Available online at www.sciencedirect.com

ScienceDirect

Journal of the Franklin Institute 357 (2020) 667–686

www.elsevier.com/locate/jfranklin

Urban acoustic classification based on deep feature transfer learning

Yexin Shen[a,b], Jiuwen Cao[a,b,*], Jianzhong Wang[a,b], Zhixin Yang[c]

[a] *Key Lab for IOT and Information Fusion Technology of Zhejiang, Hangzhou Dianzi University, Zhejiang 310018, China*

[b] *Artificial Intelligence Institute, Hangzhou Dianzi University, Zhejiang 310018, China*

[c] *State Key Laboratory of Internet of Things for Smart City, Faculty of Science and Technology, University of Macau, Macau, China*

Received 31 March 2019; received in revised form 14 July 2019; accepted 11 October 2019
Available online 24 October 2019

Abstract

Urban acoustic classification (UAC) plays a vital role in smart city engineering, urban security, noise pollution analysis, etc. Convolutional neural networks (CNNs) based feature transfer learning have been shown competitive performance in many applications but little attention has been paid to UAC. In this study, a novel UAC algorithm exploiting the deep CNNs based feature transfer learning and the deep belief net (DBN) based classification is developed. The spectrogram is first adopted for the urban acoustic stream representation. Then, three deep CNNs pre-trained on ImageNet database are applied as feature extractors. The extracted features are concatenated and fed to a DBN for classifier learning. To achieve a good generalization performance, three restricted boltzmann machines (RBM) trained by the contrastive divergence algorithm (CD) followed by a back-propagation (BP) based fine parameter tuning is adopted in DBN. The proposed UAC is evaluated on a real acoustic database, including 11 categories of acoustic signals recorded from the urban environment. Performance comparisons to many state-of-the-art algorithms are presented to demonstrate the superiority of the proposed method.

* Corresponding author at: Key Lab for IOT and Information Fusion Technology of Zhejiang, Hangzhou Dianzi University, Zhejiang 310018, China.

E-mail address: jwcao@hdu.edu.cn (J. Cao).

https://doi.org/10.1016/j.jfranklin.2019.10.014

1. Introduction

UAC aims at the recognition of acoustic streams frequently encountered in the urban environment, which are found crucial to smart city engineering [1–11]. For instances, Asensio [1] organized a special issue on "acoustics in smart cities", which the recent achievements on urban sounds monitoring, soundscape assessment, analyses and management were reported. Particularly in the special issue, Bellucci et al. [2] and Mydlarz et al. [3] studied the development and implementation of low-cost sensors for urban acoustic measuring and monitoring. Aumond et al. [4] investigated the mobile phone based urban noise pollution monitoring and Agha et al. [5] developed an intelligent surveillance camera to automatically capture passing vehicles causing noise pollution to urban cities. Ye et al. [6], Piczak [7], Cao et al. [8], and Zhao et al. [12] focused on UAC with the classical methods and the popular deep learning algorithms, respectively. But UAC is still a challenging task as urban acoustical environment are complex, changeable and normally mixed with a large variety of different background noises.

Effective acoustical signal representations are crucial to UAC. The conventional time- and frequency-domain hand-crafted features, which are proved to be effective in speech/audio recognition, have been studied in UAC [6,11,13–16], including the zero-crossing rate, frequency energy, linear prediction cepstral coefficients (LPCC) [19], Mel-frequency cepstral coefficient (MFCC) [20], etc. But traditional hand-crafted features may suffer degraded generalization performance when dealing with large and complex urban acoustics [8,12,15]. Deep neural networks (DNN), which are powerful in complex and high-dimensional data learning [17,18], have been recently investigated for environment acoustic classification. To name a few, Huzaifah [19] compared the performance on acoustic recognition with a 3-layered CNN on various classical hand-crafted features. Tak et al. [20] proposed to use the phase encoded Mel filter bank energies with CNN. Sailor et al. [22] used the convolutional RBM to learn filter banks from raw acoustic signals for classification. Xue and Su [27] adopted a DBN learning the contextual correlations of audio clips to perform classification. Comparing to the hand-crafted features, the acoustic spectrogram, which has a better description on the time-frequency property of acoustic streams, has been investigated for UAC. Boddapati et al. [21] studied the environment recognition performance with an AlexNet on spectrograms [7,8,12,21]. Zhao et al. [12] studied the wavelet transform based scalogram with deep networks. Piczak [7] and Cao et al. [8] presented the UAC results using CNN with the Mel filter bank spectrogram. But existing deep CNNs spend a long time in model tuning when dealing with a complex urban acoustic database, and the resultant model is generally not robust in parameter updation.

Feature transfer learning with pre-trained DNNs on large scale databases [23–26,47] have been shown effective in discriminative feature learning for many applications [39,45]. DNNs trained on the ImageNet [31] and other large image databases have been widely studied as feature extractors for image recognition. For examples, Nguyen et al. [39] applied three pre-trained deep CNNs for feature learning on microscopic image classification. Deng et al. [40] exploited the pre-trained VGG-16 network for feature extraction on face images. Krizhevsky et al. [41] studied the deep feature extraction for audio waveforms using a pre-trained AlexNet. Akilan et al. have comprehensively studied the deep CNN based transfer feature learning and feature fusion methods in [42,43]. For instance, various late feature fusion strategies combining with the transfer learning models of the pre-trained Inception-v3, VGG-16 and AlexNet have been investigated in [42]. Although fruitful results

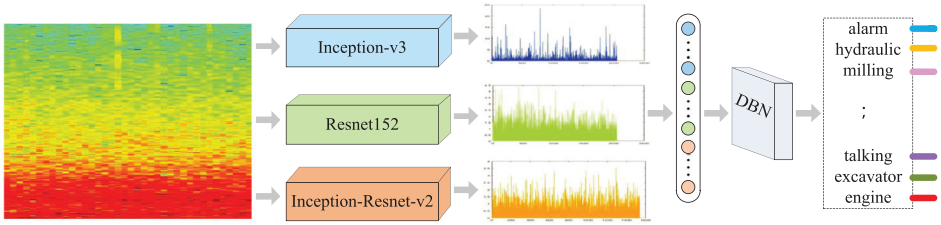


Fig. 1. The proposed UAC.

on deep transfer learning have been presented in the past, few attention has been paid to UAC.

In this paper, a novel UAC algorithm exploiting the deep CNNs based feature transfer learning and the DBN based classification is proposed. The acoustic streams are first transformed into the spectrogram image to have a good time-frequency characterization. Three representative deep CNNs, namely, the Inception-v3 [28], ResNet152 [29], and Inception-ResNet-v2 [30], pre-trained using the ImageNet database, are then adopted as feature extractors. The Inception-v3 is an extension of the popular GoogLeNet [44] with a good generalization performance. The ResNet152 is a special structure of the ResNet families [29], which enhances the performance of deep networks by adopting the residual units to alleviate the degradation. The Inception-ResNet-v2, exploiting the merits of combining the inception structure with the residual unit, has shown good generalization performance. The extracted features of the three deep CNNs are then concatenated and fed to a DBN for classifier training. To achieve a good generalization performance, three RBMs trained by the contrastive divergence (CD) algorithm is adopted in DBN. The flowchart of the proposed UAC is minutely shown in Fig. 1. For the performance evaluation, a real acoustic stream database consisting of 11 categories of frequently encountered acoustics in urban environment is recorded for experiments. The urban acoustic signals include the acoustic waves produced by five mostly used machines in urban constructions, the engine sounds and horns of various passing vehicles, the wind noises, the sound of high-rating generator, human talking and music. A plenty of comparisons to many existing hand-crafted features, such as the LPCC and MFCC, and the deep CNNs based transfer learning features, combining with state-of-the-art classifiers, including the support vector machine (SVM), artificial neural network, and DBN, are provided in the paper. The proposed algorithm achieves an average of 98.55% classification accuracy, which generally outperforms all compared methods.

The rest of the paper is organized as follows: in Section 2, we introduce the proposed UAC algorithm, in Section 3, we show the detailed experiments, results, and comparisons, and the conclusions and future works are given in Section 4.

2. The proposed UAC

2.1. Spectrograms of urban acoustic waveforms

Spectrograms are generated by stacking the spectra of consecutive frames and are shown effective in characterizing the time-frequency energy variations of acoustic signals. In this paper, all the raw acoustic streams are first filtered by a high-pass filter for noise reduction

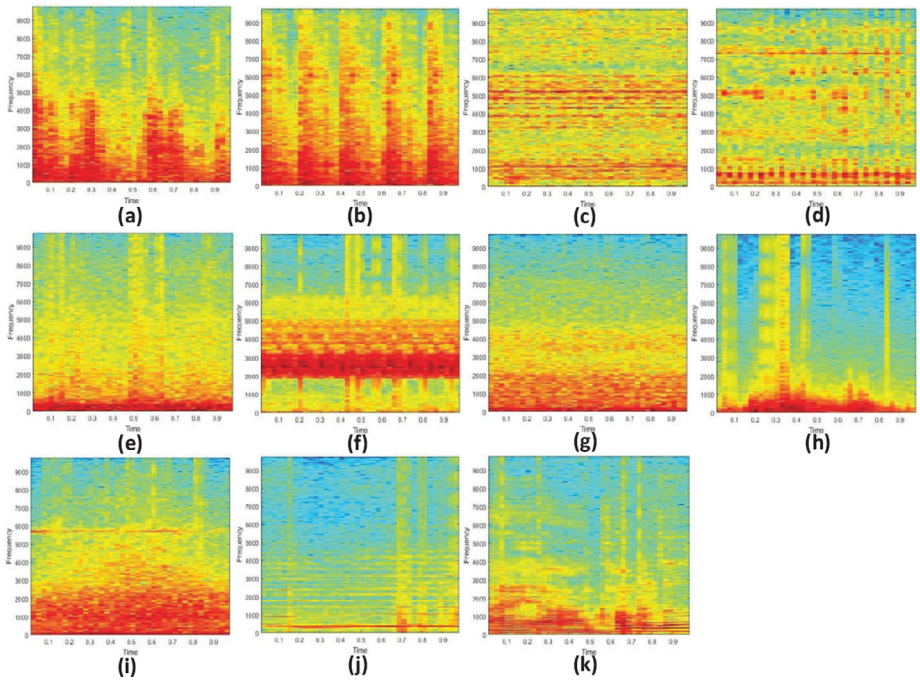


Fig. 2. Spectrograms of 11 urban acoustic waveforms: (a) excavator (b) hydraulic hammer (c) cutting machine (d) electric hammer (e) milling machine (f) horns (g) generator (h) noises (i) engine sound (j) music (k) talking.

and then the short-time Fourier transform (STFT) is adopted to calculate the spectra as

$$X(\omega) = \sum_{m=-\infty}^{\infty} x(m)w(n-m)e^{-j\omega m}, \tag{1}$$

where $x(m)$ is the acoustic signal, $w(n)$ is the Hanning-window function, ω is the frequency and $X(\omega)$ is the associated Fourier transform. The logarithmic power scale (dB) is finally computed to generate the spectrograms. For each acoustic signal, 1024 samples are used in each frame to compute the STFT and the consecutive frame overlap is set to be 50%. Fig. 2 plots the spectrograms of 11 urban acoustic waveforms studied in this paper, where the time duration for each spectrogram image is set to be 1 second. As shown in the figure, different urban acoustics exhibit different power spectra distributions within the time-frequency scale. For instances, the power spectra of the acoustic waveforms generated by cutting machine and electric hammer distributes relatively even within the frequency ranges. While for excavator and hydraulic hammer, several clear pulse-shaped power spectra bands ranging from the low to high frequencies can be found.

2.2. Feature transfer learning with deep CNNs

To exploit an effective representation for urban acoustic streams using spectrograms, we investigate the feature transfer leaning approach based on three popular deep CNNs pre-trained on the ImageNet database, namely, the Inception-v3 [28], ResNet152 [29] and Inception-ResNet-v2 [30], which are briefly described in the following, respectively.

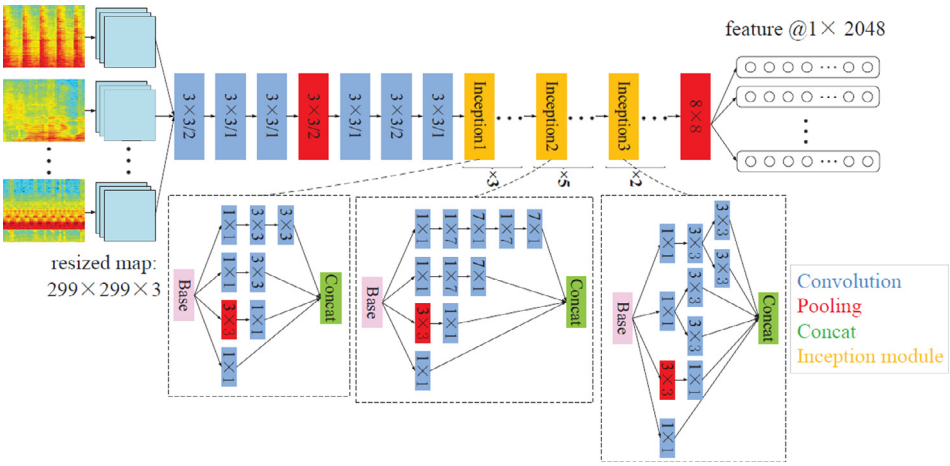


Fig. 3. Inception-v3.

The Inception-v3 [28] is a novelly extended architecture based on the popular GoogLeNet [44]. In Inception-v3, several design principles are exploited to scale up networks in an efficient way and at the meantime, to achieve an improved performance. The large spatial convolution filters in Inception-v3 are factorized into multilayer networks with small convolutions to reduce the computational cost. As stated in [28], replacing a 5×5 convolution with two layers of 3×3 convolutions can achieve a net of $\frac{18}{25}\times$ reduction. Meanwhile, the grid reduction technique using the parallel stride blocks was adopted in the pooling operation to decrease the feature sizes. The label-smoothing regularization with batch-normalized auxiliary classifiers enabled the Inception-v3 a good generalization performance. In this paper, we adopt the pre-trained Inception-v3 as a feature extractor on the spectrograms of urban acoustic streams. The spectrogram images are firstly converted into the size of $299 \times 299 \times 3$. The same structure to [28] is used in this paper (the detailed structure is drawn in Fig. 3), where the convolution factorization and the traditional reception module reduction with the grid size reduction technique are adopted. The resultant features for each spectrogram image are 2048 dimensions.

The ResNet [29] addressed the degradation problem suffered in DNNs that the network accuracy gets saturated when the depth increases. Instead of using stacked networks to approximate the desired mapping $\mathcal{H}(\mathrm{x})$, ResNet aims to fit the residual $\mathcal{F}(\mathrm{x}) := \mathcal{H}(\mathrm{x}) - \mathrm{x}$. It is evident [29] that optimizing the residual mapping is easier than the original mapping for some applications. Such structure was also referred as the shortcut connection. The residual network was then used as the basic building block in ResNet as

$$\mathrm{y} = \mathcal{F}(\mathrm{x}, \{W\}) + \mathrm{x}, \tag{2}$$

where x, y, and $\mathcal{F}(\mathrm{x}, \{W\})$ represent the input, output and residual mapping to be learned in the block, respectively. Evaluations conducted on various benchmark databases have shown that the ResNet with 152 layers obtained the best accuracy. To exploit the capability of residual learning, we adopt the ResNet152 pre-trained on ImageNet as a feature extractor on the urban acoustic spectrogram, where similarly, each spectrogram image is first converted into a size of 229×299 and a resultant feature vector with 2048 dimensions is then obtained.

The Inception-ResNet-v2 [30] made use of the Inception structure and the residual connections, where the residual learning is employed in the Inception modules. It consists of three inception blocks with different interior grid modules. Benefiting from the merits of residual learning and inception module, the resultant Inception-ResNet-v2 achieved a significantly improved performance. We use the pre-trained Inception-ResNet-v2 to extract features on the spectrogram image, where the size of each acoustic spectrogram is 299×299 and the dimensions of the final extracted feature vector are 1536.

2.3. DBN based UAC

The extracted features by the aforementioned three deep CNNs are concatenated into a long vector for the urban acoustic representation, where the dimension of the concatenated feature vector is 5632. Instead of directly using the concatenated features for classifier training, a discriminative feature learning and classifier training algorithm based on a DBN [45–50] is adopted for UAC, where for the DBN, a popular and generic structure consisting of three RBMs followed by a discriminative fine-tuning based on BP is used in this paper. As stated in [45], DBN exhibits significant advantages because of its efficient and layer-wised learning procedure. Particularly, the learning phase of DBN is completed by two stages, namely the pre-training of RBM and the fine-tuning by BP.

DBN is a composition of RBM modules learned in a layer-wised top-down procedure. An RBM is an energy-based module with unsupervised learning, which generally consists of two layers of binary units, including a visible layer for input data x and a hidden layer to be learned for feature representation h. Defining an energy function on state variables as $\mathbb{E}(\mathrm{x}, \mathrm{h}; \Theta) = -\mathrm{h}^T W \mathrm{x} - \mathrm{h}^T \mathrm{b} - \mathrm{x}^T \mathrm{c}$, the joint probabilities $P(\mathrm{x}, \mathrm{h})$ can be obtained

$$P(\mathrm{x}, \mathrm{h}) = \frac{1}{Z(\Theta)} e^{-\mathbb{E}(\mathrm{x},\mathrm{h};\Theta)}, \tag{3}$$

where $Z(\Theta) = \sum_{(\mathrm{x},\mathrm{h})} e^{-\mathbb{E}(\mathrm{x},\mathrm{h};\Theta)}$ is the normalizing constant, and $\Theta = \{W, \mathrm{b}, \mathrm{c}\}$ denotes the parameters, including the weight W connecting the visible layer and the hidden layer as well as their associated biases b, c. Thus, the probability of a input data x can be expressed as the marginal $p(\mathrm{x}) = \sum_{\mathrm{h}} P(\mathrm{x}, \mathrm{h}) = \frac{\sum_{\mathrm{h}} e^{-\mathbb{E}(\mathrm{x},\mathrm{h};\Theta)}}{Z(\Theta)}$. Given a certain training data $\{\mathrm{x}_1, \ldots, \mathrm{x}_K\}$, the negative log-likelihood function is expressed as

$$-\ell(\Theta) = -\log \prod_k \sum_{\mathrm{h}} p(\mathrm{x}_k, \mathrm{h}; \Theta). \tag{4}$$

Suppose p^0 is the distribution of the given training data, which is independent of Θ, it is known that $\sum_k \log p^0(\mathrm{x}_k)$ is a constant. Adding $\sum_k \log p^0(\mathrm{x}_k)$ into (4) will not affect the optimal Θ and thus, (4) becomes $-\tilde{\ell}(\Theta) = \sum_k \log p^0(\mathrm{x}_k) - \log \prod_k \sum_{\mathrm{h}} p(\mathrm{x}_k, \mathrm{h}; \Theta)$. Computing the expectation of $-\tilde{\ell}(\Theta)$ with respect to p^0, denoted as $\langle -\tilde{\ell}(\Theta) \rangle_{p^0}$, leads to

$$\begin{aligned} \langle -\tilde{\ell}(\Theta) \rangle_{p^0} =: & \sum_k p^0(\mathrm{x}_k) \log p^0(\mathrm{x}_k) \\ & - \sum_k p^0(\mathrm{x}_k) \log \frac{\sum_{\mathrm{h}} e^{-\mathbb{E}(\mathrm{x}_k,\mathrm{h};\Theta)}}{Z(\Theta)}, \end{aligned} \tag{5}$$

which becomes computing the Kullback-Leibler divergence between the distribution $p^0(\mathrm{x})$ of the training data and the model distribution $p^\Theta(\mathrm{x})$. Hence, the minimization of (5) becomes

finding Θ that the model distribution can approximate the training data distribution as better as possible.

As pointed out in [45,46], to minimize (5), one can recur to calculating the gradient of $\langle -\tilde{\ell}(\Theta)\rangle_{p^0}$ with respect to Θ, that is

$$\frac{\partial\langle -\tilde{\ell}(\Theta)\rangle_{p^0}}{\partial\Theta} = \left\langle \frac{\sum_{\mathrm{h}} e^{-\mathbb{E}(\mathrm{x}_k,\mathrm{h};\Theta)}\partial_\Theta\mathbb{E}(\mathrm{x}_k,\mathrm{h};\Theta)}{\sum_{\mathrm{h}} e^{-\mathbb{E}(\mathrm{x}_k,\mathrm{h};\Theta)}} \right\rangle_{p^0} - \frac{\sum_{(\mathrm{x},\mathrm{h})} e^{-\mathbb{E}(\mathrm{x},\mathrm{h};\Theta)}\partial_\Theta\mathbb{E}(\mathrm{x},\mathrm{h};\Theta)}{\sum_{(\mathrm{x},\mathrm{h})} e^{-\mathbb{E}(\mathrm{x},\mathrm{h};\Theta)}}, \tag{6}$$

where the first term of (6) is the expectation of the input data, which is easy to calculate. But the second term, with a simplified expression $\sum_{(\mathrm{x},\mathrm{h})} P(\mathrm{x},\mathrm{h})\partial_\Theta\mathbb{E}(\mathrm{x},\mathrm{h};\Theta)$, is practically intractable as the integration is related to $2^{|\mathrm{x}|+|\mathrm{h}|}$ combination of the data in visible and hidden units.

To approximate the second term of (6), the CD algorithm has been developed [45], where the Markov Chain Monte Carlo with the Gibbs sampling approach is adopted for the approximation iteratively. That is, with the given data and initialized network parameters, the probability of hidden units is firstly estimated. Then, the sampled hidden units from the estimated probability with Gibbs sampling can be used to reconstruct the visible units and to estimate its associated probability. The derived probability of the visible units can be further used to calculate the probability of hidden units, where these estimations are adopted to approximate the gradient and perform the parameter updation. Assume that a sample x^0 is randomly selected, after k times Gibbs sampling with each sampling step obeying

$$\mathrm{h}^{t-1} = P\left(\mathrm{h}\middle|\mathrm{x}^{t-1}\right), \tag{7}$$

$$\mathrm{x}^{t} = P\left(\mathrm{x}\middle|\mathrm{h}^{t-1}\right), \tag{8}$$

we can get the sample x^k. It can be further used to approximate the gradients associated to the weight and biases $\Theta = \{W, \mathrm{b}, \mathrm{c}\}$ as

$$\frac{\partial \ln P(\mathrm{x})}{\partial w_{ij}} \approx P\left(h_i = 1\middle|\mathrm{x}^0\right)x_j^0 - P\left(h_i = 1\middle|\mathrm{x}^k\right)x_j^k, \tag{9}$$

$$\frac{\partial \ln P(\mathrm{x})}{\partial b_i} \approx x_i^0 - x_i^k, \tag{10}$$

$$\frac{\partial \ln P(\mathrm{x})}{\partial c_i} \approx P\left(h_i = 1\middle|\mathrm{x}^0\right) - P\left(h_i = 1\middle|\mathrm{x}^k\right). \tag{11}$$

Instead of using a single sample, a mini-batch set is adopted in RBM training. The above gradient will be then approximated using the average result of the mini-batch samples. The parameters Θ updation is expressed as

$$\Theta \leftarrow \Theta + \epsilon \cdot \partial_\Theta, \tag{12}$$

where Θ represents the set of parameters, ϵ is the learning rate and ∂_Θ is the approximated gradient. The approximation is performed recursively in RBM.

Table 1
Specifications of UAC datasets (Training+Testing).

Datasets	Spectrogram	LPCC	MFCC
Excavator	1713 (1285+428)	68307 (51230+17077)	68307 (51230+17077)
Hydraulic hammer	4810 (3608+1202)	185185 (138889+46296)	185185 (138889+46296)
Cutting machine	954 (716+238)	36762 (27572+9190)	36762 (27572+9190)
Electric hammer	383 (287+96)	14554 (10916+3638)	14554 (10916+3638)
Milling machine	733 (550+183)	28048 (21036+7012)	28048 (21036+7012)
Horns	2287 (1715+572)	86906 (65180+21726)	86906 (65180+21726)
Sound of generator	5181 (3886+1295)	196880 (147660+49220)	196880 (147660+49220)
Wind noise	2792 (2094+698)	106627 (79970+26657)	106627 (79970+26657)
Engine sound	1488 (1116+372)	56809 (42607+14202)	56809 (42607+14202)
Music	4434 (3226+1108)	170709 (128032+42677)	170709 (128032+42677)
Talking	1617 (1213+404)	61446 (46085+15361)	61446 (46085+15361)

In the proposed UAC, the learned features of the hidden layer for each RBM are fed to the subsequent RBM module as the inputs. Finally, the parameters of DBN are tuned by the BP algorithm to minimize the error between the network outputs and the desired targets. For UAC, the extracted features from the three pre-trained CNNs are 2048, 2048, and 1536 dimensions, respectively. The concatenated long feature vector is employed for representation and classifier learning.

3. Experiments and discussions

3.1. Database description

We evaluate the performance of the proposed UAC algorithm on a real recorded database, consisting of 11 frequently encountered urban acoustic signals. Acoustic waveforms generated by five construction machines are captured as it is reported that an intelligent monitoring and management to these machines is crucial to the urbanization construction and urban security [14]. Horns produced by various passing vehicles, including the cars, motorcycles, electrombiles, etc., are another important acoustics in smart transportation and urban noise control. The rest categories include the sound of generator, wind noises, engine sound of vehicles, music and human talking. A portable device equipped with a cross microphone array and data processing terminal is developed, where the sampling frequency is set to be 19.53 kHz. For each category, the acoustic streams are recorded under various conditions. For instances, all acoustic signals are captured under different propagation distances to the sources, and the wind noises, engine sound of vehicles, music and human talking are recorded from many urban streets under various sound field environment.

Table 1 shows the numbers of sample in each urban acoustic dataset. For spectrogram, the size of Hanning Window function is set to be 1024 samples for calculating the STFT and the time duration of the acoustic streams is set to be 1 second, where the frame overlap is 50%. The frame length of LPCC and MFCC features is the same to the spectrogram. For LPCC, 16-ordered features are extracted and for MFCC, 12-ordered filter coefficients are obtained. For each experiment, the UAC datasets are partitioned into training and testing datasets with a ratio 3: 1. The four-fold cross-validation is adopted, in which 75% and 25% of the training dataset are employed for the classifier training and performance validation,

respectively. For all algorithms, the experiments are conducted on the software platform using Python 3.5 + Tensorflow. A PC with an Intel(R) Core(TM) i7-6700 CPU and a NVIDIA GeForce GTX 1070 Graphics card is adopted for experiment. The operating system is 64-bit Windows 10.

3.2. Experimental setups

We compare the performance of the proposed UAC algorithm with several state-of-the-art acoustic feature extraction and classification methods, which are listed in details in the following

- *Hand-crafted features*: the popular LPCC [19] and MFCC [20] acoustic features mentioned above;
- *Transfer learning features by single CNN*: the features extracted by single CNN on spectrogram with Inception-v3 [28], ResNet152 [29], and Inception-ResNet-v2 [30], respectively;
- *SVM*: SVM trained on hand-crafted features;
- *Extreme learning machine (ELM)*: the artificial neural network based ELM [33] trained on hand-crafted features;
- *DBN*: DBN [45,46] trained on hand-crafted features and transfer learning features by single CNN, respectively;
- *Hierarchical extreme learning machine (H-ELM)*: H-ELM [51] trained on concatenated transfer learning features by three deep CNNs used in the paper;
- *Multilayer extreme learning machine (ML-ELM)*: ML-ELM [52] trained on concatenated transfer learning features by three deep CNNs used in the paper;
- *Sparse Autoencoder (SAE) with softmax layer*: the classifier with SAE and the softmax function [53] trained on the concatenated transfer learning features by three deep CNNs used in the paper.
- *Fully-connected layer*: the classifier with two fully-connected layers with the softmax function [39] trained on the concatenated transfer learning features by three deep CNNs used in the paper.
- *PCA+DBN*: the concatenated transfer learning features reduced by principle component analysis (PCA) and trained by DBN.
- *PCA+SVM*: the concatenated transfer learning features reduced by PCA and trained by SVM.

Feedforward neural networks (FNNs) with random parameters can be tracked back to [32]. ELM is one of the popular learning algorithms for random FNNs [33–38] and becomes attractive for its fast classifier training speed and less human-intervention parameters. ELM states that all hidden node parameters, including input weights and hidden biases, can be randomly generated and remain unchanged during the training. Only the hidden node size needs to be tuned and the least-squares (LS) approach is employed for solving the output weight. In this paper, the LPCC and MFCC features are trained with ELM, respectively, for the performance comparisons of urban acoustic classification.

In the proposed UAC, the hidden unit sizes of three RBMs are set to be 2000, 1000, and 500, respectively. For LPCC+DBN, various DBN structures with different hidden nodes in RBM are adopted for testing, including ‘16-16-16’, ‘50-30-20’, ‘100-50-25’, ‘150-100-50’, ‘200-150-100’, ‘500-250-100’, ‘1000-500-250’, and ‘2000-1000-500’, respectively. The same

structures of RBM are studied in MFCC+DBN except for the first one, which is replaced by '12-12-12'. The structure of DBN tested in the proposed UAC includes '300-200-100', '500-300-200', '750-500-250', '1000-500-250', '1200-600-300', '1500-750-350', '1800-900-450', and '2000-1000-500', respectively. In RBM, the activation functions used in hidden layer and output layer are Sigmoid and Sotfmax, respectively, and the loss function is cross entropy. The batch sizes in learning are 32. The initial learning rate is 1e-3 and 10 iteration epochs are used in the CD algorithm. For the fine tuning BP algorithm in DBN, the following parameters are adopted: the initial learning rate 1e-3, the iteration epochs 30, the dropout 0.12, and the momentum 0.5. The weights in RBMs are randomly initialized with the Normal distribution with a standard deviation of $\sqrt{2/(n_v + n_h)}$, where n_v and n_h are the sizes of visible and hidden layers, respectively. The offsets of visible and hidden layers are set as 0. The activation function used in ELM is Sigmoid and its hidden node size is optimized by searching within the region $\{100, 200, \ldots, 1000, 2000, \ldots, 6000\}$. The Gaussian kernel function is used in SVM, where the regularization parameter and the kernel size are optimized within the grid $\left\{2^{-2}, 2^{-1}, \ldots, 2^{12}\right\} \times \left\{2^{-10}, 2^{-9}, \ldots, 2^{4}\right\}$. For H-ELM, the linear function and Sigmoid function are used as the activation function in the AE and classifier layers, respectively. The number of hidden nodes of AE is optimized from 1000 to 4000, the number of hidden nodes of ELM classifier is optimized from $\{500, 600, \ldots, 1000, 2000, \ldots, 5000\}$. For ML-ELM, The Sigmoid function is used as the activation of AE and ELM classifier. Three hidden layers of AEs are used in ML-ELM, where the number of hidden nodes of AEs are optimized from '1000-500-200', '1000-700-500', '2000-1000-500', '2000-1000-700', '2000-1000-800', '3000-1500-700', '3000-1500-700', '3000-1500-800', '3000-1500-900', '3000-1500-1000', and '4000-2000-1000', respectively. For SAE with softmax classifier, the learning rate, sparsity parameter, lambda, beta are set to be $\{0.01, 0.1, 1e-4, 3\}$, correspondingly. The number of hidden nodes of SAE is optimized from 200 to 1000. For the fully-connected layer classifier, the hidden nodes, learning rate, batch size, dropout, learning rate decay and epoches are set to be {100, 0.01, 1, 0.8, 3000}, correspondingly.

3.3. Results and comparisons

Four experiments are carried out for the performance comparisons in this section. The first one focuses on the comparisons with hand-crafted acoustic features, where six algorithms, namely, employing LPCC and MFCC with SVM, ELM, and DBN, respectively, are adopted. The detailed combinations of these methods are shown in Table 2. For each algorithm, the best average results on multiple trials are reported. That is, for the DBN classifier based algorithms, the RBM structure with the highest classification accuracy is reported in Table 2. Similarly, for ELM, the optimal hidden neuron size with the best classification performance is adopted, and for SVM, the parameter combination of regularization coefficient and kernel size with the highest UAC rate is used.

Table 2 lists the average classification accuracy offered by the proposed UAC and the six compared algorithms on 11 urban acoustic datasets, respectively. As highlighted, comparing with the six hand-crafted feature based algorithms, the proposed UAC achieves the highest classification rate on 9 out of 11 datasets. It performs consistently good that the classification accuracies on all datasets are higher than 92%. The average classification rate obtained by the proposed UAC is higher than 98.5%, which offers 8.94%, 11.83%, 15.77%, 9.33%, 13.5%, and 12.9% increments over the six compared algorithms, respectively. One can readily find that the conventional hand-crafted feature based algorithms perform relatively poor on the acoustic

Table 2
Comparison with Hand-crafted features.

Datasets	Proposed UAC	LPCC+SVM	LPCC+ELM	LPCC+DBN	MFCC+SVM	MFCC+ELM	MFCC+DBN
Excavator	**96.33**	72.00	63.25	65.50	62.38	50.75	44.00
Hydraulic hammer	**99.80**	84.62	83.00	76.00	82.13	73.25	77.75
Cutting machine	98.40	97.88	97.50	**99.13**	96.63	95.13	94.88
Electric hammer	98.57	98.38	98.25	**99.50**	94.25	93.75	89.88
Milling machine	**97.79**	95.75	96.37	94.50	94.28	92.88	97.38
Horns	**100.00**	99.88	99.87	**100.00**	**100.00**	**100.00**	**100.00**
Sound of generator	**99.51**	94.38	94.50	94.50	95.63	94.00	96.50
Wind noise	**95.33**	88.50	90.12	83.88	88.88	87.63	87.88
Engine sound	**92.88**	85.12	80.75	75.38	78.50	71.38	73.00
Music	**99.67**	94.25	91.87	79.38	96.25	94.75	94.50
Talking	**99.69**	77.87	68.63	55.88	87.63	78.50	72.00
Average	**98.55**	89.61	86.72	82.78	89.22	85.05	85.65

datasets of excavator, hydraulic hammer, engine sound, and talking. The less discriminative capability of LPCC/MFCC in characterizing these four datasets would be the cause of low classification rate.

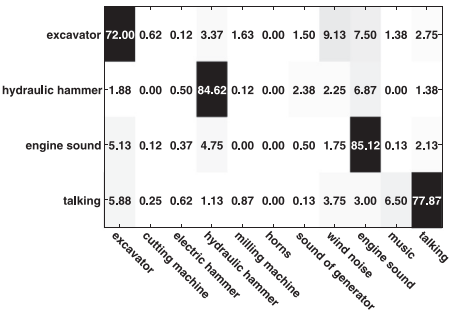
Further, the confusion matrices of hand-crafted feature based algorithms on aforementioned four datasets are presented in Fig. 4. With LPCC/MFCC, there exists a large amount of misclassification among excavator, wind noise, engine sound, hydraulic hammer, talking, and music. For instances, using LPCC, the acoustic samples of excavator are mostly misclassified to wind noise and engine sound, where the misclassification percentages are 9.13% and 7.5% for LPCC+SVM, 11.88% and 7.88% for LPCC+ELM, 13.25% and 6.5% for LPCC+DBN, respectively. While using MFCC, the acoustic samples of excavator are mostly misclassified to hydraulic hammer, milling machine, wind noise, and engine sound. The misclassification percentages are 5%, 7.13%, 11.48% and 7.38% for MFCC+SVM, 7.38%, 7.88%, 12.25% and 8% for MFCC+ELM, 9.75%, 13.13%, 14.63% and 9.75% for MFCC+DBN, respectively. Similar observations can be found to other classes of urban acoustics.

For each feature dataset (LPCC, MFCC, the deep CNNs based transfer learning feature), the average intra- and inter-class Euclidean distances are calculated for investigation and validation in Fig. 5, respectively. The definitions of the average intra- and inter-class Euclidean distances d_{tra} and d_{ter} are

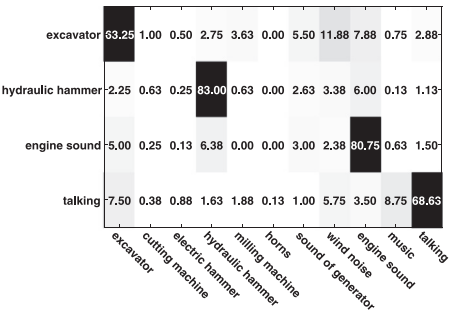
$$d_{tra} = \frac{1}{N_i \cdot N_j} \sum_{k=1}^{N_i} \sum_{l=1}^{N_j} \left\| \mathrm{x}_k^{(i)} - \mathrm{x}_l^{(j)} \right\|_2^2, \tag{13}$$

$$d_{ter} = \frac{1}{N_i^2} \sum_{k=1}^{N_i} \sum_{l=1}^{N_i} \left\| \mathrm{x}_k^{(i)} - \mathrm{x}_l^{(i)} \right\|_2^2, \tag{14}$$

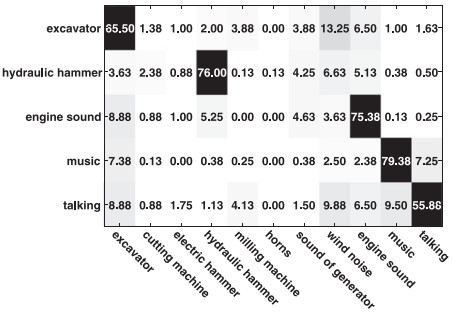
where $\mathrm{x}_k^{(i)}$ and $\mathrm{x}_k^{(j)}$ denote the samples from the i-th and j-th classes, respectively, N_i and N_j represent the total samples of the i-th and j-th classes, respectively. In general, a discriminative feature should produces a small inter-class Euclidean distance, and meanwhile, the intra-class Euclidean distance should be as large as possible. The results shown in Fig. 5 well support the



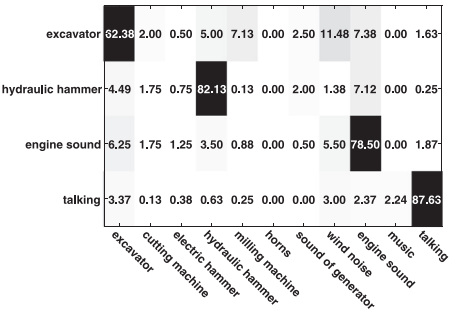
(a) LPCC+SVM



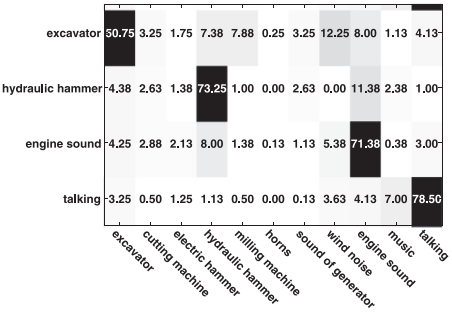
(b) LPCC+ELM



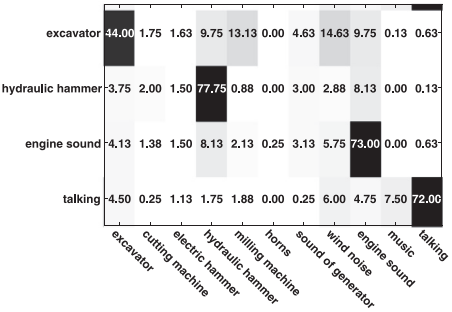
(c) LPCC+DBN



(d) MFCC+SVM



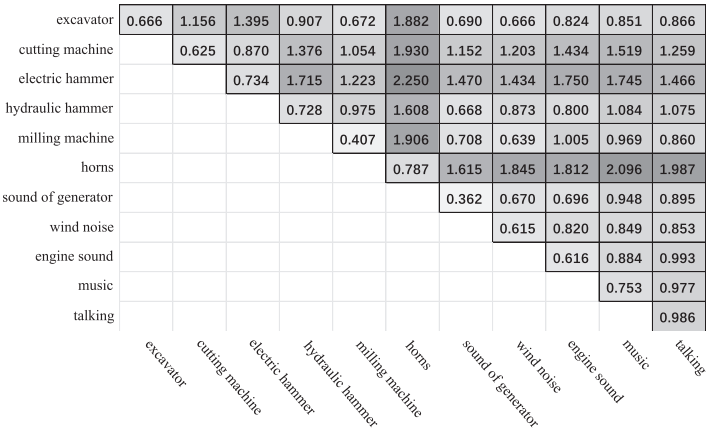
(e) MFCC+ELM



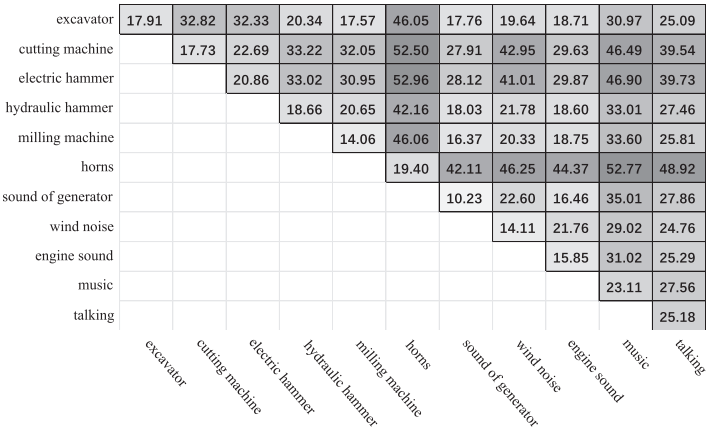
(f) MFCC+DBN

Fig. 4. Confusion matrices of classification accuracy by LPCC/MFCC based algorithms.

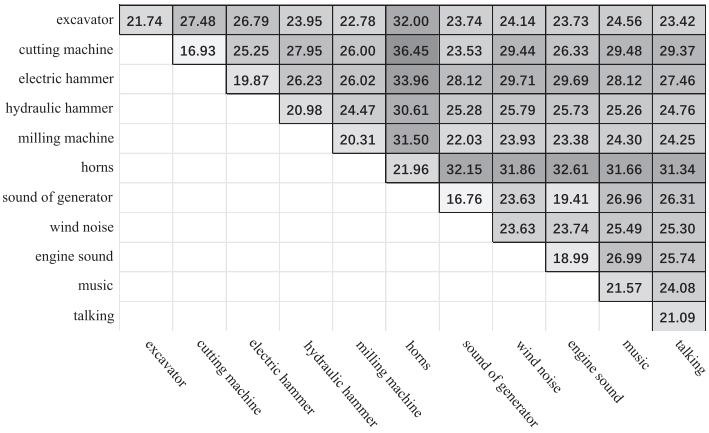
previous observations. One can find that the LPCC/MFCC feature based intra-class Euclidean distances of excavator, hydraulic hammer and talking are not the smallest when comparing with its associated inter-class Euclidean distances. On the contrary, the intra-class Euclidean distances of the concatenated deep CNNs based transfer learning features are always smaller than the associated inter-class distances.



(a) LPCC



(b) MFCC



(c) Transfer learning features

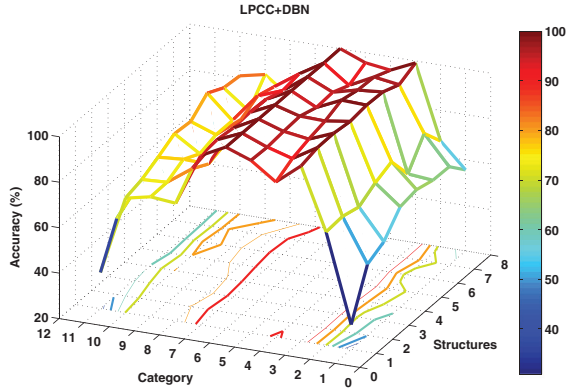
Fig. 5. Intra- and inter-class Euclidean distance comparisons of the LPCC, MFCC, and concatenated transfer learning features by deep CNNs, respectively.

The sensitive and robustness of DBN with respect to (w.r.t.) different node sizes in RBM are also studied in the first experiment, where RBM structures presented in Section 3.2 are investigated. The classification accuracies obtained by LPCC+DBN, MFCC+DBN, and the proposed UAC are presented in Fig. 6, where the X-axis denotes the 8 RBM structures used in DBN (the label '1∼8' corresponds to the aforementioned 8 structures of RBM for LPCC/MFCC+DBN and the proposed UAC, respectively) and the Y-axis represents the 11 urban acoustic classes (the label '1∼11' corresponds to the datasets listed in Table 1 from top to bottom), respectively. For each class, it is observed that the affection of RBM structures used in DBN is small. For instances, using LPCC+DBN, the classification rates of excavator and hydraulic hammer are generally within 40%–65% and 65%–76% w.r.t. different RBM structures. For MFCC+DBN, the classification rates of these two categories are within 40%–50% and 70%–78%, respectively, except for the first structure in RBM, which has a very worse performance with the classification accuracies low to 9.25% and 47.63% correspondingly. Similar observation that the variation of RBM structure has little impacts on the classification performance is also found in the proposed UAC. However, comparing with LPCC/MFCC+DBN, the proposed UAC offers a more reliable and convincing performance, where the classification rates of 11 datasets are all higher than 92% under all tested RBM structures.

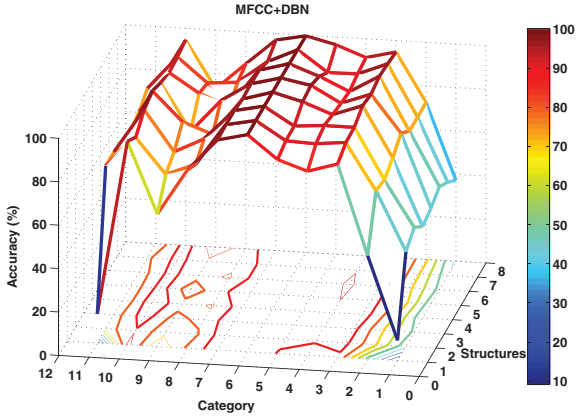
Besides the studies on the UAC performance with different RBM structures, different training iterations adopted in RBM are also presented. It is well known that too many epoches in training may cause the overfitting issue while too few epochs are unable to reach the optimal model. In this study, three layers of RBM are used in DBN with the fixed hidden node sizes as '1500-750-350'. The experiments on 6 different learning epochs 1, 5, 10, 20, 30, 50, in RBM are carried out while the rest parameters, including the mini-batch size, the initial learning rate, the initialization methods of weights and biases, etc., remain the same to the above experiment. The concatenated transfer learning features obtained by the three pre-trained DNN's are used for the study. Fig. 7 draws the accuracy curves on different epochs for all 11 urban acoustic datasets. It is found that for some datasets, such as cutting machine, horns, music, sound of generator, the accuracy variations are generally small when different epochs are used in RBM. But for some datasets, the affection of learning epochs is very obvious. For example, the accuracy of talking suffers a significant reduction when the epochs are increasing and the accuracy of engine sound experiences an apparent drop when 20 epochs are used. But overall, the proposed algorithm reaches the highest average accuracy of all datasets when 10 epochs training are applied in RBM.

The second experiment shows the effectiveness of using concatenated transfer learning features from three deep CNNs over the one only using a single CNN. The three pre-trained CNNs, Inception-v3, ResNet152, and Inception-ResNet-v2, are adopted as the single feature extractor on the spectrograms of urban acoustic streams, respectively, and are compared in this experiment. For each spectrogram image, there are 2048, 2048, and 1536 dimensions of features are obtained by the three CNNs, respectively. Similar to the proposed UAC, DBN is employed as the classifier for feature learning and classification, where the RBM structures tested in the proposed UAC are also studied for the single CNNs based transfer learning features and the one with the highest classification accuracy is reported.

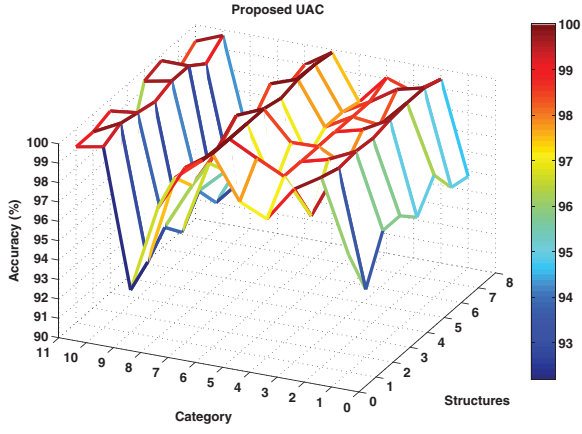
As shown in Table 3, using the concatenated transfer learning features generally offers better performance than using a single CNN. It is no doubt that the more generic features are extracted, the more solid classification performance of the proposed UAC can achieve. Particularly, the proposed UAC wins the highest classification rate on 10 out of 11 datasets,



(a) LPCC+DBN



(b) MFCC+DBN



(c) Proposed UAC

Fig. 6. Classification accuracy comparisons on LPCC+DBN, MFCC+DBN and proposed UAC with different RBM structures.

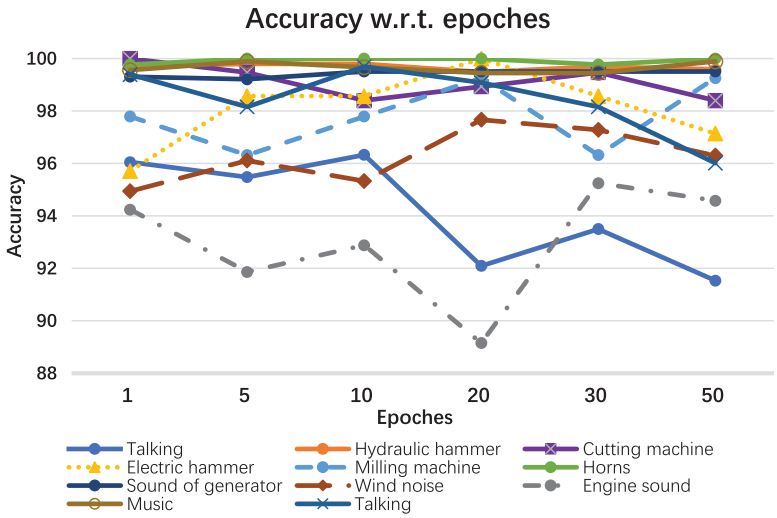


Fig. 7. Accuracy comparisons on different training epochs in RBM.

Table 3
Comparison with the transfer learning features by single CNN.

Datasets	Inception-v3 [28]	ResNet152 [29]	Inception -ResNet-v2 [30]	Proposed UAC
Excavator	92.94	88.42	86.44	**96.33**
Cutting machine	97.47	96.26	96.26	**98.40**
Electric hammer	97.14	97.14	95.71	**98.57**
Hydraulic hammer	99.18	99.59	99.39	**99.80**
Milling machine	96.32	**97.79**	94.12	**97.79**
Horns	**100**	99.54	99.77	**100**
Sound of generator	98.92	98.23	97.15	**99.51**
Wind noise	95.33	94.16	**95.72**	95.33
Engine sound	87.80	90.85	90.85	**92.88**
Music	98.68	98.03	98.68	**99.67**
Talking	97.85	95.40	97.55	**99.69**
Average	97.42	96.80	96.73	**98.55**

which on average, offers 1.13%, 1.73%, and 1.83% increments of classification accuracy than the Inception-v3, ResNet152, and Inception-ResNet-v2 feature extractor based algorithms, respectively. Some of the datasets, which have poor performance with the LPCC/MFCC representations mentioned in the first experiment, also suffer a relatively low classification rate when using the single CNN based feature extractor. For example, the accuracies of excavator with the ResNet152 and Inception-ResNet-v2 feature extractors are only 88.42% and 86.44%, respectively, far less than the proposed UAC. Similar observation can be found from the engine sound dataset.

The third experiment shows the comparison to the algorithm in different classifiers. These classifiers trained on the same features of urban acoustic streams to the proposed UAC. Table 4 lists the average classification rate on 11 urban acoustic datasets. As highlighted, DBN is more effective in feature learning and classification, which provides higher classification accuracy

Table 4
Comparison with the Fully-connected layer based classifier.

Datasets	H-ELM	ML-ELM	SAE+Softmax	Fully-connected layer [39]	Proposed UAC
Excavator	89.83	68.08	87.85	84.24	**96.33**
Cutting machine	**99.48**	98.78	99.08	98.75	98.40
Electric hammer	98.93	94.12	97.33	**99.52**	98.57
Hydraulic hammer	97.14	91.43	97.14	98.78	**99.80**
Milling machine	**97.79**	65.44	96.32	97.55	**97.79**
Horns	**100**	**100**	**100**	99.77	**100**
Sound of generator	**99.51**	97.35	99.21	97.19	**99.51**
Wind noise	95.13	85.02	93.77	86.66	**95.33**
Engine sound	91.19	73.56	88.81	92.35	**92.88**
Music	**99.78**	97.59	99.45	97.45	99.67
Talking	97.55	88.04	96.32	95.6	**99.69**
Average	97.82	91.73	97.05	95.56	**98.55**

Table 5
Comparison with PCA feature fusion based algorithms.

Datasets	PCA+DBN	PCA+SVM	Proposed UAC
Excavator	94.91	94.35	**96.33**
Cutting machine	**99.69**	**99.69**	98.40
Electric hammer	**99.46**	98.93	98.57
Hydraulic hammer	97.14	97.14	**99.80**
Milling machine	96.32	97.05	**97.79**
Horns	**100**	**100**	**100**
Sound of generator	99.50	98.92	**99.51**
Wind noise	93.49	93.57	**95.33**
Engine sound	92.54	92.66	**92.88**
Music	**99.89**	99.67	99.67
Talking	97.85	97.23	**99.69**

on 8 datasets than other classifiers. Meanwhile, the H-ELM suffers a relative poor performance on the excavator datasets, where this classification rate is 6.5% lower than the ones obtained by the proposed UAC. The ML-ELM suffers a relative poor performance on the excavator, milling machine, wind noise, engine sound and talking datasets, where their classification rates are 28.25%, 32.35%, 10.31%, 19.32% and 11.65% lower than the ones obtained by the proposed UAC, respectively. Similar results can be observed by other classifiers. For the classification accuracy of all datasets, on average, the proposed UAC offers 0.73%, 6.82%, 1.5%, and 2.99% increments, respectively, over other classifiers.

The last experiment shows the feature fusion performance on the deep transfer learning features. Particularly, PCA is adopted for feature reduction, and then SVM and DBN, are applied for feature learning and classification, respectively. The parameter setups of these two classifiers are the same to the above experiments. For PCA, the concatenated feature vector is reduced such that 90% information are preserved. Table 5 presents the average testing accuracies of PCA+DBN, PCA+SVM, and the proposed UAC on 11 datasets. It is found that in general, the proposed UAC outperforms PCA+DBN and PCA+SVM on most of the datasets. It is also observed that for the datasets of Cutting machine, Electric hammer, and Music, PCA not only reduces the feature dimension, but also enhances the classification

performance. For the rest datasets, the PCA based algorithms perform either the same to or slightly worse than the proposed UAC method, but win at a lower computational complexity.

4. Conclusions

In this paper, we proposed a novel UAC algorithm using deep CNNs based feature transfer learning and DBN based classification. Through the study, we have the follow observations: 1) The proposed UAC achieved the highest 98.55% classification rate on the real urban acoustic database; 2) The three pre-trained deep CNNs are effective in acoustic spectrogram feature extraction and outperform conventional LPCC/MFCC based methods; 3) The concatenated transfer learning features are more discriminative than the ones extracted with a single CNN; 4) DBN performs better than many compared state-of-the-art classifiers. For a real UAC system, the deep CNN based feature transfer learning is easy to be implemented and extended. However, the complex urban noise environment as well as the unpredictable interference presents challenges to the UAC algorithm. Future work will focus on: 1) the adaptive background noise filtering for performance enhancement, and 2) modeling the classification using only the target data with the one-class anomaly detection algorithms.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (61503104, U1509205, U1609218), by the State Key Program of National Natural Science of China (61333009), and by the Science and Technology Development Fund of Macao SAR (FDCT) under SKL-IOTSC-2018-2020, MoST-FDCT Joint Grant 015/2015/AMJ, and Grant FDCT/194/2017/A3.

References

[1] C. Asensio, Acoustics in smart cities, Appl. Acoust. 117 (2017) 191–192.
[2] P. Bellucci, L. Peruzzi, G. Zambon, LIFE DYNAMAP project: the case study of rome, Appl. Acoust. 117 (2017) 193–206.
[3] C. Mydlarz, J. Salamon, J. Bello, The implementation of low-cost urban acoustic monitoring devices, Appl. Acoust. 117 (2017) 207–218.
[4] P. Aumond, C. Lavandier, C. Ribeiro, et al., A study of the accuracy of mobile technology for measuring urban noise pollution in large scale participatory sensing campaigns, Appl. Acoust. 117 (2017) 219–226.
[5] A. Agha, R. Ranjan, W. Gan, Noisy vehicle surveillance camera: a system to deter noisy vehicle in smart city, Appl. Acoust. 117 (2017) 236–245.
[6] J. Ye, T. Kobayashi, M. Murakawa, Urban sound event classification based on local and global features aggregation, Appl. Acoust. 117 (2017) 245–256.
[7] K. Piczak, Environmental sound classification with convoltional neural networks, in: Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing, 2015, pp. 1–6.
[8] J. Cao, M. Cao, J. Wang, C. Yin, D. Wang, P. Vidal, Urban noise recognition with convolutional neural network, Multimed. Tools Appl. 78 (20) (2019) 29021–29041.
[9] C. Yin, X. Huang, S. Dadras, Y. Cheng, J. Cao, H. Malek, J. Mei, Design of optimal lighting control strategy based on multi-variable fractional-order extremum seeking method, Inf. Sci. 465 (2018) 38–60.
[10] C. Yin, S. Dadras, S. Huang, J. Mei, H. Malek, Y. Cheng, Energy-saving control strategy for lighting system based on multivariate extremum seeking with newton algorithm, Energy Convers. Manag. 142 (2018) 504–522.
[11] S. Ntalampiras, Universal background modeling for acoustic surveillance of urban traffic, Digit. Signal Process. 31 (2014) 69–78.
[12] R. Zhao, K. Qian, Z. Zhang, V. Pandit, A. Baird, B. Schuller, Deep scalogram representations for acoustic scene classification, IEEE/CAA J. Autom. SINICA 5 (3) (2018) 662–669.

[13] F. Eyben, F. Weninger, F. Groß, B. Schuller, Recent developments in openSMILE, the munich open-source multimedia feature extractor, in: Proceedings of the 21st ACM International Conference on Multimedia, Barcelona, Spain, 2013, pp. 835–838.
[14] J. Cao, W. Wang, J. Wang, R. Wang, Excavation equipment recognition based on novel acoustic statistical features, IEEE Trans. Cybern. 47 (12) (2017) 4392–4404.
[15] K. Piczak, ESC: dataset for environmental sound classification, in: Proceedings of the 23rd ACM international conference on Multimedia, Brisbane, Australia, 2015, pp. 1015–1018.
[16] J. Cao, T. Wang, L. Shang, X. Lai, C.-M. Vong, B. Chen, An intelligent propagation distance estimation algorithm based on fundamental frequency energy distribution for periodic vibration localization, J. Frankl. Inst. 355 (4) (2018) 1539–1558.
[17] L. Cai, J. Zhu, H. Zeng, J. Chen, C. Cai, K. K. Ma, HOG-assisted deep feature learning for pedestrian gender recognition, J. Frankl. Inst. 355 (4) (2018) 1991–2008.
[18] J. Duan, Financial system modeling using deep neural networks DNNs for effective risk assessment and prediction, J. Frankl. Inst. (2019), doi:10.1016/j.jfranklin.2019.01.046.
[19] M. Huzaifah, Comparison of time-frequency representations for environmental sound classification using convolutional neural networks, CoRR (2017). http://arxiv.org/abs/1706.07156
[20] R. Tak, D. Agrawal, H. Patil, Novel phase encoded mel filterbank energies for environmental sound classification, in: Proceedings of the International Conference on Pattern Recognition Machine Intelligence, Kolkata, India, 2017, pp. 317–325.
[21] V. Boddapati, A. Petef, J. Rasmusson, L. Lundberg, Classifying environmental sounds using image recognition networks, Procedia Comput. Sci. 112 (2017) 2048–2056.
[22] H. Sailor, D. Agrawal, H. Patil, Unsupervised filterbank learning using convolutional restricted Boltzmann machine for environmental sound classification, Proceedings of the Interspeech (2017) 3107–3111.
[23] J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (10) (2010) 1345–1359.
[24] J. Yosinski, J. Clune, Y. Benglo, How transferable are features in deep neural networks 2 (2014) 3320–3328.
[25] A. Razavian, J.S. Azizpour H., et al., CNN features off-the-shelf: an astounding baseline for recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 806–813.
[26] J. Cao, J. Zhu, W. Hu, A. Kummert, Epileptic signal classification with deep EEG features by stacked CNNs, IEEE Trans. Cognit. Dev. Syst. (2019), doi:10.1109/TCDS.2019.2936441.
[27] L. Xue, F. Su, Auditory scene classification with deep belief network 8935 (2015) 348–359.
[28] C. Szegedy, V. Vanhoucke, S. Ioffe, Rethingking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.
[29] K. He, X. Zhang, S. Ren, et al., Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[30] C. Szegedy, S. Ioffe, et al., Inception-v4, inception-resnet and the impact of residual connections on learning, in: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, 2017, pp. 4278–4284.
[31] J. Deng, R.S. Dong W., et al., Imagenet: a large-scale hierarchical image database, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.
[32] B. Igelnik, Y. Pao, Stochastic choice of basis functions in adaptive function approximation and the functional-link net, IEEE Trans. Neural Netw. 6 (6) (1995) 1320–1329.
[33] G.B. Huang, H. Zhou, X. Ding, R. Zhang, Extreme learning machine for regression and multiclass classification, IEEE Trans. Syst. Man Cybern. Part B: Cybern. 42 (2) (2012) 513–529.
[34] J. Cao, K. Zhang, H. Yong, X. Lai, B. Chen, Z. Lin, Extreme learning machine with affine transformation inputs in an activation function, IEEE Trans. Neural Netw. Learn. Syst. 30 (7) (2019) 2093–2107.
[35] T. Wang, J. Cao, X. Lai, B. Chen, Deep weighted extreme learning machine, Cognit. Comput. 10 (6) (2018) 890–907.
[36] H. Dai, J. Cao, T. Wang, M. Deng, Z. Yang, Multilayer one-class extreme learning machine, Neural Netw. 115 (2019) 11–22.
[37] X. Lai, J. Cao, X. Huang, T. Wang, Z. Lin, A maximally split and relaxed ADMM for regularized extreme learning machines, 2019, IEEE Trans. Neural Netw. Learn. Syst., In Press. doi:10.1109/TNNLS.2019.2927385.
[38] J. Yang, J. Cao, T. Wang, A. Xue, B. Chen, Regularized correntropy criterion based semi-supervised ELM, Neural Netw. 122 (2020) 117–129.
[39] L.D. Nguyen, D. Lin, et al., Deep CNNs for microscopic image classification by exploiting transfer learning and feature concatenation, Proceedings of the IEEE International Symposium on Circuits and Systems (2018) 1–5.

[40] J. Deng, N. Cummins, et al., The university of Passau open emotion recognition system for the multimodal emotion challenge, Chin. Conf. Pattern Recognit. 663 (2016) 652–666.
[41] A. Krizhevsky, I. Sutskever, G. Hinton, Imagenet classification with deep convolutional neural networks, in: Proceedings of the International Conference Neural Information Processing System, Nevada, USA, 2012, pp. 1097–1105.
[42] T. Akilan, Q. Wu, A. Safaei, W. Jiang, A late fusion approach for harnessing multi-CNN model high-level features, in: Proceedings of the IEEE International Conference on Systems, Man, Cybernetics (SMC), 2017, pp. 566–571.
[43] T. Akilan, Q. Wu, H. Zhang, Effect of fusing features from multiple DCNN architectures in image classification, IET Image Process. 12 (7) (2018) 1102–1110.
[44] C. Szegedy, W. Liu, et al., Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.
[45] Y. Bengio, G.D. Guyon I., et al., Deep learning of representations for unsupervised and transfer learning, Workshop on Unsupervised and Transfer Learning 7 (2011) 1–20.
[46] H. Lee, R. Grosse, et al., Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations, in: Proceedings of the 26th Annual International Conference on Machine Learning, 2009, pp. 609–616.
[47] J. Schmidhuber, Deep learning in neural networks: an overview, Neural Netw. 61 (2015) 85–117.
[48] E. Horster, R. Lienhart, Deep networks for image retrieval on large-scale database, Proceedings of the 16th ACM international conference on Multimedia (2008) 643–646.
[49] P. Hamal, D. Eck, Learning features from music audio, with deep belief networks, Proceedings of the 11th International Society for Music Information Retrieval Conference (2010) 339–344.
[50] T. Liu, A novel text classification approach based on deep belief network, in: Proceedings of the 17th international conference on Neural Information Processing: Theory and Algorithms - Volume Part I, 2013, pp. 314–321.
[51] W. Zhu, L.Q.J. Miao, et al., Hierarchical extreme learning machine for unsupervised representation learning, in: Proceedings of the International Joint Conference on Neural Networks, 2015, doi:10.1109/IJCNN.201507280669.
[52] L. Qi, J. Wang, X. Wang, Surface EMG signals based motion intent recognition using multi-layer ELM, in: Proceedings of the Society of Photo-optical Instrumentation Engineers Conference Series, 2017, doi:10.1117/12.2288037.
[53] Y. Chen, Y.L. Jiao, et al., Multilayer projective dictionary pair learning and sparse autoencoder for poISAR image classification, IEEE Trans. Geosci. Remote Sens. 55 (12) (2017) 6683–6694.