Learning to Separate: Soundscape Classification using Foreground and Background

Dhanunjaya Varma Devalraju, Padmanabhan Rajan, Dileep A.D.
School of Computing and Electrical Engineering
Indian Institute of Technology, Mandi
s18023@students.iitmandi.ac.in, padman@iitmandi.ac.in, addileep@iitmandi.ac.in

***Abstract*—This paper applies the framework of robust principal components analysis (RPCA) to the problem of classifying acoustic soundscapes. RPCA provides a mechanism to decompose a data matrix as the sum of a low-rank matrix and a sparse matrix. In the context of data representing acoustic soundscapes, the low-rank matrix represents the slow-changing background sound events, and the sparse matrix represents the occasional foreground sound events. The data representations are obtained as feature embeddings from pretrained deep convolutional networks. The paper investigates the effectiveness of classifying acoustic soundscapes by using the foreground or background information alone. Further, by using the subspace projection technique of nuisance attribute projection (NAP), the undesired components from the foreground or background are removed. Our results indicate that RPCA and subspace projections indeed provide benefits in improving discrimination for classifying acoustic soundscapes.**

***Index Terms*—Acoustic scene classification, robust PCA, subspace projections**

I. Introduction

Real-world data in many situations can be seen as occasional events happening in a (relatively) slow-changing environment. For example, such a situation can be in a surveillance video captured by closed-circuit camera. Here, there is a mostly constant background, inter-spaced by occasional movements of people or objects in the video. Another example is the acoustic environment inside a bus. There is the constant sound of the engine (the slow-changing component), interspaced by the door opening or people talking (occasional events). Yet another example is the acoustic environment in a restaurant. There is the constant murmur of diners and the clink of cutlery (the slow-changing part) and the occasional sounds of laughter or more unusual sounds.

For analysing and classifying acoustic soundscapes such as these, it sometimes may be useful to separate the slow-changing background sounds (the sound of the bus engine or the murmur of diners) from the occasional sound events (opening of the bus door or laughter in the restaurant). The framework of *robust principal component analysis* (RPCA) provides a method to perform this separation [1] [2]. RPCA decomposes a set of observations represented by a data matrix M as

$$M = L + S,$$

where L is a low-rank matrix representing the slowly-changing background events, and S is a sparse matrix representing the outliers or occasional events.

This paper examines the task of soundscape classification or acoustic scene classification (ASC) utilising the RPCA framework. The paper also examines if the removal of background and using only the foreground (and vice versa) helps in better classification. Furthermore, by using the subspace projection method of nuisance attribute projection (NAP), it is also explored if partial removal of either can be useful. Individual acoustic soundscapes may be classified into broader categories such as indoor, outdoor or vehicle etc. In this situation, the background may be very different across categories. At the same time, soundscapes within a category may have similar backgrounds, and hence the foreground may be crucial in discriminating between them.

Vaswani et al. [1] [2] in their review papers gave a detailed description about RPCA, the problems in learning the subspace, and various algorithms to solve the low-rank and sparse matrix decomposition. They also provide details about various RPCA applications in computer vision and video analytics, dynamic and functional MRI and detecting anomalies in computer and social networks. Huang et al. [3] applied RPCA to separate singing voice from monaural recordings, where they use the augmented Lagrange multiplier method borrowed from Lin et al. [4] to solve the low-rank and sparse matrix decomposition.

Source separation is a related problem. There are a few works where the authors apply source separation for ASC, though the explicit low-rank and sparse formulation as in RPCA is not applied. Mun et al. [5] trained a recurrent neural network (RNN) for source separation and used mid-layer features from RNN as novel discriminative features. Han et al. [6] proposed a neural network based ensemble model trained with spectrograms generated from binaural audio, background subtraction and harmonic-percussive source separation to achieve better classification accuracy.

In recent years, most of the works in ASC uses time-frequency representations such as log-mel spectrogram [7] [8], constant-Q transform spectrogram [9] and scalogram [10]. These time-frequency representations are treated as images and features (also called embeddings) are extracted using convolution neural networks (CNN) for downstream classification. However, a few methods use raw audio signals for extracting features. Arshdeep et al. [11] used SoundNet, a deep convolution neural network (DCNN) for extracting features

from raw audio signals. In this paper, we explore both time-frequency representation (log-mel spectrogram) as well as raw audio signals for extracting embeddings by deploying pre-trained deep convolution neural networks (DCNN) as feature extractors.

The rest of this paper is organized as follows. Section II describes the RPCA framework in brief. Section III introduces the subspace projection method we employ to manipulate the fineness of the background or foreground supression . Section IV describes the proposed framework. Section V presents the experimental results. We conclude the paper in Section VI.

II. Robust Principal Component analysis

Traditionally, principal component analysis (PCA) is used to learn a representation or basis for a given set of observations. By utilising the significant columns of the PCA matrix, a subspace can be formed where most of the observations lie. But PCA is sensitive to outliers or data corruptions. This situation can arise easily in real-world data. Robust PCA (RPCA) overcomes this issue by assuming that outliers are additive and sparse. RPCA is a convex program that recovers low-rank matrices when a fraction of their entries are corrupted [3]. The low-rank matrix can be recovered by solving the following convex optimization problem:

$$\text{minimize } ||L||_* + \lambda||S||_1, \tag{1}$$

$$\text{subject to } L + S = M, \tag{2}$$

where M, L and $S \in \mathbb{R}^{n1 \times n2}$, $||\cdot||_*$ and $||\cdot||_1$ denote the nuclear norm (sum of singular values) and the L1-norm (sum of absolute values), respectively. $\lambda > 0$ is a trade-off parameter between the rank of L and the sparsity of S. For more details about RPCA and the algorithms used to solve the decomposition please refer to [1].

Fig 1 illustrates the procedure of RPCA applied to an audio signal. The spectrogram representation M of the audio signal is approximated as the spectrograms $L + S$. Here L represents the background sounds and S represents the foreground sounds. Time-domain signals are obtained from L and S by utilising the inverse short-time Fourier transform and the phase of the original audio signal. Spectrograms from an audio example including those of foreground and background are illustrated in Figure 2.

III. Subspace projections

In some situations, the RPCA solution may not give the best desirable separation. For example, in acoustic scenes such as a library or a metro station, complete separation of the background from the foreground may not be possible. In these cases, it may be useful to be able to separate a part of the background or the foreground. One way to achieve this is by using subspace projections as described below.

Nuisance attribute projection (NAP) is a technique used in speaker recognition to compensate channel effects on the speech signal. This is achieved by removing dimensions that are irrelevant to the task [12] [13]. A similar technique was

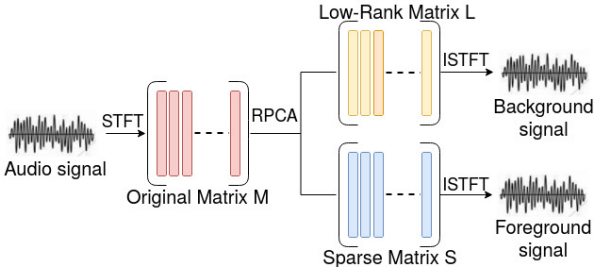


Fig. 1: RPCA based foreground and background separation. The phase of original audio signal is used to reconstruct foreground and background signals [3].

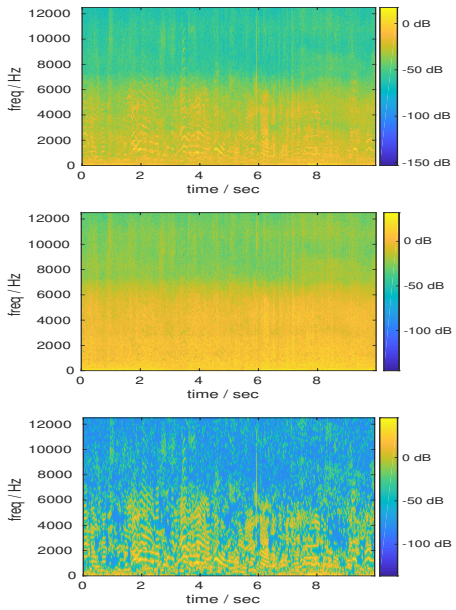


Fig. 2: Mel spectrograms, for an audio example from soundscape "beach", before and after RPCA. Top: The original audio M. Middle: The background L. Bottom: The foreground S. The spectrograms are shown in log scale for better visualization.

applied in face recognition to remove illumination artifacts [14]. NAP is used to remove the unwanted session/nuisance variations from a vector representation of the data item of interest.

NAP attempts to remove the unwanted variations by applying the following transform to the data vector X, to get the nuisance-removed vector $\widetilde{X}$:

$$\widetilde{X} = X - BB^T X = (I - BB^T)X. \tag{3}$$

Here B is an orthogonal basis matrix whose columns span the nuisance space, and I is the identity matrix. B can be estimated by learning a suitable representation from a collection of nuisance data.

In our context, the nuisance factor could be either the background of the acoustic scene or the foreground, depending on what we are trying to suppress when the classification

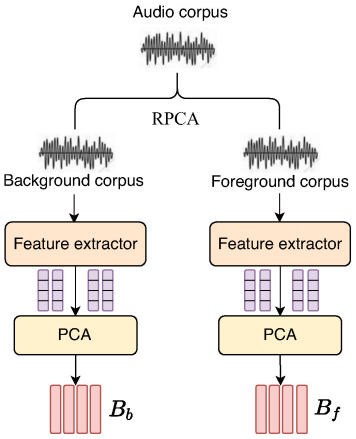


Fig. 3: Illustration of constructing background basis and foreground basis for NAP.

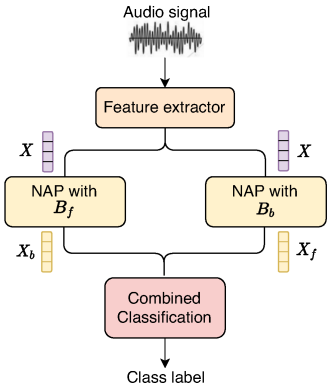


Fig. 4: Proposed framework, using NAP to suppress foreground or background. A combination is used to make the final decision.

is performed. The process of learning the basis matrix is illustrated in Fig 3. Taking enough examples of signals containing only background signals, a basis is learnt using PCA, to give the background basis B_b. A similar process is done for developing the the foreground basis B_f.

IV. The Proposed Framework

The proposed framework is illustrated in Figure 4. After feature extraction, the input audio signal is represented by the vector X. This vector contains information from both the foreground and the background. If we assume that most of the information for classifying an acoustic scene comes from the background, we can consider the foreground sound events as the nuisance attributes. In this case, NAP is performed on X by subtracting the nuisance components of X:

$$X_b = X - B_f B_f^T X \tag{4}$$

Here X_b is the vector that represents the "foreground-removed" representation. Similarly, if we assume that the foreground contains most of the information to classify the scene, we consider the background sound events as the nuisance attributes. NAP is then performed on X by subtracting the nuisance components of X:

$$X_f = X - B_b B_b^T X \tag{5}$$

Here X_f is the "background-removed" representation. By varying the number of components (in other words, columns) in the NAP basis, we can control the amount of foreground or background to be suppressed in X_b or X_f respectively.

In the above equations, B_f represents foreground basis, B_b represents background basis and X represents the feature embeddings for the original (nonseparated) audio signal.

The information from the resulting vectors X_b and X_f can be combined further by using feature fusion or decision fusion. This has the effect of classifying with both the foreground as well as the background, although in separate streams. Moreover, NAP can be used to control the amount of nuisance information being suppressed in either case.

V. Experimental Evaluation

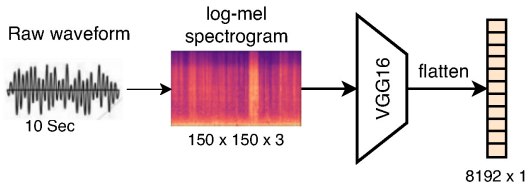
We describe here the experimental evaluation of the proposed method for acoustic scene classification. The primary purpose of the experiments are to determine if separating the slow-changing background and the sparse foreground is effective when compared to not separating them. Hence, our baseline systems extract embeddings from the input audio signal without any separation and classifies them. This is compared with various schemes where the background and foreground are separated, including: (1) individual classification using only background and only foreground, (2) applying NAP, followed by individual classification using foreground and background, and (3) combining the information from foreground and background using feature fusion or decision fusion. We also compare our results with the non-ensemble methods reported in [8] and [6]. The compared system in [8] uses PLDA classifier trained on features extracted from CNN based models. These CNN models are trained on spectrogram image features using log scaled filter-banks. The compared system in [6] uses neural network trained on spectrograms from background subtraction with a median filter.

A. Dataset

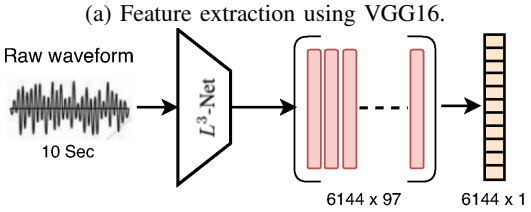
We used DCASE 2017 ASC (task 1) development dataset [15]. This dataset has 15 scenes which are broadly categorized in to three categories namely vehicle (Bus, Car, Tram, Train), outdoor (Urban park, Residential area, Lakeside beach, City center, Forest path) and indoor (Grocery store, Cafe/Restaurant, Home, Metro station, Library, Office). Audio examples are recorded at 44.1 kHz sampling rate with a binaural microphone then the recordings are split into audio segments of length 10 seconds. We train and evaluate our models (4-fold cross-validation) as per the guidelines provided in the DCASE 2017 ASC task 1 [15].

B. Feature extraction and classification

We use two different feature extractors namely VGG16 and L^3-Net to derive embeddings from audio signals. For the embeddings from VGG16, we computed log-mel spectrogram



(a) Feature extraction using VGG16.



(b) Feature extraction using L^3-Net.

Fig. 5: Feature extraction procedure used to derive embeddings from the input audio signal.

for each audio signal using Hann window with window size of about 40 ms and hop length 10 ms. These log-mel spectrograms saved as colour PNG images are input into the VGG16 model [16] as shown in Fig. 5a. We use the same model and weights which are trained on Imagenet dataset and extracted features from the last convolution block. The extracted feature embedding is a column vector of length 8192 after flattening.

Similarly, we used Openl3 library which is an open source implementation of L^3-Net from the authors of [17]. L^3-Net is a pre-trained model trained on the AudioSet dataset. Openl3 library accepts raw audio signal as input, computes mel-spectrogram with 128 bands and returns embedding of length 6144×1 for each audio frame. In our case it accepts raw wave form (10 sec) as input and outputs embedding of size 6144×97 where 97 corresponds to the time dimension. We average the embedding across time to get a column vector of length 6144 as shown in Fig. 5b.

In all cases, our classifier is a simple support vector machine (SVM) trained one-against-one on the input embeddings, using a linear kernel. Since the objective of the paper is to determine the effectiveness of foreground-background separation, we did not perform any fine-tuning of the classifier.

The baseline systems described above also uses the same feature extraction procedure. The results of the two baseline systems are given in Table I. It can be seen that the L^3-Net embeddings perform better than VGG16 embeddings, as it is trained with the large-scale AudioSet dataset, which is fine-tuned with the proper and better design choices for audio signals [17]. All results are in terms of classification accuracy.

C. Effect of separating foreground and background

We now investigate the effect of the separation of foreground and background using RPCA. RPCA is applied in the manner depicted in Fig 1, where the input audio signal is given to the algorithm, and two separate audio signals (one for background and one for foreground) are generated. We utilise the implementation of RPCA in [3], which utilised the phase of the original audio signal to obtain the foreground and background signals. Once the foreground and background signals are obtained, we train separate SVM classifiers for each. The results using only foreground or only background using the two different embedding schemes are also given in Table I. The results seem to indicate that the background is more useful for classification than the foreground.

TABLE I: Results without NAP: baseline (no separation of foreground and background), using only background, using only foreground.

	Baseline	*Background only*	*Foreground only*
VGG16 + SVM	68.05	65.79	55.79
L^3-Net + SVM	84.70	83.23	74.97

D. Effect of NAP

Learning basis for NAP: The nuisance basis for the foreground B_f and the background B_b is learnt as depicted in Fig 3. RPCA is performed on the input audio signal as shown in Fig 1 to generate foreground and background audio signals. Feature extraction is done on these signals using with either VGG16 or L^3Net, and PCA is performed to determine the columns of B_f and B_b. The corpora used for this consists of all examples from all classes in the training set. By varying the number of principal components to use (which form the the columns of B_f or B_b), we control the amount of foreground or background that will be removed while performing NAP with B_f or B_b respectively.

We now examine the effect of partially suppressing the background or foreground using NAP. The results of applying NAP are shown in Fig 6. This demonstrates that there is an improvement in classification accuracy on VGG16 embeddings by 0.65% as a result of suppressing the foreground, and 0.41% as a result of suppressing the background. For L^3-Net embeddings, there is an improvement of 0.38% and 0.17% by suppressing the foreground and background respectively. The number under each bar corresponding to NAP in Fig 6 indicates the size of the basis used for the projection. Only the best performing results are shown. The results also indicate that considering the foreground as the nuisance attribute (the unwanted attribute) leads to higher performance gains in classification, when compared to the background. This again indicates that the background carries more useful information for classification. The methods in [8] and [6] achieve similar results, though they use more sophisticated feature extraction and classification pipelines.

E. Fusion Based Analysis

Since the application of NAP did not deteriorate performance, it motivates the combination of information from the foreground and background. For this, we next apply fusion in both, the embedding space (also termed early fusion) and the decision space (also termed late fusion).

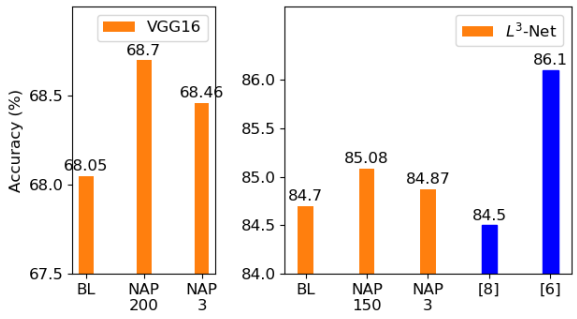


Fig. 6: Classification results after suppression via NAP. The left subplot gives results for VGG16: baseline, after suppressing foreground, after suppressing the background. The right subplot gives the results for L^3-Net: baseline, after suppressing the foreground, after suppressing the background, and results from [8], [6]. The number under NAP gives the number of components in the NAP basis.

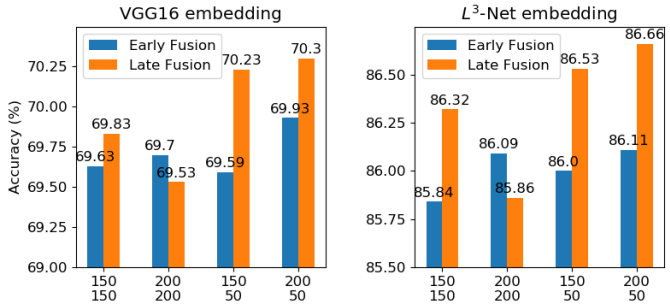


Fig. 7: Results of fusion after NAP. The numbers under each bar gives the number of components in the NAP basis. The number above is for the foreground basis and that below is for the background basis.

The results of applying fusion are shown in Fig. 7. All the results are higher than the respective baselines. The largest improvement obtained for VGG16 embeddings was 2.25% with decision fusion, using 200 components for the foreground basis, and 50 components for the background basis. For L^3-Net embeddings the improvement was 1.96% with decision fusion, using 200 components for the foreground basis and 50 components for the background basis. These are depicted in the last orange bars in both subfigures of Fig 7. These results are close to those achieved in [6].

F. Discussion

The improvements after applying NAP are modest, though promising. The RPCA algorithm does introduce some artifacts in both the generated foreground and background signals. Thus the bases learnt for NAP are probably capturing these artifacts as well. Suppressing the artifacts while learning the NAP basis is part of future investigation. Also, choosing the correct number of components in the NAP basis is crucial.

VI. Conclusion

This paper explored the use of RPCA to separate foreground and background components of an audio signal to perform acoustic scene classification. Our experiments indicate that by using NAP subspace projections, suppression of the foreground components achieve improvements in classification accuracy. They also indicate that the background is generally more useful than the foreground, but the foreground also contains important cues that help in discrimination.

References

[1] N. Vaswani and P. Narayanamurthy, "Static and Dynamic Robust PCA and Matrix Completion: A Review," in Proceedings of the IEEE, vol. 106, no. 8, pp. 1359-1379, Aug. 2018.

[2] N. Vaswani, T. Bouwmans, S. Javed and P. Narayanamurthy, "Robust Subspace Learning: Robust PCA, Robust Subspace Tracking, and Robust Subspace Recovery," in IEEE Signal Processing Magazine, vol. 35, no. 4, pp. 32-55, July 2018.

[3] P. Huang, S. D. Chen, P. Smaragdis and M. Hasegawa-Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, 2012, pp. 57-60.

[4] Z. Lin, M. Chen, L. Wu, and Y. Ma, "The augmented Lagrange multiplier method for exact recovery of corrupted lowrank matrices," Tech. Rep. UILU-ENG-09-2215, UIUC, Nov.2009.

[5] Mun, S., Shon, S., Kim, W., Han, D. K., and Ko, H. (2017). A novel discriminative feature extraction for acoustic scene classification using RNN based source separation. IEICE Transactions on Information and Systems, E100D(12), 3041-3044.

[6] Y. Han, J. Park and K. Lee, "Convolutional Neural Networks with Binaural Representations and Background Subtraction for Acoustic Scene Classification,", IEEE AASP Challenge on DCASE 2017 technical reports, 2017.

[7] M. Valenti, S. Squartini, A. Diment, G. Parascandolo and T. Virtanen, "A convolutional neural network approach for acoustic scene classification," 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, 2017, pp. 1547-1554.

[8] Hyder, Rakib, Shabnam Ghaffarzadegan, Zhe Feng, and Taufiq Hasan. "BUET bosch consortium (B2C) acoustic scene classification systems for DCASE 2017." IEEE AASP Challenge on DCASE 2017 technical reports (2017).

[9] Lidy, Thomas, and Alexander Schindler. "CQT-based convolutional neural networks for audio scene classification." In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016), vol. 90, pp. 1032-1048. 2016.

[10] Ren, Zhao, Kun Qian, Yebin Wang, Zixing Zhang, Vedhas Pandit, Alice Baird, and Bjorn Schuller. "Deep scalogram representations for acoustic scene classification." IEEE/CAA Journal of Automatica Sinica 5, no. 3 (2018): 662-669.

[11] A. Singh, A. Thakur, P. Rajan and A. Bhavsar, "A Layer-wise Score Level Ensemble Framework for Acoustic Scene Classification," 2018 26th European Signal Processing Conference (EUSIPCO), Rome, 2018, pp. 837-841.

[12] A. Solomonoff, C. Quillen, W.M. Campbell, "Channel Compensation for SVM Speaker Recognition", in: Proc. of Odyssey, pp. 5762, 2004.

[13] A. Solomonoff, W. M. Campbell and I. Boardman, "Advances in channel compensation for SVM speaker recognition," Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005., Philadelphia, PA, 2005, pp. I/629-I/632 Vol. 1.

[14] V. Štruc, B. Vesnicer, F. Mihelič and N. Pavešić, "Removing illumination artifacts from face images using the nuisance attribute projection," 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, TX, 2010, pp. 846-849.

[15] Mesaros, Annamaria, Toni Heittola, Aleksandr Diment, Benjamin Elizalde, Ankit Shah, Emmanuel Vincent, Bhiksha Raj, and Tuomas Virtanen. "DCASE 2017 challenge setup: Tasks, datasets and baseline system." 2017.

[16] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in arXiv:1409.1556, 2014.

[17] J. Cramer, H. Wu, J. Salamon and J. P. Bello, "Look, Listen, and Learn More: Design Choices for Deep Audio Embeddings," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom, 2019, pp. 3852-3856.