Audio-Based Vehicle Detection Implementing Artificial Intelligence

Oleg Golovnin[1(✉)], Artem Privalov[1], Anastasiya Stolbova[1], and Anton Ivaschenko[2]

[1] Samara University, 34, Moskovskoye Shosse, Samara 443086, Russia
golovnin@ssau.ru

[2] Samara State Technical University, 244, Molodogvardeyskaya Str, Samara 443100, Russia

Abstract. This paper presents a method for audio-based vehicle detection within the urban traffic flow analysis in Smart Cities. The proposed technology implements artificial neural networks to recognize and count vehicle sounds on audio recordings using mel-frequency cepstral coefficients. Nowadays there are a lot of different approaches for sound recognition but convolutional neural networks (CNN) have the greatest accuracy among the others. In this study, we compared CNN to a classic multilayer perceptron in the case of audio events recognition. The method was tested on the UrbanSound8K dataset and a mixed dataset combined by authors to be similar to actual conditions. Evaluation of possible intelligent solutions using the same UrbanSound8K dataset demonstrated that CNN have higher classification accuracy: 92.0% for CNN against 87.6% for multilayer perceptron. For the mixed dataset CNN presented the average vehicle detection accuracy of about 84.2%. Therefore, the proposed method allows simplification of traffic surveillance and reducing its costs and total information processing time.

Keywords: Traffic flow · Sound recognition · CNN · MFCC

1 Introduction

Nowadays one can say with confidence that the information technologies of the Smart City have changed a lot in the life of a city dweller; and these positive changes are becoming even more visible every day. Implementation of Smart City technologies in practice requires a deep analysis of various characteristics that describe the functioning of the urban environment, and, in particular, traffic flows as having the highest impact on the transport function and city infrastructure.

Online processing and analysis of the traffic flow characteristics are possible only with an advanced technical supply. Information about the current state of traffic flows is traditionally collected using various technical facilities, for example, loop detectors and radars. However, the recent increase in the number of video cameras on the urban streets makes it possible to use video records to analyze traffic flows. This approach has certain drawbacks since video recordings contain redundant information and require significant costs for data processing, storage, and analysis. It seems promising to use the

O. Dolinina et al. (Eds.): ICIT 2020, SSDC 337, pp. 627–638, 2021.
https://doi.org/10.1007/978-3-030-65283-8_51

audio signal instead captured by Smart City's video cameras or microphones, since the audio signal is less redundant and does not depend on visibility conditions.

The purpose of this work is the development and experimental testing of a method for detecting acoustic emission of vehicles in audio recordings as a part of the traffic flow analysis. The method is planned for implementation under the Smart City framework. The proposed method for detecting patterns of acoustic emission from vehicles uses mel-frequency cepstral coefficients (MFCCs) to identify the classification features using an artificial neural network. Two classes of neural networks are considered: multilayer perceptron (MLP) and convolutional neural network (CNN).

2 State-of-the-Art

For successful traffic management, it is necessary to determine the density of traffic using the technologies of video analysis [1, 2]. To solve the problem of detecting vehicles, lidar-based computer vision methods are often used [3]. As a rule, the proposed solutions include two stages [4]: research and evaluation of lidar methods, and then training and tuning of neural networks to improve detection quality.

Detection and monitoring of vehicles in real-time are one of the difficult problem domains where CNN demonstrate high efficiency and performance in the field of detection and identification of objects [5, 6]. In [7], there is conducted a review of CNN-based vehicle detection methods for monitoring a traffic situation. Note that CNN are used not only for the analysis of video images but also for aerial photographs [8].

A separate area for research is the task of detecting vehicles in video and images at night. So, in [9], a detection method is proposed, which is based on video and laser data processing. This technology uses the Gabor filter and the support vector method.

To analyze video records in order to determine traffic, a wavelet transform is implemented in [10]. Spectral analysis methods are used for pre-processing and post-processing of data for vehicles detection. For example, in [11], fast Fourier transforms and prescribed smart solutions enhance traffic detection of non-contact microwave radars. The wavelet transform is used for image preprocessing in traffic monitoring systems, where it highlights the characteristics of the vehicles [12].

Widespread methods for detecting vehicles in aerial photographs have many applications for vehicle monitoring [13] and urban planning [14, 15]. The analysis of aerial photos has certain difficulties because of the small size of the objects, the complex background, and the different orientation of the images. Effective methods are significantly different from the methods for detecting objects in images from the ground. In [12], to solve this problem, the authors propose a new CNN structure with double focal loss, and the authors of [16] propose a method for obtaining rotation-invariant descriptors.

The considered technologies use video data and images as initial data [17], which is often redundant and requires large resources to process. The use of audio analysis is a promising way to identify vehicles since it does not require expensive recording devices and a large amount of memory for data storage. In addition, such an important indicator as poor illumination does not affect the quality of data.

The work [18] demonstrates the possibilities of classifying vehicles by analyzing audio signals. The authors of [19] developed a system for detecting vehicles by the sound

signal received from two microphones located on the sidewalk. Analysis of audio signals based on a combination of frequency, time, and frequency-time characteristics is used to identify dangerous events on the roads, such as drifts, accidents, in conditions of poor visibility [20]. It is shown in [21] that for detecting events by the sound the combination of gammatone frequency cepstral coefficients and discrete wavelet transform coefficients can be used.

The estimation of the traffic flow by the audio signal is carried out using the support vector regression method in [22]. Assessment of the traffic by analyzing the total acoustic signal received from the smartphones based on a wavelet packet transform is given in [23].

3 Audio-Based Vehicle Detection Method

3.1 General Information

In order to efficiently and accurately recognize audio events that describe the appearance of a vehicle of various kinds, the following method was proposed, consisting of five steps, described below

The first step is the conversion of the original audio signal into a set of frames with overlapping.

The second step is pre-processing, which includes filtering and window weighing. This step is necessary for spectral smoothing of the signal. In this case, the signal becomes less susceptible to various noises arising during processing. In addition, during pre-processing, the audio is resampled, as well as its conversion to one channel. The bit depth is also normalized, so the values range from -1 to 1. This removes the complexity of data processing because different audio files can have different bit ranges.

The third step is to extract the necessary features. In the third step, the signal redundancy is reduced, the most relevant information is highlighted, and irrelevant information is eliminated. The patterns describing the audio signal are combined into one vector, on the basis of which further classification takes place. As symbols, it is proposed to use mel-frequency cepstral coefficients extracted from the audio signal. MFCCs summarize the frequency distribution according to window size, so we can analyze both the frequency and time characteristics of the sound. These audio presentations make it possible to identify the characteristics necessary for classification.

At **the fourth step**, the post-processing of the attributes occurs. After extracting the patterns of the signal for their further use, the patterns are normalized so that each component of the feature vector has an average value and a standard deviation of 1. Dimension reduction is used to significantly increase the speed and accuracy of the learning process, and the accuracy of machine learning algorithms, due to getting rid of an excess of patterns and highlighting significant patterns. It is proposed to use the principal component method at this step since it makes it possible to reduce the dimension of the feature vector by identifying independent components, which maximally covers the scatter for all events.

At **the last fifth step** of the proposed method, a training model is selected. For various types of audio events, it is worthwhile to select a specific classifier, because this

can provide a large increase in accuracy in the classification process. The paper discusses the use of a multilayer perceptron and a convolutional neural network as a classifier.

3.2 Calculation of Mel-Frequency Cepstral Coefficients

We detail some aspects of the presented method in part of MFCCs calculation. It is necessary to reduce the problem for N-numbers to the problem with a smaller number. For $N = \text{pq}, \text{p} > 1, q > 1$ over the field of complex numbers we introduce $\varepsilon_v = e^{2\pi i/v}$, $\varepsilon_v^v = 1$, where v is any number.

Discrete Fourier Transform is used:

$$b_i = \sum_{k=0}^{p-1} \sum_{j=0}^{q-1} a_{kq} + \varepsilon_N^{(kq+j)i} = \sum_{j=0}^{q-1} \varepsilon_N^{ij} \left(\sum_{k=0}^{p-1} a_{kq+j} \varepsilon_p^{ki} \right). \tag{1}$$

Next, each b_i is calculated:

$$b_i = \varepsilon^{-i^2} \sum_{j=0}^{N-1} \varepsilon^{(i+j)^2/2} \varepsilon^{-j^2/2} a_j. \tag{2}$$

The algorithm for obtaining MFCCs is constructed as follows: first we get the spectrum of the original signal ($x[n], 0 \leq n < N$):

$$X_a(k) = \sum_{n=0}^{N-1} x(n) e^{-\frac{2\pi i}{N} kn}, 0 \leq k < N \tag{3}$$

The resulting spectrum is displayed on a chalk scale. To do this, we use the windows located on the chalk axis:

$$H_{\text{m}}(k) = \begin{cases} 0, k < f(m-1) \\ \frac{\text{k}-\text{f(m}-1)}{(f(m)-f(m-1)}, f(m-1) \leq k < f(m) \\ \frac{(\text{f(m}+1)-k)}{(f(m+1)-f(m)}, f(m) \leq k < f(m+1) \\ 0, k > f(m+1) \end{cases}. \tag{4}$$

The frequencies f(m) are obtained from the equality:

$$f(m) = 700\left(10^{\frac{m}{2595}} - 1\right). \tag{5}$$

Next, we calculate the energy of each window:

$$S(m) = \ln\left(\sum_{k=0}^{N-1} |X_a(k)|^2 H_m(k), 0 \leq m < M\right), \tag{6}$$

where M is the number of filters we want to get. After that, a discrete cosine transform is applied to obtain a set of MFCCs:

$$c[n] = \sum_{m=0}^{M-1} S(m) \cos\left(\frac{\pi n(m+0.5)}{M}\right), 0 \leq n < M. \tag{7}$$

3.3 Artificial Neural Networks Specification

As mentioned above, we use a multilayer perceptron and a convolutional neural network as a classifier. Such a comparison is needed to conclude how the choice of classifier affects the result of the method. The MLP was chosen because it is a classical architecture and is the first choice as a network architecture when a new area of problems is considered for the solution of which a neural network can be applied. The CNN architecture is proposed as an alternative to the MLP, because the CNN architecture is widely used in the task of classifying images, and this area is adjacent to the problem that is considered in this paper.

The architecture of the used MLP consists of three layers: input, hidden, and output layers. The activation function is ReLU because it is the best choice for neural networks with a similar architecture:

$$f(x) = \max(0, x) \tag{8}$$

The CNN is organized in three dimensions: width, height, and depth. The vertices in one layer are not necessarily connected to all the vertices of the next layer. The model optimizer is Adam. The ReLU function is also used as an activation function for convolutional layers. The activation function in the output layer is Softmax. The function converts a vector z of dimension K into a vector σ of the same dimension, where each coordinate σ_i of the resulting vector is represented by a real number in the interval [0,1] and the sum of the coordinates is 1.

The coordinates σ_i are calculated as follows:

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{k=1}^{K} e^{z_k}}. \tag{9}$$

4 Results

4.1 Dataset Preprocessing

For testing the proposed method, the Urbansound8K data set is used [24]. The UrbanSound8K dataset consists of 8732 short (less than 4 s) fragments of city sounds, which are divided into 10 classes, while the class labels are not balanced. The training and test samples consist of.wav files and metadata describing them, stored in a .csv table.

Each sample represents the amplitude of the wave in a particular time interval, where the depth in bits determines how detailed the sample is. Therefore, the data that we will analyze for each sound fragment, in fact, is a one-dimensional array or a vector of amplitude values.

First, buffering occurs with overlapping in the source file. The following is the signal preprocessing. To convert the data into spectrogram representations, we used the LibROSA library [25], which is an open-source package implemented in Python.

Figure 1 shows an example of the shape of an audio signal passing by car, and Fig. 2 – of a truck.

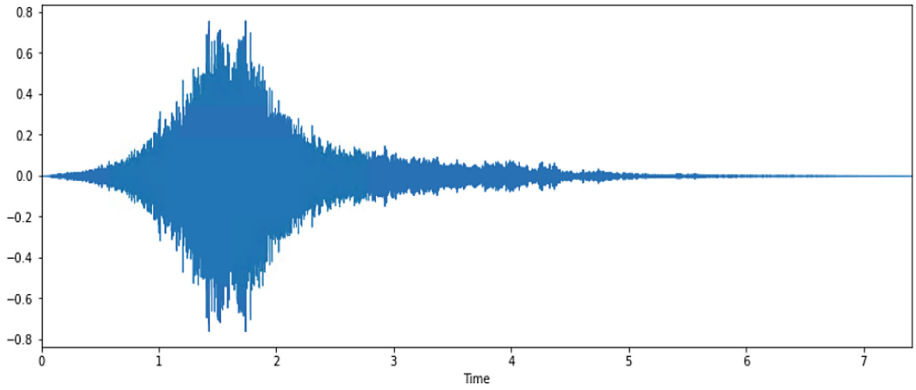

Fig. 1. Acoustic emission of a car

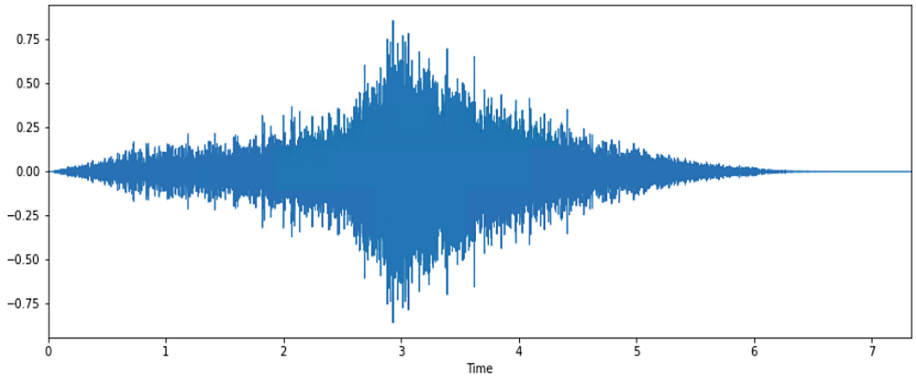

Fig. 2. Acoustic emission of a truck

From the visual analysis, it is clear that it is difficult to identify the difference between the classes of vehicles. Also, the shape of the car has similarities with street music, and the sound of children playing.

It is also worth noting that most instances of the sample have two audio channels (stereo sound), although some have only one audio channel. The easiest way to avoid this problem is to combine the two channels into one, by averaging the two channels.

4.2 UrbanSound8K Dataset Study

To compare the efficiencies of MLP and CNN architectures, we trained the networks with these architectures on the same data set. After training, the classification accuracy was measured and the neural network was selected with the greatest accuracy.

The MFCCs is used as a feature extraction tool. First, we extract the MFCCs from the instances for each frame with a window size of several milliseconds. MFCCs summarize the frequency distribution according to window size, therefore, both frequency and time characteristics of sound can be analyzed. A similar representation of audio identifies the

characteristics for classification. The following is the post-processing of the received characteristics.

Figure 3 shows the spectrograms obtained without and with scaling coefficients for a car, and in Fig. 4 – for a truck.

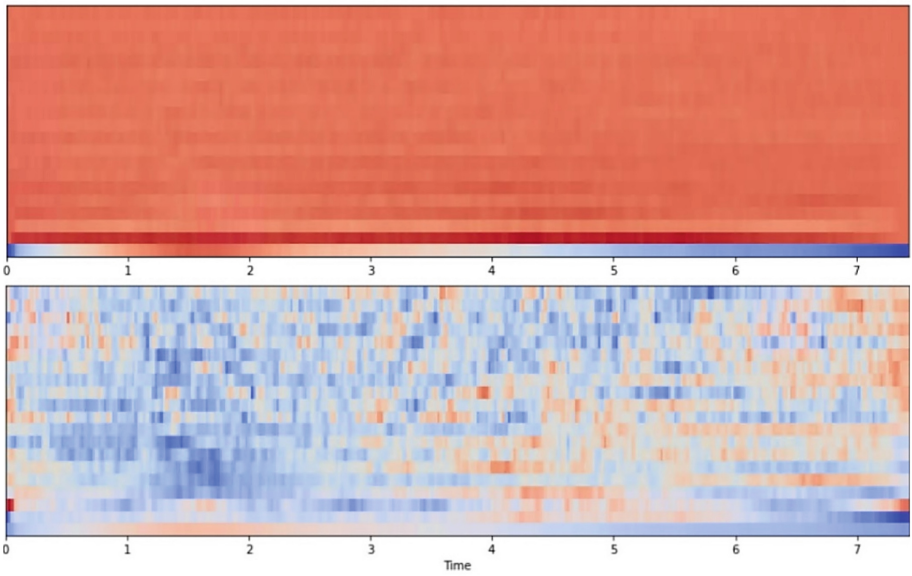

Fig. 3. Spectrogram for a car

Multilayer Perceptron. To begin with, we considered a simple structure of a neural network – MLP. The construction of the MLP is performed using Keras and TensorFlow.

The first layer is the input. Each sample contains 40 MFCCs, so the layer has a shape of 1×40. The first two layers have 256 neurons, and the activation function is ReLU. The exclusion level set to 50% to regularize the neural network during training, which leads to obtaining a network with better predictions.

The output layer has 10 neurons that correlate with the number of feature classes in the data set. The activation function for this class is Softmax. Softmax makes the output sum close to 1, so the output values can be interpreted as probabilities. Then the model will make its forecast based on which option has the greatest probability.

The results of testing the recognition accuracy of the MLP in the test and training samples are 87.6% and 92.5%, respectively. The accuracy is quite high, as well as a slight difference between samples (about 5%). Thus, the results show the MLP was not retraining.

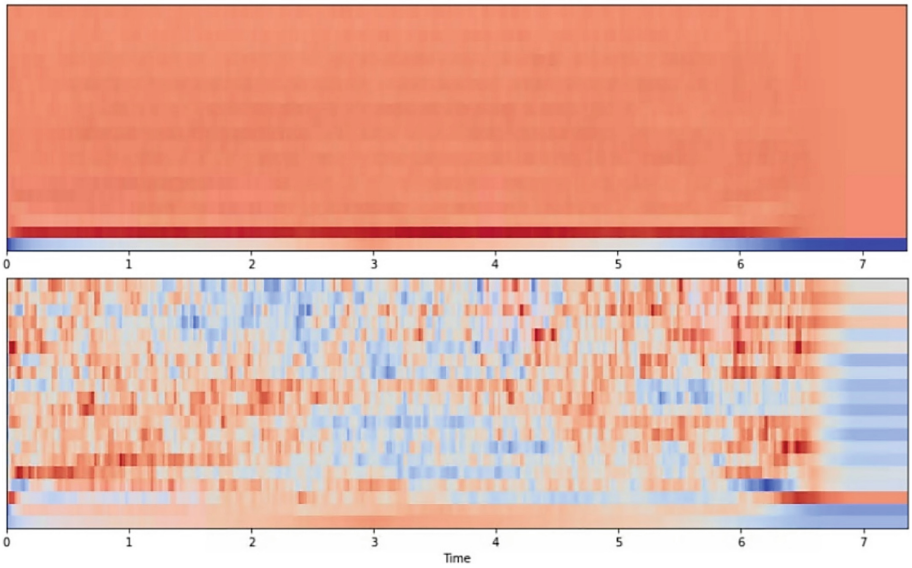

Fig. 4. Spectrogram for a truck

A graph of the classification accuracy versus the number of epochs is presented in Fig. 5.

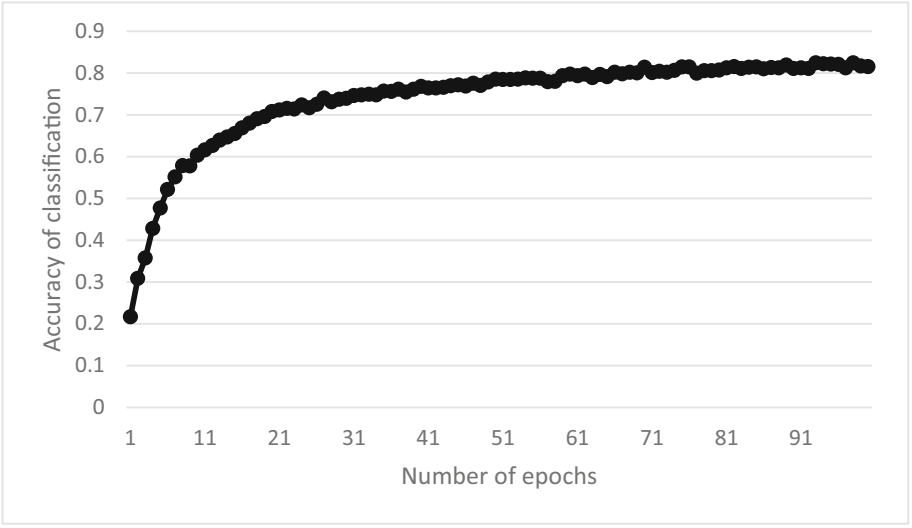

Fig. 5. The accuracy of classification vs. the number of epochs (MLP)

Convolutional Neural Network. MLP training has shown quite a good result. However, it is worthwhile to find out if another network architecture can achieve even higher

accuracy. CNN shows a good result in the classification of images, so this architecture was chosen to verify the assumption.

Since the CNN requires that the number of inputs be equal, we zero out the vectors so that they all become the same size.

The network model also remains consistent, with 4 convolutional layers and a dense output layer. The number of filters for convolution filters is defined as 16, 32, 64, and 128. The size of the core is 2x2 because the window size, in this case, is 2.

The first layer takes the form 40, 174 by 1, where 40 is the number of MFCCs, 174 is the number of frames, taking into account the filling, and 1 is the number of channels. The activation function for convolutional layers is, as in the previous model, ReLU, and the exclusion level set to 20%.

Each convolutional layer has a 2-by-2 sub-sampling layer. All of the subsampling layers are connected to one with the averaging layer, which supplies the averaged data to the output layer. Sub-sampling layers can reduce the dimension of the model (by reducing the number of parameters and sequential calculations), which leads to a decrease in training time and reduces retraining. The output layer is identical to the output layer of the MLP.

Figure 6 shows a graph of the classification accuracy versus the number of training epochs.

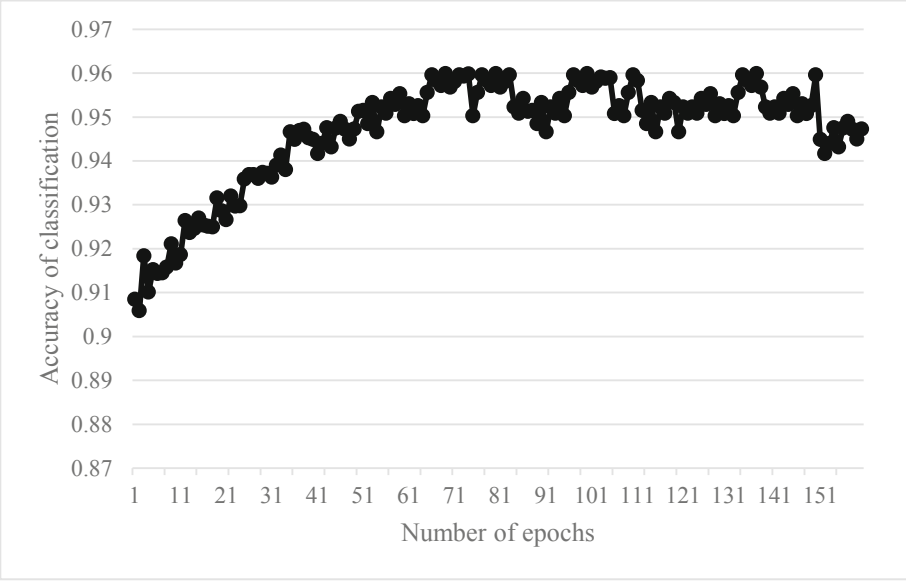

Fig. 6. The accuracy of classification vs. the number of epochs (CNN)

The results of testing the recognition accuracy of the CNN in the test and training samples are 92.0% and 98.2%, respectively. The accuracy increased by 6% for the training and about 4% for the test samples. Although the difference between them has

increased (up to 6%), this value is not so great that it is an indicator of the fact that during training the CNN was not retraining.

4.3 Mixed Dataset Study.

We tested the CNN on other data sets containing sounds of passing vehicles of varying quality and duration. Thus, we tried to simulate a real situation. Table 1 shows the parameters of the method when using trained CNN for a control set of data taken from open sources.

Table 1. Classification accuracy on open-source data set

Tittle	Number of vehicles	Number of classified vehicles	Classification accuracy, %
Record 1	68	65	92.5
Record 2	43	43	100.0
Record 3	10	10	100.0
Record 4	14	13	92.9
Record 5	33	30	90.9
Record 6	68	65	92.5
Record 7	22	18	81.8
Record 8	3	103	100.0
Record 9	55	50	90.9
Record 10	46	53	86.8

Table 2 shows the results of testing on the data recorded by the authors.

Table 2. Classification accuracy on authors' data set

Tittle	Number of vehicles	Number of classified vehicles	Classification accuracy, %
Personal record 1	15	13	86.7
Personal record 2	7	7	100.0
Personal record 3	33	30	90.9
Personal record 4	46	35	76.1
Personal record 5	18	9	50.0
Personal record 6	13	7	53.8

(*continued*)

Table 2. (*continued*)

Tittle	Number of vehicles	Number of classified vehicles	Classification accuracy, %
Personal record 7	18	15	83.3
Personal record 8	24	20	83.3
Personal record 9	68	55	80.9
Personal record 10	34	17	50.0

Thus, the resulting accuracy on the mixed data set (open-source and authors'), which is similar to actual conditions, was 84.2%.

5 Conclusion

In this paper, we proposed and investigated the method for patterns detection of vehicle acoustic emission in audio recordings using CNN and MFCCs. Studies conducted using MLP and CNN on the UrbanSound8k data set. The better results were presented by CNN that has higher classification accuracy: 92.0% against 87.6% by MLP.

In addition, a study was conducted on the authors' data set recorded, and mixed by audio files taken from open data sources. The resulting accuracy on the mixed data set using the CNN was 84.2%. A certain decrease in accuracy on the mixed set relative to the UrbanSound8k data set is due to the authors' audio recordings have overlapping of sound emission from several vehicles at the same time.

In additional, the results achieved on the classification accuracy (92.0% on the UrbanSound8k data set, 84.2% on the mixed data set) exceeds the classification accuracy of 73.5% achieved by the authors in the previous work [26], when MFCCs were not used.

As a result, the proposed method is recommended for Smart City applications to simplify the traffic surveillance and reduce the costs and total information processing time.

References

1. Swathy, M., Nirmala, P., Geethu, P.: Survey on vehicle detection and tracking techniques in video surveillance. Int. J. Comput. Appl. **160**(7), 22–25 (2017)
2. Ostroglazov, N., Golovnin, O., Mikheeva, T.: System analysis and processing of transport infrastructure information. CEUR Workshop Proceedings **2298**, 144071 (2018)
3. Li, B., Zhang, T., Xia, T.: Vehicle detection from 3d lidar using fully convolutional network. arXiv preprint arXiv 1608.07916 (2016)
4. Asvadi, A., Garrote, L., Premebida, C., Peixoto, P., Nunes, U.: Multimodal vehicle detection: fusing 3D-LIDAR and color camera data. Pattern Recogn. Lett. **115**, 20–29 (2018)
5. Bautista, C., Dy, C., Mañalac, M., Orbe, R., Cordel, M.: Convolutional neural network for vehicle detection in low resolution traffic videos. In: 2016 IEEE Region 10 Symposium, pp. 277–281. IEEE (2016)

6. Gao, S., Jiang, X., Tang, X.: Vehicle motion detection algorithm based on novel convolution neural networks. Curr. Trends Comput. Sci. Mech. Autom. **1**, 544–556 (2017)
7. Manana, M., Tu, C., Owolawi, P.: A survey on vehicle detection based on convolution neural networks. In: 3rd IEEE International Conference on Computer and Communications, pp. 1751–1755. IEEE (2017)
8. Qu, T., Zhang, Q., Sun, S.: Vehicle detection from high-resolution aerial images using spatial pyramid pooling-based deep convolutional neural networks. Multimedia Tools Appl. **76**(20), 21651–21663 (2016)
9. Zhang, R., You, F., Chen, F., He, W.: Vehicle detection method for intelligent vehicle at night time based on video and laser information. Int. J. Pattern Recognit Artif Intell. **32**(04), 1850009 (2018)
10. Golovnin, O., Stolbova, A.: Wavelet analysis as a tool for studying the road traffic characteristics in the context of intelligent transport systems with incomplete data. Trudy Spiiran **18**(2), 326–353 (2019)
11. Ho, T., Chung, M.: An approach to traffic flow detection improvements of non-contact microwave radar detectors. In: 2016 International Conference on Applied System Innovation, pp. 1–4. IEEE (2016)
12. Tang, Y., Zhang, C., Gu, R., Li, P., Yang, B.: Vehicle detection and recognition for intelligent traffic surveillance system. Multimedia Tools Appl. **76**(4), 5817–5832 (2015)
13. Razakarivony, S., Jurie, F.: Vehicle detection in aerial imagery: a small target detection benchmark. J. Vis. Commun. Image Represent. **34**, 187–203 (2016)
14. Audebert, N., Le Saux, B., Lefèvre, S.: Segment-before-detect: Vehicle detection and classification through semantic segmentation of aerial images. Remote Sens. **9**(4), 368 (2017)
15. Tang, T., Zhou, S., Deng, Z., Zou, H., Lei, L.: Vehicle detection in aerial images based on region convolutional neural networks and hard negative example mining. Sensors **17**(2), 336 (2017)
16. Ma, B., Liu, Z., Jiang, F., Yan, Y., Yuan, J., Bu, S.: Vehicle detection in aerial images using rotation-invariant cascaded forest. IEEE Access **7**, 59613–59623 (2019)
17. Peppa, M., Bell, D., Komar, T., Xiao, W.: Urban traffic flow analysis based on deep learning car detection from CCTV image series. In: SPRS TC IV Mid-term Symposium "3D Spatial Information Science–The Engine of Change", pp. 499–506. Newcastle University (2018)
18. Yang A., Goodman E.: Audio Classification of Accelerating Vehicles (2019)
19. Kubo, K., Li, C., Ishida, S., Tagashira, S., Fukuda, A.: Design of ultra low power vehicle detector utilizing discrete wavelet transform. In: Proceeding of ITS AP Forum, pp. 1052–1063. (2018)
20. Almaadeed, N., Asim, M., Al-Maadeed, S., Bouridane, A., Beghdadi, A.: Automatic detection and classification of audio events for road surveillance applications. Sensors **18**(6), 1858 (2018)
21. Waldekar, S., Saha, G.: Analysis and classification of acoustic scenes with wavelet transform-based mel-scaled features. Multimedia Tools and Appl. **79** 1–16 (2020)
22. Lefebvre, N., Chen, X., Beauseroy, P., Zhu, M.: Traffic flow estimation using acoustic signal. Eng. Appl. Artif. Intell. **64**, 164–171 (2017)
23. Vij, D., Aggarwal, N.: Smartphone based traffic state detection using acoustic analysis and crowdsourcing. Appl. Acoust. **138**, 80–91 (2018)
24. Dataset UrbanSound8k. https://urbansounddataset.weebly.com/urbansound8k.html. Accessed 04 Jun 2020
25. LibROSA, https://librosa.github.io/librosa/. Accessed Accessed 04 Jun 2020
26. Golovnin, O., Privalov, A., Pupynin, K.: Vehicle Detection in Audio Recordings by Machine Learning. In: 2019 International Multi-Conference on Industrial Engineering and Modern Technologies, pp. 1–4. IEEE (2019)