Signal, Image and Video Processing (2021) 15:1211–1219
https://doi.org/10.1007/s11760-020-01851-9

ORIGINAL PAPER

Multi-view representation for sound event recognition

S. Chandrakala[1] · Venkatraman M[1] · Shreyas N[1] · Jayalakshmi S L[2]

Received: 6 August 2020 / Revised: 29 October 2020 / Accepted: 31 December 2020 / Published online: 23 January 2021

Abstract

The sound event recognition (SER) task is gaining lot of importance in emerging applications such as machine audition, audio surveillance, and environmental audio scene recognition. The recognition of sound events with noisy conditions in real-time surveillance applications is a difficult task. In this paper, we focus on learning patterns using multiple forms (views) of the given sound events. We propose two variants of the Multi-View Representation (MVR)-based approach for the SER task. The first variant combines the auditory image-based features and the cepstral features from sound signal. The second variant combines the statistical features extracted from the auditory images and the cepstral features of sound signal. In addition to these variants, Constant Q-transform and Variable Q-transform image-based features are also explored to study the other effective forms of multi-view representations. A discriminative model-based classifier is then used to recognize these representations as environmental sound events. The performance of the proposed MVR approaches is evaluated on three benchmark sound event datasets namely ESC-50, DCASE2016 Task 2, and DCASE2018 Task 2 for the SER task. The recognition accuracy of the proposed MVR approach is significantly better than the other approaches proposed in the recent literature.

Keywords Sound event recognition (SER) · Spectrograms · Mel-frequency cepstral coefficients (MFCCs) · Histogram of oriented gradients (HOG) · Moment-based features · Constant Q-transform (CQT) · Variable Q-transform (VQT) · Support vector machine (SVM)

1 Introduction

Over the past few years, many researchers have been working on developing sound-based surveillance tools to automatically detect environmental sounds [1–4]. Developing sound surveillance system is a popular research field due to its potential benefits in both public and private environments. Recently, some efforts have been directed toward systems capable of detecting and classifying these sounds [5]. Environmental sounds exist in domestic, business, and out door environments. Most of the investigations concentrate on a restricted domain. For example, a system capable of recognizing sounds in a specific indoor environment may be of great importance for monitoring and security applications [4,6]. These functionalities can also be used in portable assistive devices to alert disabled and elderly persons with hearing impairment about specific sounds such as doorbells, alarm signals, etc.

Recently, sound event recognition (SER) has gained significant interest due to its wide applications in the field of multimedia context analysis and automated audio surveillance [2,3]. In the case of automated surveillance, audio sensors play a vital role during night-times when compared to video cameras [7,8]. SER is important to detect the environmental sounds as abnormal or normal events. For instance, door slam, knock, laughter, and coughing sounds are grouped into normal sounds and suspicious events such as glass breaking and screaming are considered as abnormal sounds. SER task also helps in recognizing the context or environment for robots and smart cars [9,10].

✉ S. Chandrakala
sckala@gmail.com; chandrakala@cse.sastra.edu

Venkatraman M
arvimani@gmail.com

Shreyas N
shreyasn9814@gmail.com

Jayalakshmi S L
jayalakshmi.sl@vit.ac.in

[1] Intelligent Systems Group, SASTRA University, Thanjavur, Tamil Nadu, India

[2] School of Computer Science and Engineering, VIT University (Chennai Campus), Chennai, Tamil Nadu, India

Some of the challenges related to SER include the following: the existence of multiple sound sources or overlapping or polyphonic events; recognition of confusable sound events; and lack of compact representation techniques for sound events. These challenges increase the complexity of learning acoustic events and complicates the real-time automated surveillance systems. Recently, SER systems focus on learning representations that can accurately capture the characteristics of a given sound event. Various audio features have been proposed for sound event recognition tasks in different applications.

The most widely used handcrafted sound features are Mel-frequency Cepstral coefficients (MFCCs). On the other hand, the spectrogram-based visual features are extracted by transforming sound signal into its two-dimensional Time-Frequency representation. Some of the visual features used for the recognition task are Scale Invariant Feature Transform (SIFT), Speeded Up Robust Features (SURF), Histogram of oriented Gradients (HOG), and Local Binary Pattern (LBP). Previous studies have analyzed the performance of audio and visual features in automatic speech recognition (ASR) [11] and Music Information Retrieval (MIR) [12] tasks.

In this paper, we focus on forming a multi-view representation by combining visual features extracted from spectrograms with the well-known MFCC features. Indeed, MFCCs essentially capture nonlinear information from the power spectrum of the signal. HOG-based features are invariant to small time and frequency translations. They include local direction of variation of power spectrum, which is not provided by MFCC [13,14]. Hence, we propose multi-view representation that combines the advantages of both MFCC and HOG features. In this work, we propose two variants of the Multi-View Representation. The first variant combines HOG features extracted from spectrogram with cepstral features such as MFCCs. Another variant combines statistical features computed from spectrogram and the MFCCs extracted from sound signal. Multi-view representations are then fed as input to discriminative classifier such as Support Vector Machines (SVM) to recognize the given sound signal.

The rest of this paper is organized as follows. Related work on sound event representations is briefly presented in Sect. 2. In Sect. 3, we describe the proposed Multi-view Representation approaches for the SER task. In Sect. 4, we present experimental studies and discussion.

2 Related work

In the recent studies, there have been a lot of work done in feature learning and recognition of sound events [15–21]. Cakir et al. [15] studied three types of features namely Mel-frequency Cepstral coefficients(MFCCs), Mel-band energies, and log Mel-band energies. Kim and Kim et al. [22] proposed a segmental 2-D MFCCs which rely on transformation of cosines. Lim et al. [23] proposed a bag-of-audio words approach in order to recognize the universal characteristics of an environmental sound event. Eronen et al. [24] used a method which proposes the combination of frequency-domain features and time-domain features such as Zero Crossing Rate (ZCR), spectral centroid, spectral roll-off, short-time average energy, MFCCs, and linear prediction coefficients to classify 24 contexts with the use of hidden Markov models (HMMs). Chu et al. [25] proposed an idea of Matching Pursuit (MP) to obtain effective time and frequency-based features. Then, MP-based feature was combined with MFCC-based features for the acoustic event recognition. Ye et al. [26] has incorporated the local statistics such as mean and standard deviation on local pixels to establish a robust Local Binary Pattern (LBP). Besides, the L2-Hellinger normalization method was applied to the proposed features to further increase the robustness and the discriminative power.

The choice of compact representation influences the outcome of any learning tasks. Most of the recent methods convert the segments of acoustic signals into spectrograms. Spectrogram is a visual time-frequency representation of a 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) [14,27]. Few other methods involve the extraction of sound features such as Mel Frequency Cepstral Coefficients, Constant-Q chromagram, and Spectral flatness directly from raw sound signals. Then, the extracted feature vectors are used as sound event descriptors to train the model.

Feroze et al. [28] 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. Jayalakshmi et al. [6] proposed an approach based on statistical moments computed from MFCC features with Support Vector Machine (SVM) classifier. This 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 DCASE2016 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 Li et al. [29] 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 learnt using deep neural networks. Finally, a combination of

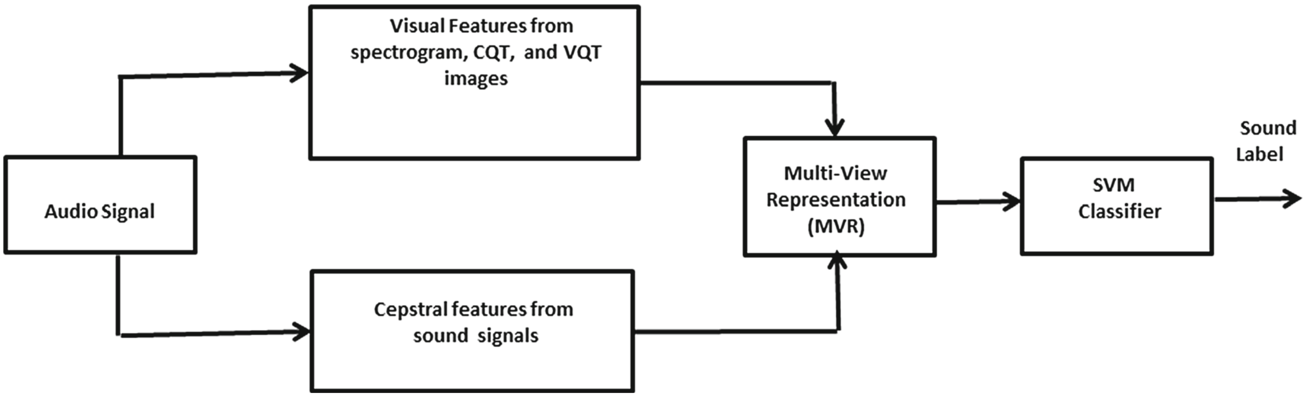
Audio Signal
Visual Features from spectrogram, CQT, and VQT images
Cepstral features from sound signals
Multi-View Representation (MVR)
SVM Classifier
Sound Label

Fig. 1 Block diagram of proposed multi-view representation (MVR) approach for sound event recognition (SER) task

LSTM and Bi-Directional Recurrent Neural Networks was used for classification. This showed moderate performance improvement over existing systems. Another deep network was built using a combination of Convolutional Neural Network (CNN) and a Recurrent Neural Network (RNN) [30]. CNN's have shown to be robust to local and temporal spectral variations and capable of extracting high-level features. RNN's can learn longer-term temporal context, are combined to form a C-RNN network for the task of Polyphonic sound event recognition. Recently, Yu et al. [31] proposed a system to classify the audio events through EEG signals by monitoring the brain activity of participants. In this work, we focus on sound event recognition by using complementary data present in two different modalities of sound signals.

3 Multi-view representation for sound event recognition

The performance of machine learning approaches is predominately dependent on the compact data representation. Effective recognition of sound events depends significantly on the representations derived from sound samples. But these techniques lack in capturing the significant discriminative patterns of sound classes in unconstrained environments because contents of sound signals highly depend on the context of an environmental scene. Therefore, we focus on combining capabilities from multiple modes (views) of input that will complement each other and can be generalized for effective representation. Generally, the multi-view representation aims to combine the multiple views into a single and compact representation to exploit the complementary knowledge contained in multiple views to comprehensively represent the data.

This paper aims to utilize the multiple views of sound data from sound events and to propose a new MVR-based system that yields better results when given as an input to traditional shallow models instead of data-hungry deep models. We propose two variants of the Multi-View Representation (MVR)-based approach for the SER task. The first approach uses the auditory image-based visual features extracted from spectrograms as one form of input and cepstral features extracted from sound signals as other form of input. Second approach uses the auditory image-based statistical features and the cepstral features of sound signal. In addition to these two variants, we have also explored Constant Q-transform (CQT) and Variable-Q Transform (VQT)-image-based visual features for multi view representations. The block diagram of the proposed approach is given in Fig. 1.

3.1 Visual features from auditory images

Spectrogram contains the spectrum of frequencies of a signal varying with time. These spectrograms are used extensively in the field of music and speech processing. Spectrogram is depicted as an image which consists of intensity shown by varying color and brightness.

Figures 2, 3 and 4 illustrate the spectrograms of 'Keyboard','Cough', and 'Laugh' sound event classes, respectively, from the DCASE2016 Task 2 dataset. It can be observed from Fig. 2 that, the spectrograms generated for sound signals of same sound class do not vary much and looks similar. Whereas for 'Cough', and 'Laugh' sounds which are acoustically similar but different sound classes, the corresponding spectrograms are dissimilar as shown in Figs. 3 and 4. Though these two sound classes sound similar (overlapping) the corresponding spectrograms show the discrimination. This helps to reduce the overlap between two different but overlapping sound classes leading to improved performance.

We explore two types of features namely Histogram of Oriented Gradient (HOG) and statistical features. Generally, several key points can be extracted from the spectrogram images. They provide the temporal analysis of the sound event thus giving an auditory image that is easier to interpret. One of the visual feature descriptor popularly used in

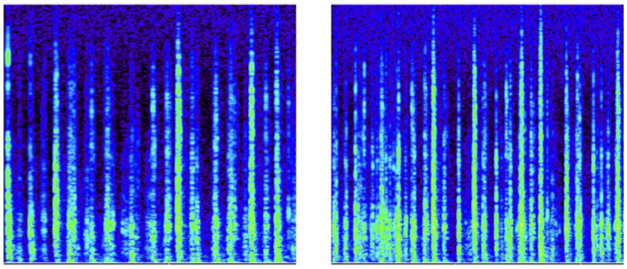

Fig. 2 Spectrograms of two samples of 'Keyboard' sound Class from DCASE2016 Task 2 dataset

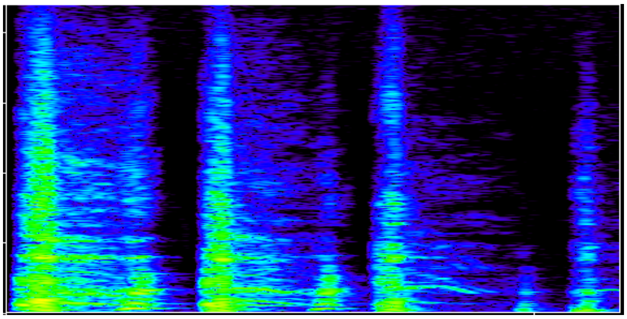

Fig. 3 Spectrogram of 'Cough' sound

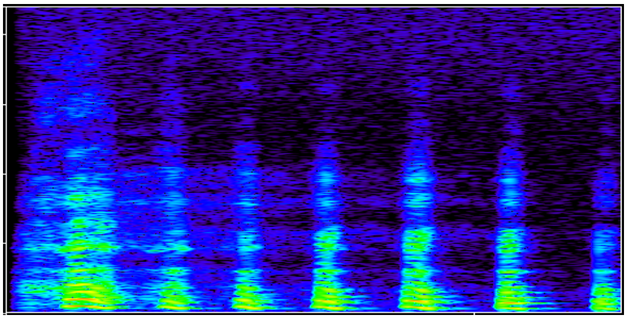

Fig. 4 Spectrogram of 'Laugh' sound

computer vision and object recognition is the histogram of oriented gradients (HOG). This method counts occurrences of gradient orientation in localized portions of an image. It describes the local characteristics of a given spectrogram image in the gradient directions. HOG can be used to capture the modulation of the scene along the temporal axis. The steps involved in computing HOG are listed as follows: (1) The gradient of either spectrogram or constant Q-transform representation is calculated; (2) The angles of all pixel gradients are calculated; (3) Non-overlapping cells of the images are formed; (4) Each cell histogram obtained is normalized based on the histogram of its neighbors. Finally, filtering and pooling are performed to optimize the HOG descriptor.

Apart from spectrograms, other auditory images generated using Constant Q Transform (CQT) and Variable Q Transform (VQT) have also been analyzed. The CQT is given by the formula mentioned below[32] :

$$x[k] = \frac{1}{L[k]} \sum_{n=0}^{L[k]-1} S[k,n] s[n] e^{-j2\pi \frac{Qn}{L[k]}} \tag{1}$$

where 2π Qn/l[k] gives the frequency of the kth component and s[n] is the sample of the digitized time-frequency. The S[k,n] represents the window function that depends on k as well as n. Similarly, Variable Q Transform (VQT) has been implemented that is similar to CQT but allows for different filterbanks to be used in each downsampled octave.

3.2 Cepstral features from sound signals

Handcrafted feature extraction methods have proved to be effective for tasks such as object recognition and audio tagging. Visual features have proved to work better in unconstrained and generalized environments. From this, it can be inferred that some generic audio and visual feature extraction techniques complement each other based on their characteristics. In the cepstral feature extraction, the N-dimensional Mel-frequency Cepstral Coefficient (MFCC) features are extracted from a short-term power spectrum of sound with a Mel-frequency scale[33]. MFCC features are proved to be effective in many of the sound-related surveillance applications [3,12].

In this work, we propose an efficient multi-view representation (MVR) for sound events as shown in Fig. 1. Proposed Multi-View Representations combine both handcrafted cepstral features and auditory image-based visual features. The multi-view representations involve the combination of features that are complementary to each other. Here we propose two such combinations based on auditory images of sound. The first approach combines the statistical moment-based features with MFCC features. This approach involves the computation of moments from pixel values along either of the axes and using those moments as a feature vector. Statistical features are computed from pixel values along either of the axis. We have experimented using both the axes (X-axis as well as Y-axis) to find which axis gives a better result. It was evident that along the y-axis the results were better compared to the x-axis. In this approach, parameters such as skewness, mean, median, variance, minimum, maximum, and kurtosis are calculated for each column along X axis or each row along Y axis. Different sound events often have different sound properties, which lead to the change of the texture pattern of the auditory images and alters the image intensity. The multi-view representation is then formed by combining the statistical moment-based features with MFCC features.

Another variant of the MVR-based approach combines the HOG feature with MFCC. The Multi view representations of sound signals are given as input to the Support Vector Machine (SVM) to recognize the sound events. In addition to the spectrogram images, we have also experimented with other images formed using Constant Q transform and variable-Q transform (VQT) coefficients for further analysis.

4 Experimental studies and discussions

4.1 Datasets used for studies

We have used sound events from the following three different datasets, namely the Environment Sound Classification-10 (ESC-50) dataset [12], DCASE2016 Task 2 [34] , and DCASE2018 Task 2 dataset [35]. These sounds are recorded in different environments with different subjects and noise levels. Each dataset contains various types of sounds recorded in different scenarios with different noise levels.

ESC-50 Dataset This dataset consists of 2000 labeled environmental sound event examples. It contains 50 classes with 40 instances per class [12]. The data are 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. It consists of animal sounds such as dog barking, the sound made by cows, frogs, etc., outdoor sound events such as rain, sea waves, fire crackling, etc., human sounds such as snoring, clapping, snoring, etc., indoor sounds such as vacuum cleaner, washing machine, alarm clock, etc., other sounds such as vehicles, church bell, hand saw, etc. 5-fold cross validation is carried out to evaluate the proposed approach.

DCASE2016 Task 2 Dataset This dataset is provided for sound event recognition in synthetic audio [34]. The audio dataset consists of isolated sound events for 11 different sound event classes with 20 samples per class related to office environment: clearing throat, coughing, door knock, door slam, drawer, human laughter, keyboard, keys (placed on a table), page-turning, the phone ringing, and speech. We used all the data in DCASE2016 Task 2 and carried out five-fold cross-validation.

DCASE2018 Task 2 dataset This dataset contains 41 sound classes [35]. We used training data of 41 classes of sound events for our experimental studies. Some of the sound events in the dataset are shattering, fireworks, keyboard sound, keys jingling, etc. The sound events for this training dataset are composed of 9473 examples. Fivefold cross-validation is carried out to evaluate the performance of the proposed approach.

4.2 Feature extraction

To capture the characteristics of sound events, we have used 26-dimensional MFCC features as sound features. For each sound event, we computed the statistical features across the 26-dimensional MFCC features. Mean, median, minimum, maximum, variance, skewness, and kurtosis were computed as statistical features. These statistical features, computed across 26-dimensional features, were then concatenated to form the fixed dimensional representation (26*7). Similarly, for all the examples, the spectrogram-based visual feature HOG is extracted. Besides, from each spectrogram, we also extract seven moment-based statistical features such as mean, median, minimum, maximum, standard deviation, skewness, and kurtosis. These features were combined to form a fixed dimensional vector for each spectral image. CQT and VQT image-based feature extraction are also performed on the given input sound signal. After the feature extraction step, the extracted visual features were combined with sound features to form a multi-view representation for effective learning.

4.3 Performance analysis

In all experiments, MFCC features are used as the basic sound features. The performance of proposed approach and other conventional approaches are shown in Tables 1, 2 and 3 for the datasets ESC-50, DCASE2016 Task 2, and DCASE2018 Task 2, respectively. The first method is spectrogram with HOG-based approach as single-view representation. In HOG computation, cell size is fixed as 16*16, number of orientations as 8 and pooling operator as average operator(pooling over frequency) as given in [39]. The second method uses statistical features in two ways: i) statistical features computed along pixel values of every row (Spectrogram+moments (X axis) + MFCC + SVM) and ii) statistical features computed along pixel values of every column (Spectrogram + moments (Y axis) + MFCC + SVM).

The performance of the proposed Spectrogram + HOG + MFCC-based multi-view representation approach gives better results compared to spectrogram with HOG-based approach. The proposed approach gives an accuracy of 72.7 %, 91.3 %,and 80.17% for the datasets ESC-50, DCASE2016 Task 2, and DCASE2018 Task 2, respectively. Even though the size of the dataset increases, the method is consistent in its performance compared to single-view representations. The second approach is based on statistical moments that achieved an accuracy of 68.9 %, 83%, and 80.17% for the datasets ESC-50, DCASE2016 Task 2, and DCASE2018 Task 2, respectively. Here we can observe that the proposed approaches work better when compared to other conventional approaches such as CQT- and VQT-based representations.

In the case of the ESC-50 dataset, Table 1 shows the baseline performance reported in [12]. Piczak [36] evaluated the

Table 1 Comparison of recognition accuracy (%) for ESC-50 dataset

Method	Accuracy (%)
ZCR + MFCC + SVM [12]	39.6
ZCR + MFCC + Random forest ensembler[12]	44.3
Log Mel scaled spectrogram +CNN [36]	64.45
MFCC + SVM	61.2
CQT + HOG + SVM	65.1
VQT + HOG + SVM	65.9
Spectrogram + moments (X axis)+ SVM	65.4
Spectrogram + moments (Y axis) + SVM	67.4
Spectrogram + Hog + SVM	71.9
CQT + HOG + MFCC + SVM	67.4
VQT + HOG + MFCC + SVM	67.7
Spectrogram+moments (X axis) + MFCC + SVM	66.1
Spectrogram+moments (Y axis) + MFCC + SVM	68.9
Spectrogram + HOG + MFCC + SVM	72.7

Table 2 Comparison of recognition accuracy (%) for DCASE2016 Task 2 dataset

Method	Accuracy (%)
Spectral template + NMF [37]	41.6
CQT + Recurrent neural networks (RNN) [34]	52.8
Gammatone cepstrum + Random forests [34]	64.8
Mel-filter bank + BLSTM [34]	78.1
Mel energy + Deep Neural Networks (DNN) [34]	78.7
NMF + MLD [38]	80.2
CQT + HOG + SVM	76.3
VQT + HOG + SVM	77.1
Spectrogram + moments (X axis) + SVM	73.9
Spectrogram + moments (Y axis) + SVM	79.0
Spectrogram + HOG + SVM	90.0
CQT + HOG + MFCC + SVM	78.1
VQT + HOG + MFCC + SVM	79.9
Spectrogram + moments (X axis) + MFCC + SVM	76.3
Spectrogram + moments (Y axis) + MFCC + SVM	83.0
Spectrogram + HOG + MFCC + SVM	91.3

Table 3 Comparison of recognition accuracy (%) for DCASE2018 Task 2 dataset

Method	Accuracy (%)
CQT + HOG + SVM	61.1
VQT + HOG + SVM	61.1
Spectrogram + moments (X axis) + SVM	63.2
Spectrogram + moments (Y axis) + SVM	66.0
Spectrogram + HOG + SVM	66.9
CQT + HOG + MFCC + SVM	62.9
VQT + HOG + MFCC + SVM	63.3
Spectrogram+moments (X axis) + MFCC + SVM	63.3
Spectrogram+moments (Y axis) + MFCC + SVM	68.1
Spectrogram + HOG+ MFCC + SVM	80.17

Table 4 Confusion matrix of single-view representation-based approach (Spectrogram + HOG + SVM) for DCASE2016 Task 2 dataset

Sound classes	Clearing throat	Coughing	Door knock	Door slam	Drawer	Human laughter	Keyboard	Keys	Page turning	Phone ringing	Speech
Clearing throat	**86.36**	0	0	4.54	0	0	0	0	4.54	0	4.54
Coughing	9.1	**95.45**	0	0	0	0	0	4.54	9.1	0	0
Door knock	0	0	**95.45**	0	4.54	0	9.1	0	0	0	0
Door slam	0	4.54	0	**90.9**	0	0	0	4.54	0	0	0
Drawer	0	0	0	0	**90.9**	0	0	0	0	0	0
Human laughter	0	0	0	0	0	**90.9**	0	0	9.1	0	0
Keyboard	0	0	4.54	0	0	0	**90.9**	0	0	0	0
Keys	4.54	0	0	0	4.54	0	0	**90.9**	0	0	0
Page turning	0	0	0	0	0	4.54	0	0	**77.27**	9.1	0
Phone ringing	0	0	0	0	0	4.54	0	0	0	**90.9**	4.54
speech	0	0	0	4.54	0	0	0	0	0	0	**90.9**

The best results are highlighted in bold

Table 5 Confusion matrix of multi-view representation-based approach (Spectrogram + HOG + MFCC + SVM) for DCASE2016 Task 2 dataset

Sound classes	Clearing throat	Coughing	Door knock	Door slam	Drawer	Human laughter	Keyboard	Keys	Page turning	Phone ringing	Speech
Clearing throat	**86.36**	0	0	4.54	0	0	0	0	4.54	0	4.54
Coughing	9.1	**95.45**	0	0	0	0	0	4.54	9.1	0	0
Door knock	0	0	**95.45**	0	4.54	0	9.1	0	0	0	0
Door slam	0	4.54	0	**90.9**	0	0	0	4.54	0	0	0
Drawer	0	0	0	0	**90.9**	0	0	0	0	0	0
Human laughter	0	0	0	0	0	**95.45**	0	0	9.1	0	0
Keyboard	0	0	4.54	0	0	0	**90.9**	0	0	0	0
Keys	4.54	0	0	0	4.54	0	0	**90.9**	0	0	0
Page turning	0	0	0	0	0	0	0	0	**77.27**	0	0
Phone ringing	0	0	0	0	0	4.54	0	0	0	**100**	4.54
Speech	0	0	0	4.54	0	0	0	0	0	0	**90.9**

The best results are highlighted in bold potential of convolutional neural networks with log-scaled Mel-spectrograms for recognizing the short duration environmental sounds. The proposed MVR approach produced a much better performance with an accuracy of 33 % improvement when compared to the baseline system reported in [12].

Similarly, Table 2 shows the recognition accuracy of the proposed approach and some of the state-of-the-art methods reported in the literature for the DCASE2016 Task 2 sound event dataset. The proposed approach outperforms the baseline system by giving a 50% increase in classification accuracy. Table 2 adds some of the systems submitted in the DCASE2016 Task 2 challenge for sound event recognition [34]. The baseline system [37] of DCASE2016 Task 2 uses a dictionary of spectral template with supervised NMF approach. The proposed MVR approach outperformed the following systems in the DCASE2016 Task 2 challenge: representations of constant-Q transform (CQT) with RNN classifier, Gammatone cepstrum 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 handcrafted features with shallow models trained on meaningful multi-view representations rather than using data-hungry deep feature learning techniques.

The results of single view and multi-view representations studied for three datasets are analyzed with the help of confusion matrices. As an example, in case of DCASE2016 Task 2 dataset, the proposed HOG+MFCC-based MVR approach reduces the overlap between classes such as {*Human laughter, Page turning*}, and {*Phone ringing, Page turning*} which can be seen in Tables 4 and 5. This leads to slightly improved performance compared to single view HOG-based approach.

Table 3 shows the recognition accuracy of the proposed MVR-based approaches and single-view approaches. There are 11 sound event classes in DCASE2016 Task 2 dataset and 41 sound event classes in DCASE2018 Task 2 dataset. As the number of sound classes increases the MVR using HOG+MFCC approach outperforms single-view with HOG only approach as given in Table 3.

5 Conclusion

In this paper, a Multi-View Representation (MVR)-based approach for Sound Event Recognition has been proposed. The proposed approach combines auditory image-based visual features with cepstral features to form compact and effective representation. The proposed handcrafted feature-based MVR representation with simple shallow model such as SVM as classifier leads to improved performance over other state-of-the-art methods for ESC-50, DCASE2016 Task 2, and DCASE2018 Task 2 datasets. The proposed approach is more suitable for recognizing acoustically similar but different sound classes.

Acknowledgements The authors would like to acknowledge the financial support vide No.DST/CSRI/2017/131(G) Project under the 'Cognitive Science Research Initiative (CSRI)' by the Department of Science and Technology, Government of India to carry out this work.

Code availability The code is available from the corresponding author upon request.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Yang, W., Krishnan, S.: Sound event detection in real-life audio using joint spectral and temporal features. Signal Image Video Process. **12**(7), 1345–1352 (2018)
2. Kong, Q., Xu, Y., Sobieraj, I., Wang, W., Plumbley, D.M.: Sound event detection and time-frequency segmentation from weakly labelled data. IEEE/ACM Trans. Audio Speech Lang. Process. (TASLP) **27**(4), 777–787 (2019)
3. Chandrakala, S., Jayalakshmi, S.L.: Generative model driven representation learning in a hybrid framework for environmental audio scene and sound event recognition. IEEE Trans. Multimed. **22**(1), 3–14 (2020)
4. Shreyas, N., Venkatraman, M., Malini, S., Chandrakala, S.: 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, pp. 95–106. Elsevier, (2020)
5. Wang, C.-Y., Tai, T.-C., Wang, J.-C., Santoso, A., Mathulaprangsan, S., Chiang, C.-C., Chung-Hsien, W.: Sound events recognition and retrieval using multi-convolutional-channel sparse coding convolutional neural networks. IEEE/ACM Trans. Audio Speech Lang. Process. **28**, 1875–1887 (2020)
6. Jayalakshmi, S.L., Chandrakala, S., Nedunchelian, R.: Global statistical features-based approach for acoustic event detection. Appl. Acoust. **139**, 113–118 (2018)
7. Atrey, P.K., Maddage, N.C., Kankanhalli, M.S.: Audio based event detection for multimedia surveillance. In: 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, vol. 5, p. V. IEEE, (2006)
8. Dennis, J., Tran, H.D., Li, H.: Spectrogram image feature for sound event classification in mismatched conditions. IEEE Signal Process. Lett. **18**(2), 130–133 (2010)
9. Do Ha, M., Sheng, W., Liu, M., Zhang, S.: Context-aware sound event recognition for home service robots. In: 2016 IEEE International Conference on Automation Science and Engineering (CASE), pp. 739–744. IEEE, (2016)
10. Singh, S., Payne, R.S., Jennings, A.P.: Toward a methodology for assessing electric vehicle exterior sounds. IEEE Trans. Intell. Transp. Syst. **15**(4), 1790–1800 (2014)
11. Graves, A., Mohamed, A., Hinton, G.: Speech recognition with deep recurrent neural networks. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 6645–6649. IEEE, (2013)

12. Piczak, K.J.: ESC: dataset for environmental sound classification. In: Proceedings of the 23rd ACM International Conference on Multimedia, pp. 1015–1018. ACM, (2015)
13. Rakotomamonjy, A., Gasso, G.: Histogram of gradients of time-frequency representations for audio scene classification. IEEE/ACM Trans. Audio Speech Lang. Process. **23**(1), 142–153 (2014)
14. Cowling, M., Sitte, R.: Comparison of techniques for environmental sound recognition. Pattern Recognit. Lett. **24**(15), 2895–2907 (2003)
15. Cakir, E., Heittola, T., Huttunen, H., Virtanen, T.: Polyphonic sound event detection using multi label deep neural networks. In: 2015 International Joint Conference on Neural Networks (IJCNN), pp. 1–7. IEEE, (2015)
16. Jeong, I.-Y., Lee, S., Han, Y., Lee, K.: Audio event detection using multiple-input convolutional neural network. In: Detection and Classification of Acoustic Scenes and Events (DCASE) (2017)
17. Chen, Y., Zhang, Y., Duan, Z.: DCASE2017 sound event detection using convolutional neural network. In: Detection and Classification of Acoustic Scenes and Events (2017)
18. Adavanne, S., Parascandolo, G., Pertilä, P., Heittola, T., Virtanen, T.: Sound event detection in multichannel audio using spatial and harmonic features. *arXiv preprint* arXiv:1706.02293, (2017)
19. Parascandolo, G., Huttunen, H., Virtanen, T.: Recurrent neural networks for polyphonic sound event detection in real life recordings. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6440–6444. IEEE, (2016)
20. Lu, R., Duan, Z.: Bidirectional GRU for sound event detection. In: Detection and Classification of Acoustic Scenes and Events (2017)
21. Zhou, J.: Sound event detection in multichannel audio LSTM network. In: Proceedings of Detection Classification Acoustic Scenes Events, (2017)
22. Myung Jong Kim and Hoirin Kim: Audio-based objectionable content detection using discriminative transforms of time–frequency dynamics. IEEE Trans. Multimed. **14**(5), 1390–1400 (2012)
23. Hyungjun, L., Kim, M.J., Kim, H.-R.: Bag-of-audio-words feature representation using GMM clustering for sound event classification. In: *ICEIC2015*, pp. 170–175, (2015)
24. Heittola, T., Mesaros, A., Eronen, A., Virtanen, T.: Context-dependent sound event detection. EURASIP J. Audio Speech Music Process. **2013**(1), 1 (2013)
25. Chu, S., Narayanan, S., Jay Kuo, C.-C.: Environmental sound recognition with time–frequency audio features. IEEE Trans. Audio Speech Lang. Process. **17**(6), 1142–1158 (2009)
26. Ye, J., Kobayashi, T., Wang, X., Tsuda, H., Masahiro, M.: An automatic taxonomy approach. In: IEEE Transactions on Emerging Topics in Computing, Audio Data Mining for Anthropogenic Disaster Identification (2017)
27. Serizel, R., Bisot, V., Essid, S., Richard, G.: Acoustic features for environmental sound analysis. In: Computational Analysis of Sound Scenes and Events, pp. 71–101. Springer, (2018)
28. Grzeszick, R., Plinge, A., Fink, G.A.: Bag-of-features methods for acoustic event detection and classification. IEEE/ACM Trans. Audio Speech Lang. Process. **25**(6), 1242–1252 (2017)
29. Li, Y., Li, X., Zhang, Y., Wang, W., Liu, M., Feng, X.: Acoustic scene classification using deep audio feature and BLSTM network. In: 2018 International Conference on Audio, Language and Image Processing (ICALIP), pp. 371–374. IEEE, (2018)
30. Vesperini, F., Gabrielli, L., Principi, E., Squartini, S.: Polyphonic sound event detection by using capsule neural networks. IEEE J. Sel. Top. Signal Process. **13**(2), 310–322 (2019). https://doi.org/10.1109/JSTSP.2019.2902305
31. Yu, Y., Beuret, S., Zeng, D., Oyama, K.: Deep learning of human perception in audio event classification. In: 2018 IEEE International Symposium on Multimedia (ISM), pp. 188–189. IEEE, (2018)
32. Brown, J.C.: Calculation of a constant q spectral transform. J. Acoust. Soc. Am. **89**(1), 425–434 (1991)
33. Hanyu, Z., Shengchen, L.: A system for DCASE challenge using 2018 CRNN with MEL features. Technical report, DCASE2018 Challenge (2018)
34. Mesaros, A., Heittola, T., Benetos, E., Foster, P., Lagrange, M., Virtanen, T.: Detection and classification of acoustic scenes and events: outcome of the DCASE 2016 challenge. IEEE/ACM Trans. Audio Speech Lang. Process. (TASLP) **26**(2), 379–393 (2018)
35. Fonseca, E., Plakal, M., Font, F., Ellis, D.P.W., Favory, X., Pons, J., Serra, X.: General-purpose tagging of freesound audio with audioset labels: task description, dataset, and baseline. *arXiv preprint* arXiv:1807.09902, (2018)
36. Piczak, K.J.: Environmental sound classification with convolutional neural networks. In: 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP), pp. 1–6. IEEE, (2015b)
37. Benetos, E., Lafay, G., Lagrange, M.: DCASE2016 task 2 baseline. Technical report, DCASE2016 Challenge (2016)
38. Komatsu, T., Toizumi, T., Kondo, R., Senda, Y.: 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), pp. 45–49, (2016)
39. Rakotomamonjy, A., Gasso, G.: Histogram of gradients of time–frequency representations for audio scene classification. IEEE/ACM Trans. Audio Speech Lang. Process. (TASLP) **23**(1), 142–153 (2015)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.