA i.e. Smart Ambient Sound Analyzes to support different ambient mining task. It extracts a variety of features from different sound component and translates into structured information.

3. Sound Recognition classification methods

A. Matching Pursuit(MP)algorithm with MFCC

MP algorithm was introduced for decomposing signals in an over complete dictionary of function supplying sparse linear expansion of waveforms [4]. The MP features are obtained by-

i) Extracting MP features.
MP is used for feature extraction for classification. It provides an excellent way to accomplish either of these tasks. The Feature extraction is based on the highest energy signal lies in leading synthesizing atoms.

ii) MP dictionary Selection
Decomposition of signals using MP with different dictionaries i.e. Fourier, Haar & Gabor. The Gabor atoms results in the lowest reconstruction error, so it is benefited for the non-homogeneous environmental sound. The Gabor functions are sin modulated Gaussian function that are scaled and translated providing joint time-frequency localization.

iii) Computational Cost of MP feature
The computational cost is a linear function of the total length of signal.

B. Probabilistic distance Supper vector machine

A novel classification method based on probabilistic SVM using the generalized gamma modeling of STE (Sub band Temporal Envelope) features. In this method, authors use the distribution of the STE as sound characterization .The parametric probabilistic distance between the STE distributions of the sound samples are then used to build the SVM classifiers. The probabilistic distance SVM classified into linear SVM and kernel SVM [5] [11].

The sound sample is dissolved into complex Gabor wavelets then the STE are calculated in each of the sub bands. The SPD (Sub band Probabilistic Distance) kernel is then constructed using either the mapping or direct kernel approach.

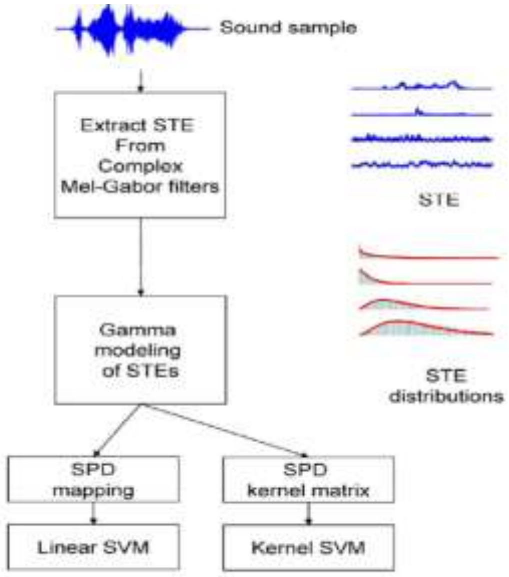

Fig -1: Diagram of Sound event recognition framework [5].

C. Frequency vector principle component analysis

It can be a simple and reliance acoustic identification method. This method consider a frames noise frequency spectrum with R components as an R Dimensional vector then each frame can be considered as a point in this R dimensional frequency spectrum space.

The average adjusted sound spectrum distribution of the set

$\overline{\Phi}1, \overline{\Phi}2 ... \overline{\Phi}N$ is defined by

$$\overline{\Psi} = \frac{1}{N}\sum_{n=1}^{n}\overline{\Phi}_n$$

…………………………… Equation(1)

Each sample differs from the average by a variance vector. This vector variance then subject to PCA i.e. principle component analysis which seek a set of orthonormal vector and their associated eigen values [2].

4. Result Analysis

The MP algorithm with MFCC compares MP and MFCC separately. The MFCC features tend to operate on the extreme which is better than MP feature in some classes but MP features perform better in some different classes. So by combining MP and MFCC features then to achieve an average accuracy rate. They examine the results from varying the number of neighbor's k and using the same k for each environment type [4]. The overall recognition accuracy using KNN with varying number of k as shown in chart 1.

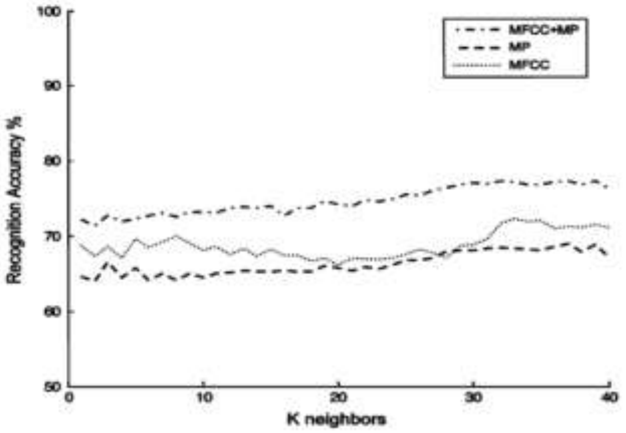

Chart -1: Overall recognition accuracy using KNN with varying number of K [4].

In probabilistic distance SVM method, compare the STE distribution versus STE mean and MFCC. The result can be show that the STE features outperform the MFCC feature in SVM classification test while polynomial SVM offers slight increased [5].

Table I: Baseline Comparison [5]

Baselines

Conditions	MFCC-SVM	MFCC-GMM	STE-SVM	STE-poly SVM
 Office Bus canteen	91.2±1.5 76.4±2.8 55.3±3.1	93.1±2.9 76.7±2.8 58.1±3.9	92.4±2.1 77.2±3.1 59.2±3.5	94.1±2.3 77.0±2.8 58.1±3.2
avg	74.2±3.9	75.9±4.7	75.8±4.2	76.2±4.0
P-value	-	0.13	0.36	0.20

5. CONCLUSION

This paper is a survey of the current research on sound recognition. Sound recognition receives intensive attentions in recent years, due to many practical applications, such as video surveillance, healthcare, and Military and much more. There are different methods like MP with MFCC, probabilistic distance SVM and frequency vector PCA etc. The experimental comparison of multiple methods shows results to gain more insight.

ACKNOWLEDGEMENT

The authors would like to thank the publishers and researchers for making their resources available. We also thank the college authority for providing the required infrastructure and support. Finally we would like to extend our heartfelt gratitude to friends and family members.

REFERENCES

[1] Fangzhou Wang, Hidehisa Nagano, Senior Member, IEEE, Kunio Kashino, Senior Member, IEEE, and Takeo Igarashi, Visualizing Video Sounds with Sound Word Animation to Enrich User Experience", JOURNAL OF LATEX CLASS FILES, VOL. 14, NO. 8, AUGUST 2015.

[2] Huadong Wu, Mel Siegel, Pradeep Khosla Vehicle Sound Signature Recognition by Frequency Vector Principal Component Analysis "May18-20, 1998.

[3] AngelosPillos, Khalid Alghamidi , NouraAlza VeselinPavlov, SwethaMachanavajhala A Real-Time Environmental Sound Recognition System For The Android Os "3 September 2016, Budapest, Hungary.

[4] S. Chu, S. Narayanan, and C.-C. J. Kuo, Environmental sound recognition with time frequency audio features "IEEE Transactions on Audio, Speech, and Language Processing, 2009.

[5] H. D. Tran and H. Li, Sound event recognition with probabilistic distance svms, "IEEE Transactions on Audio, Speech, and Language Processing, 2011.

[6] J. Shen, L. Nie, and T.-S. Chua, Smart Ambient Sound Analysis via Structured Statistical Modeling. "Cham: Springer International Publishing, 2016, pp. 231243.

[7] Ruiwei SHEN, Tsutomu TERADA, Masahiko TSUKAMOTO A System for Visualizing Sound Source using Augmented Reality "MoMM December 3 - 5, 2012.

[8] Cristani, M;Bicego, M;Murino, V : Audio-visual event recognition in surveillance video sequence. IEEE Trans.Multimed.,9 (2)(2007),257-267.

[9] Tentori, Monica, and Jesus Favela. "Activity-aware computing for healthcare", IEEE Pervasive Computing 7.2 (2008): 51-57.

[10] S. Chachada and C.-C. J. Kuo, “Environmental sound recognition: a survey,” APSIPA Transactions on Signal and Information Processing, vol. 3, p. e14 (15 pages), 2014.

[11] C.-C. Chang and C.-J. Lin, “Libsvm: A library for support vector machines,” ACM Transactions on Intelligent Systems and Technology, vol. 2, no. 3, pp. 27:1–27:27, May 2011.

[12] R. Goldhor, “Recognition of environmental sounds,” in 1993 IEEE International Conference on Acoustics Speech and Signal Processing, ser. ICASSP ’93, vol. 1. IEEE, 1993, pp. 149–152.