Ann. For. Res. 65(1): 4225-4241, 2022
ISSN: 18448135, 20652445

ANNALS OF FOREST RESEARCH
https://www.e-afr.org/

BAG-OF-AUDIO-VISUAL WORDS BASED APPROACH FOR SOUND EVENT AND ACOUSTIC SCENE RECOGNITION TASKS FOR INDUSTRIAL MACHINERIES

S Chandrakala, Sreenithi B[a], G Revathy [a] & R Sathya [b]

[a]Intelligent Systems Group, School of Computing, SASTRA University, Thanjavur, Tamil Nadu, India

[b] Assistant Professor, Department of Information Technology, Kongunadu College of Engineering and Technology, Trichy, Tamil Nadu, India.

Abstract

Sound Event Recognition(SER) and Acoustic Scene Recognition(ASR) tasks are gaining more importance due to its applications in personal and public security. Some of the factors complicating the SER and ASR tasks are the quality of audio recording devices, the number of audio sources in a particular environment, and overlapping sound and scene classes. Hence there is a demand to extract different kinds of information from audio to learn a more robust representation of sound events and acoustic scenes. This can be achieved by representing sound in multiple forms to utilize complementary information present in sound data. In this paper, we propose a Bag-of-Audio-Visual Words (BoAVW) approach for the sound event and acoustic scene recognition tasks. The proposed approach constructs Bag-of-Audio words from Mel Frequency Cepstral Coefficient (MFCC) features and Bag-of-Visual words from Scale-Invariant Feature Transform (SIFT), Speeded Up Robust Features (SURF), and Moments-based visual features extracted from auditory images. The Support Vector Machine (SVM) classifier is used to recognize these representations as sound events and acoustic scenes. The proposed BoAVW approach shows improved results when trained on benchmark datasets such as ESC-50 (sound events), DCASE-2016 (sound events), and DCASE-2017 (acoustic scenes). The proposed approach gives 66.6%, 93.2% and 82.58% accuracy respectively when compared with few recent state-of-the-art methods.

*Keywords:*Sound Event Recognition(SER), Acoustic Scene Recognition(ASR), Bag-of-Audio-Visual Words(BoAVW),Mel-Frequency Cepstral Coefficients (MFCCs),Auditory image, Spectrogram, Scale Invariant Feature Transform(SIFT), Speeded Up Robust Features(SURF).

1. Introduction

Sound Event Recognition (SER) and Acoustic Scene Recognition (ASR) tasks have gained a lot of attention over the past few years in the research community [1, 2, 3, 4]. The main focus of SER and ASR tasks is to recognize a variety of environmental sound events and acoustic scenes that occur in the environment around us. The recognition of sound events and acoustic scenes plays a major role in many tasks, such as audio surveillance [5, 6, 7], sound event recognition and retrieval [8, 1, 9], acoustic scene classification [6, 10, 11] and animal sound

classification [12]. The recognition of sound event acts as an additional information to find the happenings in an acoustic scene. In most of the SER and ASR applications, standard Mel-Frequency Cepstral Coefficients (MFCC) based features may be insufficient to identify abnormal sound activities [13]. Incorporating auditory image based visual features along with acoustic features might essentially help to identify a few abnormalities. Sound Event Recognition (SER) and Acoustic Scene Recognition (ASR) tasks are complex tasks with lot of challenges [13]. Since, there are large variety of sound events and acoustic scenes, scalability is a huge factor that affects the performance of SER and ASR systems. In addition to this large diversity, there also exists the possibility of two or more sound events or acoustic scenes having overlapping information, but they belong to different classes. Additionally, noise and reverberations have an impact on sound event and acoustic scene recognition tasks. These challenges increase the complexity of learning sound events/scenes and reduce the performance of automated surveillance systems.

The robustness of sound event and acoustic scene recognition tasks is highly dependent on how sound data is represented. Approaches that utilize deep learning techniques perform reasonably well for SER and ASR tasks [14, 15]. However, these methods require a large amount of training data to efficiently train a model. Although deep learning techniques are proved to be efficient, it is hard to interpret these models, which could increase unpredictability in behaviour concerning different scenarios.

Generally, representation formed using different modalities of raw data provide complementary information for efficient and robust learning. Hence, we focus on extracting features from multiple views of raw sound data which can be used to form a robust hybrid representation of the signal. In this work, we consider two different views of sound data namely spectral sound features and spectrogram-based visual features. We construct the Bag-of-Audio-Visual Words (BoAVW) and the representations obtained are then fed to a discriminative classifier such as Support Vector Machine (SVM) for recognition. Bag-of-Audio words from Mel Frequency Cepstral Coefficient (MFCC) features and Bag-of-Visual words from Scale-Invariant Feature Transform (SIFT), Speeded Up Robust Features (SURF), and Moments-based visual features extracted from auditory images. When compared with existing works of ESC-50 Dataset[38][40], the models uses MFCC whereas our work deals with the advanced features of bag of audio words and bag of visual words showing more accuracy results compared to state of art models. The DCASE-2016 dataset[17[41] is been used with the methodologies variable Q platform and NMF , the proposed BOAWV shows better results compared to state of art models. The DCASE-2017 Dataset[15][16][42] deals with perceptrons, spectrograms, long-mel energies and long-mel band energies the BOAVW approach exhibits more results than the existing models. The rest of the paper is organized as follows: Section 2 presents the related work for the sound event and acoustic scene recognition tasks. The proposed BoAVW approach is presented in Section 3. Sections 4 presents the experimental studies carried out and the performance analysis of sound event and acoustic scene recognition tasks on various datasets.

2. Related work

For an effective recognition of sound events and acoustic scenes, the choice of feature representation plays a vital role. This can be achieved by representing sound data in multiple forms to utilize complementary information present in sound events and acoustic scenes. There are two methods for extracting features from sound examples for effective representation. The first method involves generating spectrograms and using image-based features to train a model. Another method involves using the raw sound signal itself to extract features and learning the acoustic events and scenes.

Most of the SER and ASR tasks are based on supervised machine learning methods for training the model, where a label is given to each audio sample and the machine learning model is trained on those labelled data. Some of the generative classifiers used for effective learning are Hidden Markov Model (HMM) and Gaussian Mixture Model (GMM) [2]. Similarly, discriminative classifiers [16] such as Support Vector Machine (SVM) and Artificial Neural Network (ANN) are also explored in the literature for effective recognition of sound events and acoustic scenes. Besides, several deep learning architectures such as Deep Neural Networks (DNN)[17] [18], Recurrent Neural Networks (RNN) [19] and Convolutional Neural Networks (CNN) [20] [21] are also shown to be effective for highly accurate sound recognition.

The choice of compact feature representation influences the outcome of any SER and ASR tasks. One common method is to convert the sound events and acoustic scenes into spectrograms. Spectrograms are visual representations of sound signal. Some of the visual features extracted from spectrograms are Scale- Invariant Feature Transform (SIFT), Speeded Up Robust Features (SURF) and Histogram of oriented Gradients (HOG). An- other method involves the extraction of sound features such as Mel Frequency Cepstral Coefficients, Constant-Q chromogram, and Spectral flatness directly from raw sound signals. Then, the extracted feature vectors are used as sound event descriptors to train the model.

Most of the sound event and acoustic scene recognition applications in the literature extract the Mel-frequency cepstral coefficients (MFCC's) as a powerful feature from sound signals and use that for training the classifier. Feroze *et.al* [22] proposed a method using features such as loudness, MFCC's, and perceptual linear predictive (PLP) features. It was concluded from the experimental studies that PLP based features outperformed the MFCC based features, for sound event recognition tasks. However, MFCC's are generally preferred over other audio features today. Another method used in [23] extracted features from the spectrogram representation of audio. Jayalakshmi *et.al* [2] proposed an approach based on spectral moments and MFCC features with Support Vector Machine (SVM) classifier. The SVM approach outperformed the generative model-based classifiers such as the Hidden Markov Model (HMM) and Gaussian Mixture Model (GMM) for sound event recognition.

A system submitted for DCASE 2016 Challenge for Sound Event recognition in Synthetic Audio Task 2 involved building a sound event recognition system based on semi-supervised non-negative matrix factorization (NMF), combined with local dictionaries (MLD). A system

proposed by Yanxiong Li *et al.* [24] consists of two main steps: deep audio feature (DAF) extraction and bidirectional long-short term memory classification. MFCC's were extracted from each frame of audio, and DAF features were constructed using deep neural networks. Finally, a combination of LSTM and Bi-Directional Recurrent Neural Networks (BLSTM) was used for classification. This showed moderate performance improvement over existing systems. Another deep network, built using a combination of Convolutional Neural Network (CNN) and a Recurrent Neural Network (RNN) [25]. CNN's have shown to be robust to local and temporal spectral variations and which is capable of extracting high-level features, and RNN's which can learn longer-term temporal context, are combined to form a C-RNN network for the task of Polyphonic Sound Event recognition. Recently, Yi Yu *et.al* [26] proposed a system to classify the audio events through EEG signals by monitoring the brain activity of participants.

The system proposed by Schroder *et.al* [27] for acoustic scene recognition using deep neural networks and time-delay neural networks showed significant performance improvement when compared to systems based on Hidden Markov Models and Gaussian mixture models. The features used here were based on Gabor filter banks, differing from the more traditional MFCC features. Phan *et al.* [28] proposed an approach using a convolutional neural network-label tree embeddings (CNN-LTE) approach. The CNN-LTE approach represented the features n the form of label tree embedding images. Then a simple convolutional neural network was used for learning the representations of different acoustic scenes. Mafra *et al.* [29] proposed a low-level feature representation using temporal averaged Mel-log spectrograms. Then different combinations of features were trained using a multilayer perceptron (MLP), SVM, and CNN classifiers. However, CNN with a log spectrogram approach has not produced better performance when compared to many of the classical SVM-based approaches. In this work, we focus on both SER and ASR tasks by utilizing complementary information present in two different modalities of sound data.

3. Bag-of-Audio-Visual Words (BoAVW) approach for sound event and acoustic scene recognition

The recognition of sound events and acoustic scenes depends on meaningful representations derived from sound signal. From the literature, it has been identified that the hand-crafted features such as Zero Crossing Rate(ZCR), and Spectral Flux (SF) work reasonably well for sound event and acoustic scene recognition. Spectral Flux (SF) work reasonably well for sound event and acoustic scene recognition. But these features lack in capturing the significant discriminative patterns that are present in unconstrained environments since it is highly dependent on the available acoustic events [30]. Generally, representation from different views provide complementary information and it supplies more comprehensive data for learning to enhance the performance of the SER and ASR tasks. Therefore, we focus on multiple views of sound data such as spectral sound features and spectrogram-based visual features for different acoustic environments.

In this work, we propose a Bag-of-Audio-Visual Words (BoAVW) approach with a discriminative classifier such as SVM for SER and ASR tasks. This BoAVW approach is based on the combination of the bag-of-audio-words(BoAW) and bag-of-visual words(BoVW) constructed from sound signal. The proposed BoAVW approach forms the sound event and acoustic scene representations by utilizing features extracted from multiple views of sound signals by complementing sound and visual features. The bag-of-visual words and bag-of-audio words were inspired by the original bag-of-words model of the large text document classification [31]. In this work, we use Short-Time Fourier Transform (STFT), Speeded Up Robust Features (SURF), and spectral moments as visual features from spectrograms and Mel-frequency cepstral coefficients (MFCCs) as sound signal features to construct the BoAVW representation. In the case of spectral moments-based feature extraction, moments are computed across every row or column of pixel values in the spectrogram. Number of rows /columns multiplied by the number of moment features is the dimension of this representation.

3.1. *Bag-of-Visual Words*

The Bag-of-Visual words (BoVW) approach is an extension of the bag-of-words model. The BoVW approach involves the following three steps: Feature Extraction, Codebook Generation, and Histogram Generation. In the feature extraction step, the input data for Bag-of-Visual words is a spectrogram image of audio segments. Spectrograms are a visual representation of the frequency spectrum of a sound signal and it varies with time. It is one of the fundamental and important features used by audio and speech analysis applications. A spectrogram can be defined as an intensity plot of Short-Time Fourier Transform (STFT) magnitude [32]. Fig. 1, Fig. 2, and Fig. 3 illustrate the spectrograms of 'Key-board', 'Cough', and 'Clearthroat' sound event classes respectively from the DCASE-2016 dataset. It can be observed from Fig.1 that, the spectrograms generated for the same sound class do not vary much and looks similar. Whereas for 'Cough', and 'Clearthroat' sound classes, the corresponding spectrograms are dissimilar as shown in Fig. 2, and Fig. 3. Though these two sound classes sound similar (overlapping), the corresponding spectrograms show the discrimination. This helps to reduce the overlap between two different sound classes leading to improved performance.

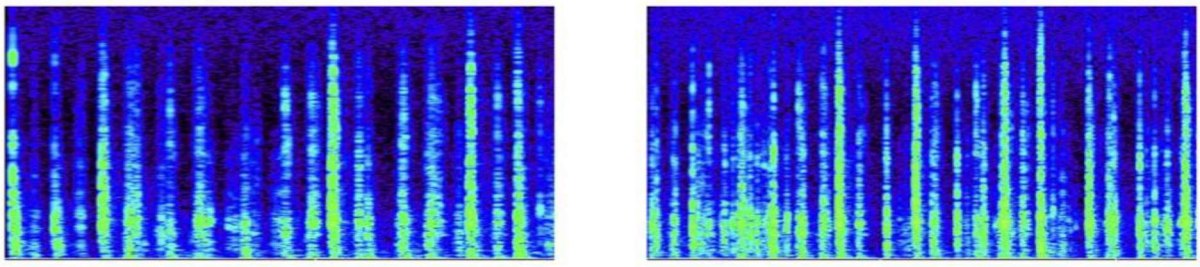

Fig. 1: Spectrograms of 'Keyboard' Class from DCASE-2016 dataset

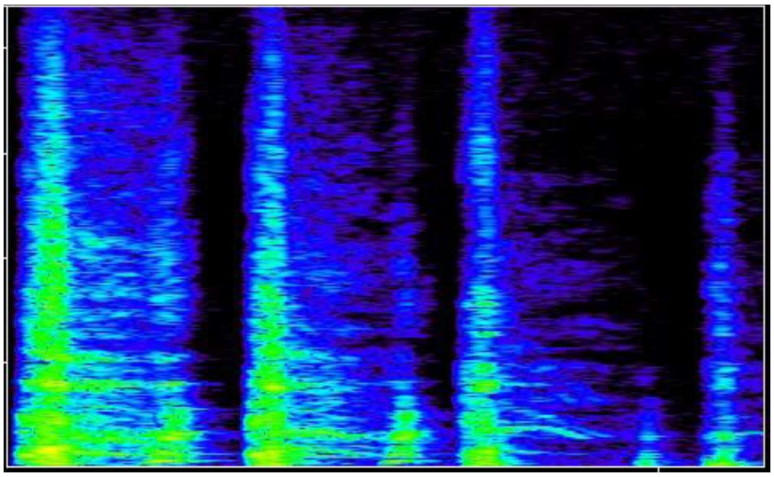

Fig. 2: Spectrogram of 'Cough' sound from DCASE-2016 dataset

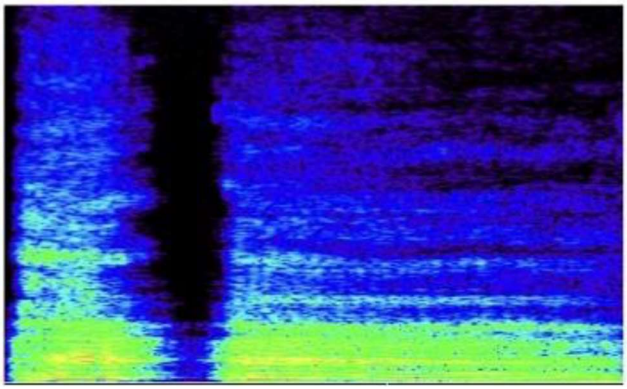

Fig. 3: Spectrogram of 'Clearthroat' sound from DCASE-2016 dataset.

Generally, several key features can be extracted from the spectrogram images. Some commonly used features include Scale-Invariant Feature Transform (SIFT) and Speeded Up Robust Feature (SURF). SIFT is an image-based feature extraction algorithm that uses Euclidean distance for feature matching [33]. SIFT provides features characterizing key points that remain constant(invariant) to changes in scale or rotation. The main motivation behind the SIFT is to find key points in two or more images and to calculate correspondences between them. It involves following 4 key steps for estimating the SIFT descriptors (1) Determine approximate location and scale of key points (2) Improve their location and scale (3) Determine orientation(s) for each key point. (4) Determine a descriptor for each key point. The approximate key point location is calculated by determining the intensity changes using the difference of Gaussians (DoG) at two nearby scales using the following

Equations 1 and 2:

$$D(x, y, \sigma) = L(x, y, k\sigma) - L(x, y, \sigma), \quad (1)$$

where $L(x, y, k\sigma)$ is the convolution of the original image $I(x, y)$ with the Gaussian blur $G(x, y, k\sigma)$ at scale $k\sigma$, i.e.,

$$L(x, y, k\sigma) = G(x, y, k\sigma) * I(x, y) \quad (2)$$

where σ indicates smoothing base scale value and 'k' indicates constant difference between the adjacent scales. The SURF algorithm is a localized feature extraction method which performs the sum of Haar Wavelet responses [34]. SURF is much faster in constructing the visual descriptors and gives optimal performance when compared to SIFT. The SURF algorithm is a localized feature extraction method which performs the sum of Haar Wavelet responses to give optimal responses. The SURF descriptors are extracted according to the equation given below: 3,

$$y = \sum_{y} d, \sum_{x} d, \sum |d|, \sum .d. \quad (3)$$

where *dx*, *dy* represent wavelet responses and |*dx*| , *dy* represent extracted from each spectrogram and each image is now seen as a collection of vectors with equal dimensions.

In the codebook generation step, the extracted features are fed into a k-means clustering algorithm to generate 'codewords' where we cluster the feature vectors. The code words are represented by the cluster centroids. The number of cluster centroids *k* correspond to the size of the codebook. Once the code book has been generated for the given set of images, each patch in an image can now be mapped to a specific codeword through clustering. By doing this, the entire image can be represented as a histogram of codewords called bag-of-Visual words (BoVW). Similarly, the histograms for all the training and test sound segments are generated.

3.2. Bag-of-Audio Words

The bag-of-audio-words (BoAW) [35, 36] is similar to the working of bag-of-visual words in computer vision and bag- of-words for text classification. The BoAW approach extracts audio features from sound data similar to image key points or text occurrences extracted from images and text, respectively. The audio features extracted from each audio segment are given as input to the BoAW approach. The Bag-of-Audio words algorithm also involves the following three steps shown in Fig.3: feature extraction, codebook generation, and histogram generation. Feature extraction plays a viral role in training the model. In speech recognition feature vector represents the speech waveforms In the feature extraction step, the N-dimensional Mel-Frequency Cepstral Coefficient (MFCC's) features are extracted from a short-term power spectrum of audio with a Mel- frequency scale [37]. MFCC features are proved to be effective in many of the sound event and acoustic scene related surveillance applications [6, 38].

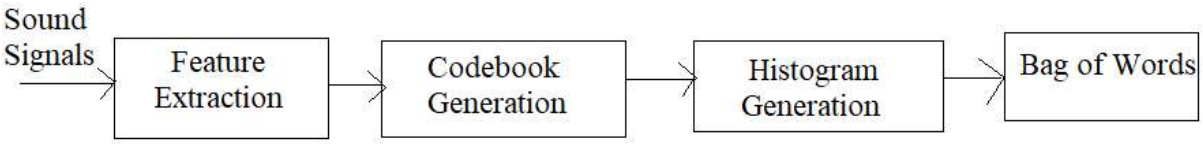

Fig.3. Steps of Bag of Audio words.

The codebooks and histograms for Bag-of-Audio Words are constructed from MFCCs using the similar methods discussed in Bag-of-Visual words approach.

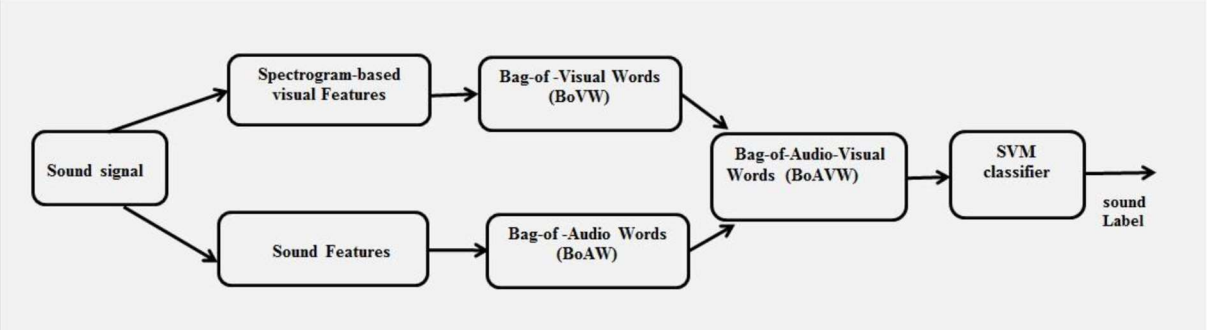

Fig. 4: Block diagram of proposed Bag-of-Audio-Visual words (BoAVW) approach

After generating bag-of-audio words and bag-of-visual words from raw sound signals and spectrograms respectively these representations are combined to form the Bag-of-Audio-Visual words (BoAVW) as shown in Fig.4 to exploit the complementary information present in both the feature spaces. The combined Bag-of-Audio-Visual words are given as input to the Support Vector Machine (SVM) to recognize sound events and acoustic scenes.

4. Experimental Studies

The experiments are carried with Python in Jupyter notebook.

4.1. Datasets Used

The proposed system was trained and tested on the following publicly available sound event and acoustic scene datasets: ESC-50 [38] sound event dataset, DCASE-2016 (Task 2) sound event dataset [17], and the DCASE-2017(Task 1) acoustic scene dataset [39].

ESC-50: This dataset consists of 2000 labelled environmental recordings. It contains 50 classes with 40 instances per class [38]. The data is grouped into 5 major categories with 10 classes for every category: Animal sounds, Natural soundscapes, and water sounds, Human (non-speech) sounds, Interior/domestic sounds, and Exterior/ urban noises.

DCASE-2016: This dataset was part of a challenge for sound event recognition in synthetic audio [17]. This task-focused on event recognition of office sounds in synthetic mixtures. The audio dataset consists of isolated sound events for 11 classes related to office environment: clearing throat, coughing, door knock, door slam, drawer, human laughter, keyboard, keys (placed on a table), page turning, phone ringing, and speech. In the given dataset, each sound class consists of 20 samples, giving a total of 220 training samples.

There are many abnormalities are considered which are not relevant to the DCASE 2016 and DCASE2017.

DCASE 2017: This dataset was used for the task of acoustic scene classification. It consists of 15 acoustic scene classes with 312 instances per class [39]. The 15 acoustic scene classes considered in this task were: Bus, Cafe/restaurant, Car, City centre, Forest path, Grocery store, Home, Lakeside beach, Library, Metro station, Office, Residential area, Train, Tram, and Urban park. Each example takes a 3-5 minutes recording from different acoustic scenes, all

having different recording locations. These recordings were split into 10-second segments. We need to use both sound and acoustic events and hence DCASE-2016 (sound events), and DCASE-2017 (acoustic scenes) both are used in this work.

4.2. Feature Extraction

In this work, we extract 39-dimensional Mel-frequency cepstral coefficients(MFCC) from each audio sample to construct the Bag-of-Audio words (BoAW). Similarly, for all the examples, the spectrogram-based visual features such as SIFT and SURF features are extracted. Besides, from each spectrogram, we also extract moment-based features such as median, mean, and standard deviations. In case of moment-based feature extraction, median, mean, and standard deviations computed across every row or column of pixel values in the spectrogram. Number of rows /columns multiplied by the number of moment features is the dimension of this representation. After the feature extraction step, the extracted visual features were used to construct the Bag-of-Visual words(BoVW), such as SIFT-visual words, SURF-visual words, and moment-based visual words from SIFT, SURF, and moment-based features, respectively. For effective learning, the Bag-of-Audio-Visual Words(BoAVW) is constructed by combining the BoAW and BoVW representations.

4.3. Performance Analysis

The performance of the proposed BoAVW approach and other representation methods are shown in Tables 1, 2, and 3 for ESC-50, DCASE-2016, and DCASE-2017 datasets respectively. Codebook sizes of 300, 500, 700, 900 and 1000 were prepared and experimented for various feature representations such as Audio Words (AW) from MFCC, Visual Words(VW) from SURF, Visual Words(VW) from SIFT, Visual Words(VW) from SURF, Audio-Visual words (SIFT), Audio-Visual Words (SURF), and AVW (Moments). These various representations are fed as input to the support vector machine classifier for recognition. The SVM was used for recognition with linear kernel and 10-fold cross-validation. The proposed approach is compared with other existing systems in the literature and are presented in Tables 4, 5, and 6.

Performance comparison for different combinations of audio and visual words for the ESC-50 dataset is shown in Table 1. We observe that a combination of audio words and visual words improved the recognition accuracy when compared to approaches trained on conventional audio words and visual words. It is also inferred that the approach that uses MFCC- audio words and SIFT-visual words with a codebook size of 700 has achieved a higher recognition rate of 66.60%, which has outperformed other state-of-the-art approaches proposed in the literature for ESC-50 dataset as shown in Table 4. Surprisingly, even with a codebook size of 300, the proposed approach can discriminate well between the sound event classes, which proves the robustness of the BoAVW-based representation.

The performance of the proposed approach for another sound event dataset DCASE-2016 is shown in Table 2. It can be observed that the audio words (AW) with SURF combination for

the codebook size of 500 achieved maximum classification accuracy of 93.2%. Significant performance improvement is achieved when compared to Random Forest Ensemble Classifiers [38] and Deep Neural Networks [40, 42, 25] shown in Table 5. Similarly, the performance of the proposed approach for the acoustic scene dataset DCASE-2017 is shown in Table 3. It can be observed that the audio words(AW) with SURF combination for the codebook size of 900 achieved maximum classification accuracy of 82.58% for DCASE-2017 dataset when compared to other methods in the literature [16, 42]. A possible justification for the improved performance of the proposed approach could be the complementary information that is exploited by combining multiple views of the raw sound data.

From Tables 1, 2, and 3 we can observe that the proposed system trained on a combination of Audio Words with SURF and SIFT Visual Words, is consistently outperforming the systems trained on audio words and visual words individually. In the case of the ESC-50 dataset, Table 4 shows the performance reported in [38]. Piczak [38] proposed a system using ZCR and MFCC features with Random Forest and SVM classifier for classifying 50 different environmental sound classes. In an- other work, Piczak [40] evaluated the potential of convolutional neural networks with log-scaled Mel-spectrograms for recognizing the short duration environmental sounds. The proposed BoAVW approach produced a much better performance with an accuracy of 26 % improvement when compared to the system reported in [38].

Similarly, Table 5 shows the recognition accuracy of the proposed approach and some of the state-of-the-art methods reported in the literature for the DCASE-2016 sound event dataset. In the case of DCASE-2016, an accuracy comparison of the proposed approach with the state-of-the-art approaches are shown in Table 5. The proposed system outperforms the Variable-Q transform (VQT) with Non-negative Matrix Factorization (NMF) system by giving a 56% increase in classification accuracy. Table 5 adds some of the systems submitted in the DCASE-2016 challenge for sound event recognition [17]. The proposed BoAVW approach outperformed the following systems in the DCASE-2016 challenge: time-frequency representations of constant-Q transform (CQT) with RNN classifier, Gammatone cestrum with Random forests classifier, the Bi-Directional Long-Short Term Memory (BSLTM) with Mel-Filter Bank features, and Non-Negative Matrix Factorization with a Mixture of Local Dictionaries (NMF-MLD). From the above observations, it is clear that significant improvement can be achieved even with simpler models trained on meaningful representations rather than using data-hungry deep feature learning techniques.

Codebook Size	Audio Words (AW)	Visual Words(VW) from SURF	Visual Words(VW) from SIFT	Visual Words(VW) from Moments	AVW(SURF)	AVW(SIFT)	AVW(Moments)

300	41.25 %	40.00%	56.80%	30.00%	51.25%	66.40%	**31.15%**
500	42.75 %	40.85%	57.15%	21.30%	50.35%	65.05%	29.50%
700	43.65 %	38.40%	57.20%	21.20%	**53.90%**	**66.60%**	27.65%
900	41.36 %	41.85%	56.57%	22.45%	48.90%	65.55%	29.65%
1000	42.45 %	42.95%	56.75%	22.15%	49.35%	65.35%	29.60%

Table 1: Performance comparison for ESC-50 sound event dataset

Codebook Size	Audio Words (AW)	Visual Words(VW) from SURF	Visual Words(VW) from SIFT	Visual Words(VW) from Moments	AVW(SURF)	AVW(SIFT)	AVW(Moments)
300	68.18 %	89.54%	84.54%	65.00%	93.18%	**85.00%**	**72.72%**
500	68.63 %	87.27%	85.45%	62.27%	**93.2%**	84.09%	72.27%
700	61.81 %	89.09%	82.72%	63.18%	90.45%	83.63%	66.81%
900	61.36 %	87.72%	83.65%	65.00%	92.01%	84.50%	70.00%
1000	61.37 %	89.00%	81.36%	65.00%	91.81%	**85.00%**	69.09%

Table 2: Performance comparison for DCASE-2016 sound event dataset

Codebook Size	Audio Words (AW)	Visual Words(VW) from SURF	Visual Words(VW) from SIFT	Visual Words(VW) from Moments	AVW(SURF)	AVW(SIFT)	AVW(Moments)
300	77.99 %	50.33%	53.33%	38.56%	79.58%	81.43%	78.78%
500	77.99 %	53.69%	55.71%	39.75%	80.32%	81.11%	79.80%

700	76.97 %	52.50%	57.21%	40.35%	80.06%	79.59%	78.11%
900	78.03 %	54.39%	54.05%	40.12%	80.00%	**82.58%**	78.78%
1000	80.58 %	53.32%	53.50%	39.63%	**81.76%**	82.15%	**80.89%**

Table 3: Performance comparison for DCASE-2017 acoustic scene dataset

Approaches	Accuracy
Zero-crossing rate and MFCC+SVM [38]	39.6%
Zero Crossing Rate and MFCC + Random Forest [38]	44.3%
Log-scaled mel-spectrograms+ CNN [40]	64.50%
Proposed BoAVW approach	**66.60%**

Table 4: Comparison of proposed approach with state-of-the-art- methods for ESC-50 dataset.

Approaches	Accuracy
Variable-Q transform (VQT)+Non-negative Matrix Factorization (NMF) [17]	37.0%
Constant-Q transform (CQT) + Recurrent neural networks(RNN)[17]	52.8 %
Gammatone cepstrum+ Random forests [17]	64.8%
Mel-Filter Bank +BLSTM [17]	78.1%
Mel energy+ Deep neural networks (DNN) [17]	78.7%
NMF + MLD[41]	80.2%
Proposed BoAVW approach	**93.2%**

Table 5: Comparison of proposed approach with state-of-the-art- methods for DCASE-2016 dataset.

Approaches	Accuracy

Log Mel-band energies+ Multilayer Perceptron(MLP) [16]	74.8%
spectrogram + CQT+ CNN [16]	77.7 %
log-mel energies+ CNN [16]	80.4 %
log-mel band energies + CRNN [15]	80.8%
MFCC and IMFCC +Deep Neural Network [42]	81.8%
Proposed BoAVW approach	**82.58%**

Table 6: Comparison of proposed approach with state-of-the-art- methods for DCASE-2017 dataset.

Table 6 summarizes the various approaches used in the literature for the DCASE-2017 acoustic scene dataset. Heittola *et al*.[16] provided the system using a multilayer perceptron architecture (MLP) trained with log Mel-band energies for different environmental audio scenes. Our proposed approach gives al- most 8% increased performance when compared to the system in [16]. The proposed BoAVW approach outperforms the various approaches in the DCASE-2017 challenge [16]. Jallet *et al.* [15] proposed the new CNN based network called convolutional recurrent neural network (CRNN). This network was trained using the log Mel-band energies of an audio signal. The highest performance was achieved in the literature by using Deep Neural Network (DNN) with MFCC and inverse MFCC features [42]. Many of the top-performing systems in the DCASE-2017 challenge were based on deep models and the analysis shows that proposed SIFT features based BoAVW approach provide better discrimination among sound classes leading to highest accuracy.

5. Conclusion

Recognition of sound event and acoustic scene is considered to be the core importance of audio surveillance applications. In this paper, we presented the Bag-of-Audio-Visual Words (BoAVW) approach for both Sound Event Recognition (SER) and Acoustic Scene Recognition (ASR) tasks. Bag-of-Audio Words(BoAW) from spectral sound features and Bag-of-Visual Words (BoVW) from spectrogram-based visual features were used to form BoAVW. Experimental studies proved that the proposed audio words with SIFT and audio words with SURF features outperformed the methods reported in the literature for ESC-50, DCASE-2016, and DCASE-2017 datasets and shows results more than state of art models. The approach is able to give a best result for the sound recognition and sound identification.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Data Availability

The data used to support the findings of this study are included within the article.

Acknowledgement

The authors would like to thank Cognitive Science Research Initiative (CSRI), Department of Science and Technology, Government of India and SASTRA University to carry out this work.

References

[1] Q. Kong, Y. Xu, I. Sobieraj, W. Wang, M. D. Plumbley, Sound event detection and time–frequency segmentation from weakly labelled data, IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) 27 (4) (2019) 777–787.

[2] S. Jayalakshmi, S. Chandrakala, R. Nedunchelian, Global statistical features-based approach for acoustic event detection, Applied Acoustics 139 (2018) 113–118.

[3] H. Liu, J. Zhou, G. Xi, B. Peng, S. Zhang, Q. Xiao, Research on acoustic events recognition method with dimensionality reduction combining attention and mutual information, IEEE Sensors Journal 22 (9) (2022) 8622–8632.

[4] C. Brignone, G. Mancini, E. Grassucci, A. Uncini, D. Comminiello, Efficient sound event localization and detection in the quaternion domain, IEEE Transactions on Circuits and Systems II: Express Briefs 69 (5) (2022) 2453–2457.

[5] N. Shreyas, M. Venkatraman, S. Malini, S. Chandrakala, Trends of sound event recognition in audio surveillance: A recent review and study, in: The Cognitive Approach in Cloud Computing and Internet of Things Technologies for Surveillance Tracking Systems, Elsevier, 2020, pp. 95– 106.

[6] S. Chandrakala, S. Jayalakshmi, Generative model driven representation learning in a hybrid framework for environmental audio scene and sound event recognition, IEEE Transactions on Multimedia 22 (1) (2020) 3–14.

[7] C. Clavel, T. Ehrette, G. Richard, Events detection for an audio-based surveillance system, in: 2005 IEEE International Conference on Multi- media and Expo, IEEE, 2005, pp. 1306–1309.

[8] H.-G. Kim, J. Y. Kim, Environmental sound event detection in wireless acoustic sensor networks for home telemonitoring, China Communications 14 (9) (2017) 1–10.

[9] C.-Y. Wang, T.-C. Tai, J.-C. Wang, A. Santoso, S. Mathulaprangsan, C.- C. Chiang, C.-H. Wu, Sound events recognition and retrieval using multi- channel sparse coding convolutional neural networks, IEEE/ACM Trans- actions on Audio, Speech, and Language Processing (2020).

[10] Y. Yin, R. R. Shah, R. Zimmermann, Learning and fusing multimodal deep features for acoustic scene categorization, in: 2018 ACM Multimedia Conference on Multimedia Conference, ACM, 2018, pp. 1892–1900.

[11] B. Elizalde, A. Kumar, A. Shah, R. Badlani, E. Vincent, B. Raj, I. Lane, Experiments on the dcase challenge 2016: Acoustic scene classification and sound event detection in real life recording, arXiv preprint arXiv:1607.06706 (2016).
[12] E. S¸ as¸maz, F. B. Tek, Animal sound classification using a convolutional neural network, in: 2018 3rd International Conference on Computer Science and Engineering (UBMK), IEEE, 2018, pp. 625–629.
[13] S. Chandrakala, S. Jayalakshmi, Environmental audio scene and sound event recognition for autonomous surveillance: A survey and comparative studies, ACM Computing Surveys (CSUR) 52 (3) (2019) 1–34.
[14] H. Purwins, B. Li, T. Virtanen, J. Schluter, S.-Y. Chang, T. Sainath, Deep learning for audio signal processing, IEEE Journal of Selected Topics in Signal Processing 13 (2) (2019) 206–219.
[15] H. Jallet, E. Cakir, T. Virtanen, Acoustic scene classification using CRNN, Tech. rep., DCASE2017 Challenge (September 2017).
[16] T. Heittola, A. Mesaros, DCASE 2017 challenge setup: Tasks, datasets and baseline system, Tech. rep., DCASE2017 Challenge (September 2017).
[17] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, M. D. Plumbley, Detection and classification of acoustic scenes and events: Outcome of the dcase 2016 challenge, IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) 26 (2) (2018) 379– 393.
[18] I. McLoughlin, H. Zhang, Z. Xie, Y. Song, W. Xiao, Robust sound event classification using deep neural networks, IEEE/ACM Transactions on Audio, Speech, and Language Processing 23 (3) (2015) 540–552.
[19] F. Colangelo, F. Battisti, M. Carli, A. Neri, F. Calabro´, Enhancing audio surveillance with hierarchical recurrent neural networks, in: 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), IEEE, 2017, pp. 1–6.
[20] O. F. Civaner, M. Kamasak, Classification of pediatric snoring episodes using deep convolutional neural networks, in: 2018 26th Signal Process- ing and Communications Applications Conference (SIU), IEEE, 2018, pp. 1–4.
[21] S. Adavanne, A. Politis, J. Nikunen, T. Virtanen, Sound event localization and detection of overlapping sources using convolutional recurrent neural networks, IEEE Journal of Selected Topics in Signal Processing 13 (1) (2018) 34–48.
[22] R. Grzeszick, A. Plinge, G. A. Fink, Bag-of-features methods for acoustic event detection and classification, IEEE/ACM Transactions on Audio, Speech, and Language Processing 25 (6) (2017) 1242–1252.
[23] E. Benetos, G. Lafay, M. Lagrange, M. D. Plumbley, Polyphonic sound event tracking using linear dynamical systems, IEEE/ACM Transactions on Audio, Speech, and Language Processing 25 (6) (2017) 1266–1277.

[24] Y. Li, X. Li, Y. Zhang, W. Wang, M. Liu, X. Feng, Acoustic scene classification using deep audio feature and blstm network, in: 2018 International Conference on Audio, Language and Image Processing (ICALIP), IEEE, 2018, pp. 371–374.
[25] F. Vesperini, L. Gabrielli, E. Principi, S. Squartini, Polyphonic sound event detection by using capsule neural networks, IEEE Journal of Selected Topics in Signal Processing 13 (2) (2019) 310–322. doi:10. 1109/JSTSP.2019.2902305.
[26] Y. Yu, S. Beuret, D. Zeng, K. Oyama, Deep learning of human perception in audio event classification, in: 2018 IEEE International Symposium on Multimedia (ISM), IEEE, 2018, pp. 188–189.
[27] J. Schroder, N. Moritz, J. Anemuller, S. Goetze, B. Kollmeier, J. Schroder, N. Moritz, J. Anemu¨ller, S. Goetze, B. Kollmeier, Classifier architectures for acoustic scenes and events: implications for dnns, tdnns, and percep- tual features from dcase 2016, IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) 25 (6) (2017) 1304–1314.
[28] H. Phan, P. Koch, L. Hertel, M. Maass, R. Mazur, A. Mertins, Cnnlte: a class of 1-x pooling convolutional neural networks on label tree embeddings for audio scene classification, in: Proc. ICASSP, 2017.
[29] G. Mafra, N. Duong, A. Ozerov, P. Pe´rez, Acoustic scene classification: an evaluation of an extremely compact feature representation, in: Detection and Classification of Acoustic Scenes and Events, 2016.
[30] Y. Li, M. Yang, Z. Zhang, A survey of multi-view representation learning, IEEE transactions on knowledge and data engineering 31 (10) (2018) 1863–1883.
[31] C. Zhang, G. Wen, Z. Lin, N. Yao, Z. Shang, C. Zhong, An effective bag- of-visual-word scheme for object recognition, in: 2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), IEEE, 2016, pp. 417–421.
[32] J. O. Smith, Mathematics of the discrete Fourier transform (DFT): with audio applications, Julius Smith, 2007.
[33] D. G. Lowe, et al., Object recognition from local scale-invariant features., in: iccv, Vol. 99, 1999, pp. 1150–1157.
[34] H. Bay, T. Tuytelaars, L. Van Gool, Surf: Speeded up robust features, in: European conference on computer vision, Springer, 2006, pp. 404–417.
[35] P. Stephanie, A. Murat, Bag-of-audio-words approach for multimedia event classification, in: Proc. of Interspeech, 2012.
[36] S. Pancoast, M. Akbacak, N-gram extension for bag-of-audio-words, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2013, pp. 778–782.
[37] Z. Hanyu, L. Shengchen, A system for dcase challenge using 2018 crnn with mel features, Tech. rep., DCASE2018 Challenge (September 2018).
[38] K. J. Piczak, Esc: Dataset for environmental sound classification, in: Proceedings of the 23rd ACM international conference on Multimedia, ACM, 2015, pp. 1015–1018.

[39] A. Mesaros, A. Diment, B. Elizalde, T. Heittola, E. Vincent, B. Raj, T. Virtanen, Sound event detection in the DCASE 2017 challenge, IEEE/ACM Transactions on Audio, Speech, and Language ProcessingIn press (2019). doi:10.1109/TASLP.2019.2907016.
[40] K. J. Piczak, Environmental sound classification with convolutional neu- ral networks, in: 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP), IEEE, 2015, pp. 1–6.
[41] T. Komatsu, T. Toizumi, R. Kondo, Y. Senda, Acoustic event detection method using semi-supervised non-negative matrix factorization with a mixture of local dictionaries, in: Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016), 2016, pp. 45–49.
[42] P. Chandrasekhar, S. V. Gangashetty, Acoustic scene classification using deep neural network, Tech. rep., DCASE2017 Challenge (September 2017).