Interspeech 2018
2-6 September 2018, Hyderabad

Robust Acoustic Event Classification using Bag-of-Visual-Words

Manjunath Mulimani and Shashidhar G. Koolagudi

National Institute of Technology Karnataka, Surathkal, India

manjunath.gec@gmail.com, koolagudi@nitk.edu.in

Abstract

This paper presents a novel Bag-of-Visual-Words (BoVW) approach, to represent the grayscale spectrograms of acoustic events. Such, BoVW representations are referred as histograms of visual features, used for Acoustic Event Classification (AEC). Further, Chi-square distance between histograms of visual features evaluated, which generates kernel to Support Vector Machines (Chi-square SVM) classifier. Evaluation of the proposed histograms of visual features together with Chi-square SVM classifier is conducted on different categories of acoustic events from UPC-TALP corpora in clean and different noise conditions. Results show that proposed approach is more robust to noise and achieves improved recognition accuracy compared to other methods.

Index Terms: Acoustic Event Classification (AEC), Bag-of-Visual-Words (BoVW), Chi-square kernel SVM, histograms of visual features

1. Introduction

Acoustic Event Classification (AEC) is the process of recognition of semantic label of an audio clip, which represents the specific sound in an environment. It has many emerging applications, such as machine hearing [1], human activity recognition [2], audio-based surveillance [3] and so on. Study on AEC is still in its infancy as compared to speech/speaker recognition tasks. However, a recent learning approach called Bag-of-Audio-Words (BoAW), inspired by well known Bag-of-Words (BoW) representation of text documents, used for AEC [4]. BoAW approach generates 'audio words (aural words)' from Low-Level Descriptors (LLDs) such as frame-wise Mel-frequency cepstral coefficients (MFCCs), spectral and temporal features using clustering algorithm. The vector quantizing the LLDs to generate histograms called as BoAW used as feature vectors to the classifier. Recently, BoAW approach even outperforms emerging Deep Neural Networks (DNNs) [5]. However, real-time acoustic events overlapped with high background noise, conventional LLDs are sensitive to noise and may not be suitable for AEC, especially in noisy conditions. In [6], image features such as the combined Histogram of Oriented Gradient (HOG) and Local Binary Pattern (LBP) descriptors from the spectrogram image are represented as BoAW. HOG and LBP features extract edge and texture information from the digital image [7] and may not be suitable for AEC.

In this paper, we use BoVW approach for AEC, widely used for object recognition in computer vision [8]. Only the difference between BoVW and BoAW is that the input feature descriptors. Unlike LLDs as BoAW, Scale Invariant Feature Transform (SIFT) descriptors are commonly represented as BoVW [9]. However, SIFT descriptors effectively recognize objects appear at the different scale, location and poses. Acoustic event in the spectrogram is free from such variations, except variation along time. Hence, SIFT descriptors may not be suitable for AEC. In this work, intensity values of grayscale spectrogram itself considered as feature descriptors (grayscale descriptors) to represent as BoVW and Chi-square kernel SVM used as the classifier. Representation of intensity values of grayscale spectrograms as BoVW omits the feature (descriptor) extraction step and effectively characterizes the acoustic event spectrograms. Robustness of the grayscale descriptors is tested in different noisy conditions.

The rest of the paper is structured as follows, Section 2 explains the proposed BoVW approach in brief. Section 3 explains the experiments carried out in this work. Results are discussed in section 4. The conclusion is given in section 5.

2. BoVW Representations

Overview of the proposed approach given in Figure 1. Initially, the grayscale spectrogram is generated from the acoustic event. Visual words are generated from the grayscale spectrogram using k-means clustering. Finally, rows of spectrograms are quantized to get BoVW as the feature vectors to SVM.

2.1. Gray-scale spectrogram image generation

A spectrogram of an acoustic event is generated using Short-Time Fourier Transform (STFT) [10]. The STFT of an acoustic event is evaluated using Hamming window of length 256 samples with 50% overlap and 44100 Hz sampling rate, which gives the spectrum of complex values (frequencies), have real and imaginary parts. The magnitude of STFT yields linear spectrogram $S(f,t)$; where f is frequency bin (total 129 frequency bins) and t is the time frame. A grayscale intensity spectrogram image is generated (see grayscale spectrogram in Figure 1b) by normalizing the values of Time-Frequency matrix $S(f,t)$ between [0, 1] as given in (1).

$$GI(f,t) = \frac{S(f,t) - min(S)}{max(S) - min(S)} \tag{1}$$

Acoustic events are highly variant to time, which may cause dimensional variations. Hence, grayscale spectrogram image $GI(f,t)$ is transposed as given in (2) to get fixed 129-dimensional row vectors (transposed grayscale spectrogram is shown in Figure 1c).

$$G(t,f) = GI(f,t)^T \tag{2}$$

Unlike 128-dimensional SIFT descriptor from each image patch in the image processing, here, each row of $G(t,f)$ is considered as a 129-dimensional feature vector of intensity values for BoVW representations.

10.21437/Interspeech.2018-1905

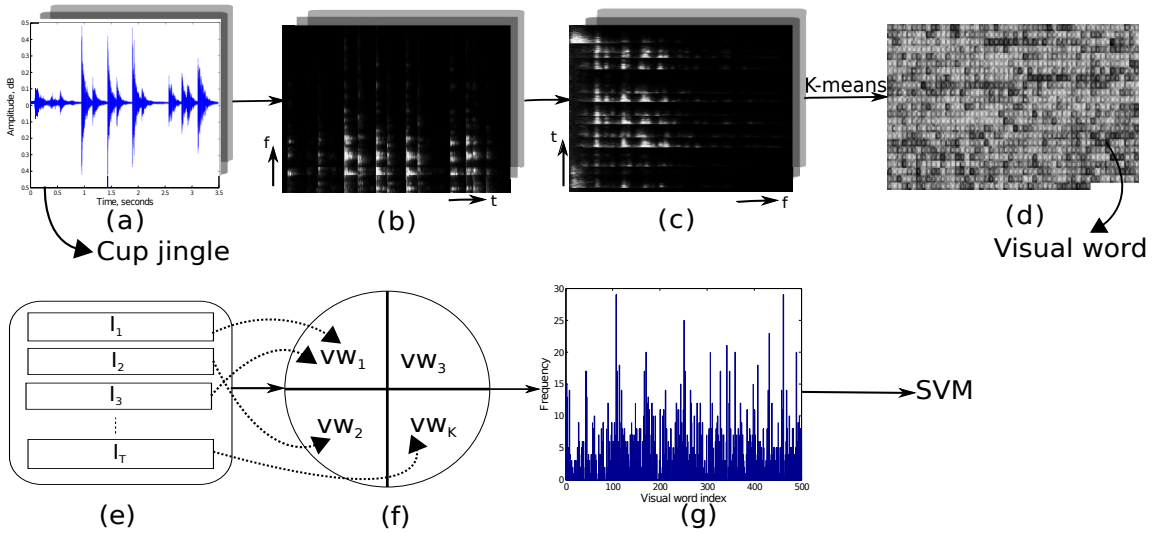


Figure 1: *Overview of proposed approach. Acoustic events (a) are converted into a grayscale spectrogram images (b). Transposed grayscale spectrograms (c). The visual codebook is generated from intensity values of few training grayscale spectrograms using k-means clustering (d). Assigning rows (I_t) of intensity values of spectrogram to the nearest visual word (vw_k) in codebook (e) and (f). Histogram of number of occurrences of each word in spectrogram image is considered as feature vector to SVM for recognition (g).*

2.2. BoVW representations of gray-scale spectrograms

Here, we represent grayscale spectrogram as BoVW. Five randomly selected grayscale spectrograms per class (small training partition) are clustered into a fixed number of clusters using the k-means clustering algorithm [11], which returns the cluster centers, each of which referred as a visual word (codeword). It is worth to point out that five grayscale spectrograms per class are efficient enough to build discriminative visual words with less computational time. All visual words together constitute a codebook or visual vocabulary (shown in Figure 1d). There is no general best practice to select the size of the visual vocabulary, i.e., the number of visual words. In this work, the size of the visual vocabulary ranging from 500 to 2000 considered and its impact on the final AEC accuracy is analyzed.

Once the visual vocabulary is generated, each row of $G(t, f)$ is quantized, i.e., assigned to the closest visual word in the vocabulary using Euclidean distance as shown in Figure 1e and 1f. At this point, a two-dimensional grayscale spectrogram $G(t, f)$ is replaced by a one-dimensional vector V of indices. Each index in V represents the nearest visual word to the corresponding row vector of $G(t, f)$.

Finally, a bag or histogram of visual words is generated from V (see histogram in 1g), which represents the count of occurrence of each visual word in the grayscale spectrogram $G(t, f)$. As we discussed, spectrograms of acoustic events are variant to time. Longer acoustic events generate overall higher histogram counts. Therefore, this time-varying nature of the acoustic event is eliminated by L1 histogram normalization. At this stage, a time-invariant histogram of the grayscale spectrogram image is considered as the feature vector to the classifier.

2.3. Classifier

Here, we use histograms of visual features to train SVM classifier. Linear kernel SVM is simple and has low computational cost; hence, it is more popular. However, linear SVM not considers the nature of input features. SVM can also perform non-linear classification using kernel trick. In this work, we computed Chi-square distance kernel using input features (normalized histograms) for SVM. Chi-square distance between any two normalized histograms h_1 and h_2 is given in (3).

$$\chi^2(h_1, h_2) = \frac{1}{2}\sum_{i=1}^{M}\frac{(h_{1i} - h_{2i})^2}{h_{1i} + h_{2i}} \tag{3}$$

Where M is the number of histogram bins, i.e., number of visual words. The Chi-square distance of histograms is computed using (3) in a pairwise manner and then, it is converted into the kernel using (4) for SVM classification.

$$K_{\chi^2}(h_1, h_2) = e^{-\alpha\chi^2(h_1,h_2)} \tag{4}$$

Where α is a constant scaling factor, which is computed as the mean of Chi-square distance between all training histograms. It is worth to note that, lower the Chi-square distance higher the match between histograms. In addition, we compare the recognition performance of Chi-square kernel with linear and histogram intersection kernel [4], for AEC.

3. Experiments

3.1. Acoustic event corpora

Performance of the proposed approach evaluated on UPC-TALP corpora [12]. A 12 different isolated meeting room acoustic events, namely, applause, chair moving, cough, door knock, door slam, keyboard typing, key jingle, laugh, paper wrapping, phone ring, spoon cup jingle and steps are selected for AEC. Approximately 60 acoustic events per class, recorded using 84 microphones: an array of 64 Mark III microphones, 12

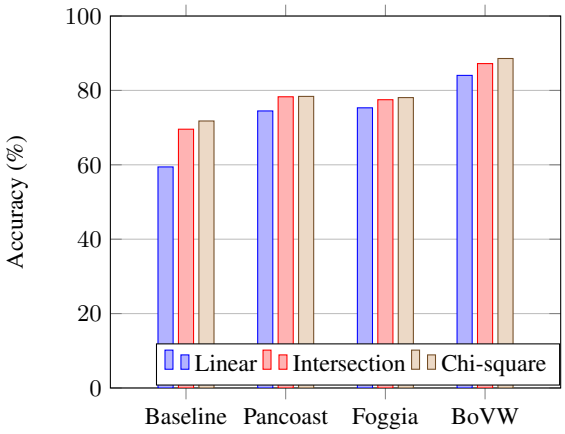


Figure 2: *Average recognition accuracy of proposed BoVW approach versus other methods using linear, intersection and Chi-square kernels SVM*

Table 1: *Comparison of overall Recognition accuracy (%) of proposed BoVW approach with other methods using Chi-square SVM at clean and different SNR.*

Method	Ref.	Clean	20dB	10dB	0dB	Average
Baseline	-	83. 76	80.01	71.71	51.67	71.78
Pancoast et al.	[4]	88.97	86.88	80.42	57.30	78.39
Foggia et al.	[13]	88.34	85.42	76.46	62.09	78.07
BoVW	-	**93.54**	**92.51**	**88.76**	**79.54**	**88.58**

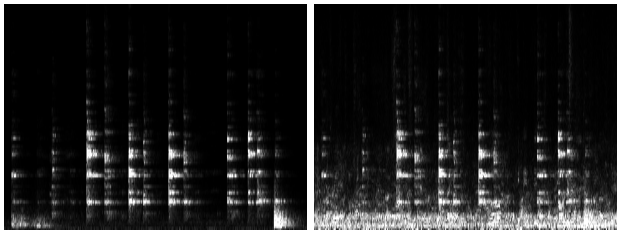
T-shape clusters microphones, 8 table top and omni-directional microphones. In this work, only the third channel of Mark III array is considered for evaluation. Acoustic events are trimmed to the length of given annotations and resulting data is divided into five disjoint folds to perform five-fold cross-validation. Each fold has the equal number of acoustic events per class.
To compare the robustness of the proposed approach, 'speech babble' noise from NOISEX'92 database is added to the acoustic events at 20, 10 and 0dB SNR. Most of the energy of 'speech babble' is distributed at lower frequencies. All acoustic events are processed at 44100 Hz sampling rate.

3.2. Performance comparison

In first set of experiments, performance of the proposed approach is compared with the following baseline system and the state-of-the-art bag of features for AEC.

1. Baseline system : Mean and standard deviation of 13 MFCCs and their first and second-order derivatives are taken over each frame, resulting in 39×2 dimensional feature vector. Further, features are normalized to zero mean and unit variance.
2. Pancoast et. al [4] consider 12 MFCCs and their first and second order derivatives with their log energies are evaluated over each frame, resulting 39-dimensional feature vector represented as Bag-of-Audio-Words (BoAW).
3. Foggia et. al [13] use spectral, temporal, energy (in short refereed as STE) features as a BoAW (aural words) to detect acoustic events in noisy environments.

The MFCCs and STE features used in our experiments are extracted using 20ms hamming window with 50% overlap. Results of all the methods are reported using linear, intersection and Chi-square kernel SVM. Five-fold cross-validation is performed to select optimal parameters of SVM.



(a) (b)

Figure 3: *Acoustic event cup jingle. (a) Grayscale spectrogram at clean condition; (b) grayscale spectrogram at 0dB SNR.*

Further, performance of proposed BoVW approach is also compared with the Spectrogram Image Features (SIFs) for AEC, which are formed by concatenating second and third order moments over 9×9 blocks of monochrome spectrogram images [14].

4. Results and Discussion

4.1. Proposed BoVW approach versus baseline and BoAW approaches

A summary of experimental results is shown in Figure 2. Detailed results at different SNR conditions are given in Table 1. The results (from Figure 2) show that the proposed Chi-square SVM for AEC slightly outperforms the Intersection and reasonably outperforms Linear SVM in all the methods. Chi-square and intersection kernels SVM learn from nature of input features and achieves high recognition rate, unlike linear SVM. Chi-square and intersection kernels SVM commonly used to learn from histogram features in computer vision. Surprisingly, Chi-square SVM with non-histogram MFCC features outperforms linear and intersection kernels. Therefore, hereafter, we consider Chi-square SVM as a competitive classifier in this work. Proposed BoVW with Chi-square SVM outperforms all the methods in clean and all noise conditions (see Table 1). As we discussed, the energy of speech babble concentrated at lower frequencies. MFCC features are sensitive to lower frequencies and noise; hence, the recognition accuracy of the baseline system significantly reduces at 0dB SNR. Noise sensitive MFCC and STE feature used in [4] and [13] as BoAW performs better than frame-average based baseline. However, it is still worse than proposed approach. At this point, we also concatenated MFCC and STE features of [4] and [13] and then represented as BoAW to build competitive method. However, it further reduces the recognition results; hence, it is not considered.

The magnitude of the spectral components of the acoustic event in the linear spectrogram $S(f, t)$ is much higher than that of the noise. Same reflected in grayscale spectrogram as the intensity values of acoustic events are much higher compared to noise. However, noise is commonly more diffuse than acoustic events and maximum energy spread over lower regions of spectrogram image. Strong peaks of acoustic events are unaffected by the noise (see Figure 3), which are effectively discriminate by BoVW from noise.

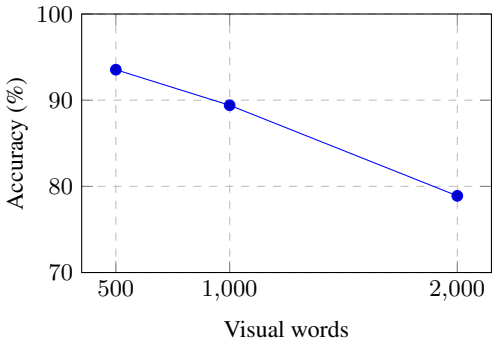


Figure 4: *Recognition accuracy versus size of vocabulary (number of visual words).*

Table 2: *Comparison of overall Recognition accuracy (%) of proposed BoVW approach with SIFs using Chi-square SVM at clean and different SNR.*

Method	Ref.	Clean	20dB	10dB	0dB	Average
SIFs	[14]	76.67	75.84	70.17	56.84	69.88
BoVW	-	**93.54**	**92.51**	**88.76**	**79.54**	**88.58**

Audio words are in the vocabulary are from the low-level speech features which are sensitive to noise. This reduces the significance BoAW model at noisy conditions and achieves poor performance. Grayscale spectrogram image effectively localizes the strongest peaks of acoustic events at 0dB SNR. Hence, visual words of acoustic events are more robust than audio words in noisy conditions. Results show that visual words outperform audio words even at clean conditions.

4.2. Size of vocabulary

The recognition accuracy of Chi-square SVM concerning vocabulary size at the clean condition is shown in Figure 4. It is observed that unlike object recognition in computer vision, the recognition accuracy does not improve as the size of vocabulary increases. Acoustic events are brief and present in the sparse frequency spectrum. Hence, acoustic events can be represented using limited visual words. The 500 visual words effectively recognize the acoustic events with higher recognition rate.

4.3. Proposed BoVW approach versus SIFs

Proposed BoVW approach fully captures the spatial information of intensities of the grayscale spectrogram by considering entire grayscale spectrogram as the feature descriptors. SIFs from two central moments over each image blocks lead to loss of important information. Hence SIFs are outperformed by the BoVW approach in all conditions (see Table 2 for comparison).

5. Conclusions

In this paper, novel BoVW and Chi-square SVM are proposed for AEC. BoVW from grayscale descriptors (rows of transposed grayscale spectrogram) discriminate the strongest peaks of acoustic events from the noise and achieve the high recognition rate compared to all other methods. The results show that BoVW approach achieves 93.54% recognition accuracy in clean condition, it indicates that BoVW approach has the significant contribution towards characterization of acoustic events. BoVW are robust to noise and achieve 79.54% accuracy at 0dB SNR. In future, concatenation of other image-specific features to grayscale descriptors may further improve recognition accuracy.

6. References

[1] R. F. Lyon, "Machine hearing: An emerging field [exploratory DSP]," *IEEE Signal Processing Magazine*, vol. 27, no. 5, pp. 131–139, 2010.

[2] J. A. Stork, L. Spinello, J. Silva, and K. O. Arras, "Audio-based human activity recognition using non-Markovian ensemble voting," in *21st IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 2012, pp. 509–514.

[3] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento, "Audio surveillance of roads: a system for detecting anomalous sounds," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 1, pp. 279–288, 2016.

[4] S. Pancoast and M. Akbacak, "Bag-of-audio-words approach for multimedia event classification," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[5] M. Schmitt, F. Ringeval, and B. W. Schuller, "At the border of acoustics and linguistics: Bag-of-audio-words for the recognition of emotions in speech." in *INTERSPEECH*, 2016, pp. 495–499.

[6] H. Lim, M. J. Kim, and H. Kim, "Robust sound event classification using LBP-HOG based bag-of-audio-words feature representation," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[7] X. Wang, T. X. Han, and S. Yan, "An hog-lbp human detector with partial occlusion handling," in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 32–39.

[8] J. Yang, Y.-G. Jiang, A. G. Hauptmann, and C.-W. Ngo, "Evaluating bag-of-visual-words representations in scene classification," in *Proceedings of the international workshop on multimedia information retrieval*. ACM, 2007, pp. 197–206.

[9] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[10] A. V. Oppenheim, "Speech spectrograms using the Fast Fourier Transform," *IEEE spectrum*, vol. 8, no. 7, pp. 57–62, 1970.

[11] L. Wang, L. Bo, and L. Jiao, "A modified k-means clustering with a density-sensitive distance metric," in *International Conference on Rough Sets and Knowledge Technology*. Springer, 2006, pp. 544–551.

[12] A. Temko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, and M. Omologo, "Clear evaluation of acoustic event detection and classification systems," in *International Evaluation Workshop on Classification of Events, Activities and Relationships*. Springer, 2006, pp. 311–322.

[13] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento, "Reliable detection of audio events in highly noisy environments," *Pattern Recognition Letters*, vol. 65, pp. 22–28, 2015.

[14] J. Dennis, H. D. Tran, and H. Li, "Spectrogram image feature for sound event classification in mismatched conditions," *IEEE Signal Processing Letters*, vol. 18, no. 2, pp. 130–133, 2011.