Application of Convolutional Neural Networks for Multimodal Identification Task

Anton Stefanidi, Artem Topnikov, Gennadiy Tupitsin, Andrey Priorov
P.G. Demidov Yaroslavl State University
Yaroslavl, Russia
antonstefanidi@mail.ru, topartgroup@gmail.com, genichyar@genichyar.com, andcat@yandex.ru

Abstract—**Currently, biometric identification systems are often used in mobile applications, banking systems, access control and management systems as well as for the management of mobile robots. In this paper, we consider the problem of personality recognition using facial images and audio signals with speech recordings. The results of the research will be used to create a system of multimodal biometric identification. Since convolutional neural networks demonstrate the highest results regarding the problems of detection, segmentation and classification of objects, this paper also proposes an approach to person identification based on convolutional neural networks. The research was carried out using modern audiovisual database VoxCeleb1. To decrease the computational capability of the experiment, the researchers reduced the number of classes from 1,251 to 200. The development results showed the possibility of using the proposed algorithm as a part of a multimodal identity identification system.**

I. Introduction

Currently, there are many approaches for personal identification and authentication, but the methods based on the analysis of biometric characteristics are the most effective. In particular, they, unlike passwords and tokens, cannot be stolen, lost or forgotten. These important properties contribute to further advanced use of biometric technologies [1, 2].

Most biometric systems are unimodal. Unimodal systems use a single source of biometric information. The choice of the type of biometric features determines the advantages and deficiencies of the identification system. For example, speaker recognition in noisy environments is extremely challenging. There is high variability of external and internal of the system parameters: background conversations, music, laughter, background vibrations, information channel and microphone effects as well as physiological features of the speaker, such as accent, emotions and intonation [4]. There is also an influence of external and internal factors in the task of user identification by a digital image: the level of illumination, the quality of the photosensitive sensor, the angle of inclination and rotation of the head, age-related changes, additional style elements such as glasses/beard/mustache, emotional activity and facial expressions. A biometric system can combine several different features to improve accuracy and compensate for deficiencies. These systems are known as multimodal [3].

This study is devoted to the development and research of personality identification methods based on the analysis of face images and audio signals. The discovered solutions will be used for creation of a multimodal biometric identification system. A block diagram of a bimodal biometric system is shown in Fig. 1. Face and voice recognition make it possible to obtain biometric parameters in the absence of physical contact of a person with the system, which expands the range of practical use of the proposed technology. The use of voice and facial images makes the system resistant to possible spoofing attacks, data fraud and unauthorized access attempts [3].

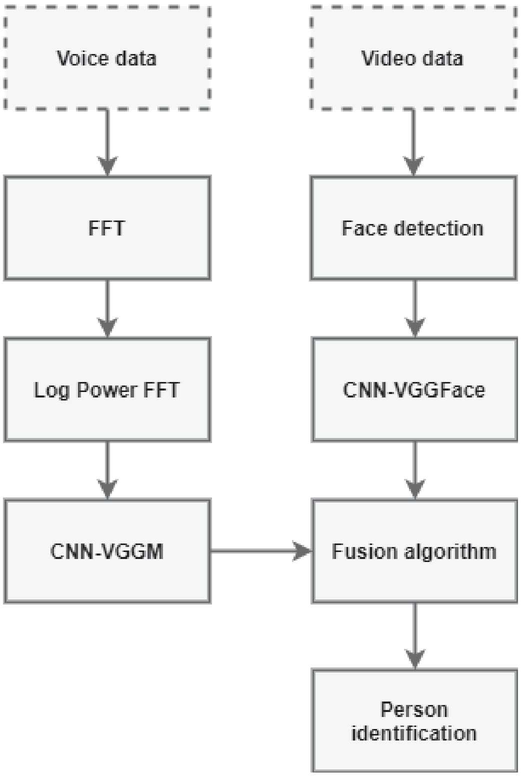

Fig1. A scheme of multimodal biometric identification system

Convolutional neural networks (CNN) have been a standard solution for face recognition problems for the past few years. Convolutional neural networks are the source that provides researchers with state-of-the-art results [1, 21, 22]. The use of mel-frequency cepstral coefficients (MFCC) and Gaussian mixtures models (GMM) was the standard solution for the task of text-independent speaker identification for many years [3, 5, 6]. The universal background model (UBM) [3, 6, 7], the joint factor analysis (JFA) [8], the total variability (TV) method [9] and the probabilistic linear discriminant analysis (PLDA) [10-12] were used to achieve greater robustness to changes in external conditions. However, the development of neural networks and deep learning systems has also affected the speaker recognition task [13-15]. In the recent years, researchers have been demonstrating the effectiveness of neural networks for speech recognition. The use of convolutional neural networks for speaker identification became possible due to the transformation of the audio signal into a two-dimensional digital signal. This is achieved by a transfer to the frequency area [5].

II. Audiovisual Dataset Description

A well-known VoxCeleb1 database was used for the experiment. It is an audiovisual dataset that consists of short segments of human speech and face images extracted from YouTube video interviews [16]. VoxCeleb1 contains over 150,000 audio samples for 1,251 celebrities that were extracted from videos uploaded to YouTube. The dataset is gender-balanced with 55% of male speakers and 45% of female speakers. This group of speakers represents a wide range of different ethnicities, accents, professions and ages. The nationality and gender of each speaker (obtained from Wikipedia) are also provided. The length of audio recordings for each celebrity varies from 4 to 145 seconds with an average length of 8.2 seconds (Fig. 2). Each person has from 45 to 250 audio samples with an average number of 123 recordings.

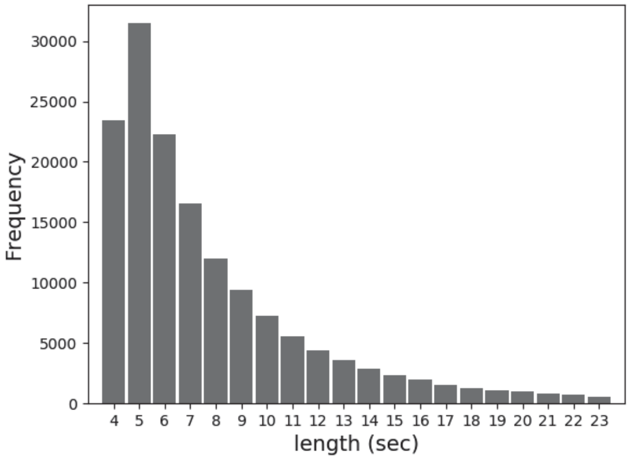

Fig. 2. A histogram of sound signals lengths from the VoxCeleb1 dataset

The fact that the research data was recorded in real-life conditions is especially important. The audio dataset examples were recorded in difficult acoustic conditions with video cameras and microphones of different specifications. These include red carpet, outdoor stadium, quiet studio interviews, speeches given to large audiences, excerpts from professionally-shot multimedia and videos shot on hand-held devices. Many audio signals contain real-world noises consisting of background chatter, laughter, overlapping speech and room acoustics [16]. The researchers note that this database is difficult for the task of speaker identification because audio tracks can contain fragments with overlapping speech, for example, when two people speak in parallel. These peculiarities make the learning process more difficult and can potentially reduce the precision of the model's performance. The researchers considered these additional noises and conditions quite common for everyday life. Therefore, it is important to develop a system that works in real conditions [17, 20].

The VoxCeleb1 database also contains a set of face images that were detected and saved during the processing of YouTube videos (Fig. 3). The total number of images is 1,217,558. This face database consists of color images. Pictures of faces are taken from different angles. Among other peculiarities, there are the rotation/tilt of the head, the color of face/hair, presence/absence of glasses/beard, mustache; different settings and the degree of illumination. As a result, the experiment conditions correlate with real-life conditions well.

Fig. 3. Examples of face images from the VoxCeleb1 dataset

The VoxCeleb1 dataset consists of face images and voice signals, and it is well-structured. As a result, researchers can create a multimodal identification system based on two biometric parameters: face and voice.

The number of defined classes was reduced from 1,251 to 200 classes for the purpose of decreasing the computational complexity and time of experimental research. The research was conducted quickly, and the researchers achieved great results in the classification of digital images and audio signals. Also, it is important to mention that the practical application of personality identification systems is narrowed to the problem of determining from a few dozens to several hundreds of objects. Table I shows the research part of the audiovisual database VoxCeleb1.

TABLE I. THE ANALYZED PART OF THE VOXCELEB1 DATASET

	Train	Cross-valid	Test	Total	Percent of all database	Classes
Images	91,331	11,417	11,314	114,062	9.36%	200
Audio signals	12,599	1,123	1,343	15,065	9.81%	200

III. Data Preprocessing and Neural Networks Architectures

The problem of facial recognition was solved using the pre-trained convolutional neural network VGGFace. This architecture shows high results regarding image classification problems [21, 22]. The researchers chose an implementation that was trained on VGG-Face dataset. VGG-Face is a database of celebrities' faces. The celebrities were chosen from the Internet movie database IMDB. The dataset consists of 2,622 classes and it has 2.6 million images [1]. Convolutional neural network VGGFace requires the input of scaled color images of size 224×224. To solve the problem of face recognition in real-life conditions, the upper layers of the pre-trained network were removed, and a new classifier was added. The classifier consists of two fully connected layers with 512 neurons and a ReLU activation function. A 200-dimensional softmax layer was used at the network output (Fig. 4). We chose Adam (adaptive amount estimation), as an optimization algorithm. It is a simple and computationally efficient algorithm for gradient-based optimization of stochastic objective functions. Also, it is robust and well-suited to a wide range of non-convex optimization problems in the field of machine learning. Adam was used with an initial learning rate of 0.001. The optimizer speed dynamically decreased by 0.9 times when a local minimum was detected [23].

Synthetic data augmentation was used to increase the number of images. The images were zoomed in on, rotated and reflected vertically and horizontally. The batch size was 128 images. The training lasted for 50 epochs, which is a small number by the standards of deep learning. The researchers did not want to train the network from scratch so they took the weights obtained from the network's analysis of the VGG-Face dataset. The time required for the network's training was greatly decreased.

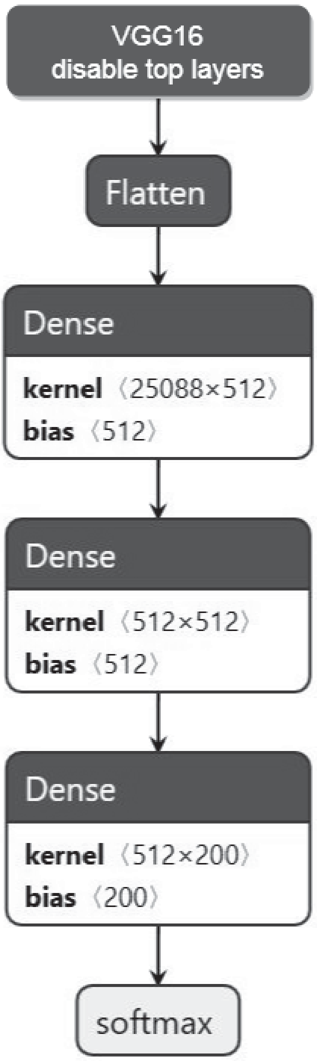

Fig. 4. The architecture of the CNN for face classification from the database VoxCeleb1

The audio data preprocessing was based on the use of the open-source librosa library. The audio signals were in WAV format with a sampling rate of 16 kHz and a quantization depth of 16 bits. A Fast Fourier transform (FFT) performed on librosa.core.stft implementation was used for the transition to a frequency area. Spectrograms were generated in a sliding window using a hanning window of width 64 ms and step 10 ms. Since the speech signals had different durations, the signals were reduced in lengths by randomly selected 4 seconds. If the original audio recording was shorter than 4 seconds, then it was enhanced with new added parts to make a recording of the average length. Therefore, the researchers got spectrograms of size 513x401x1 for each audio signal. Next, the researchers took the FFT module and converted the power spectrogram to decibels (dB) (Fig. 5). Peak power spectrogram was used as a normalization. The power spectrogram in decibels is described with the following formula:

$$S_{db} = 10 * \log_{10}(S) - 10 * \log_{10}(S_p),$$

where S is the power spectrogram, Sp is the peak power spectrogram and Sdb is the logarithm power spectrogram in decibels.

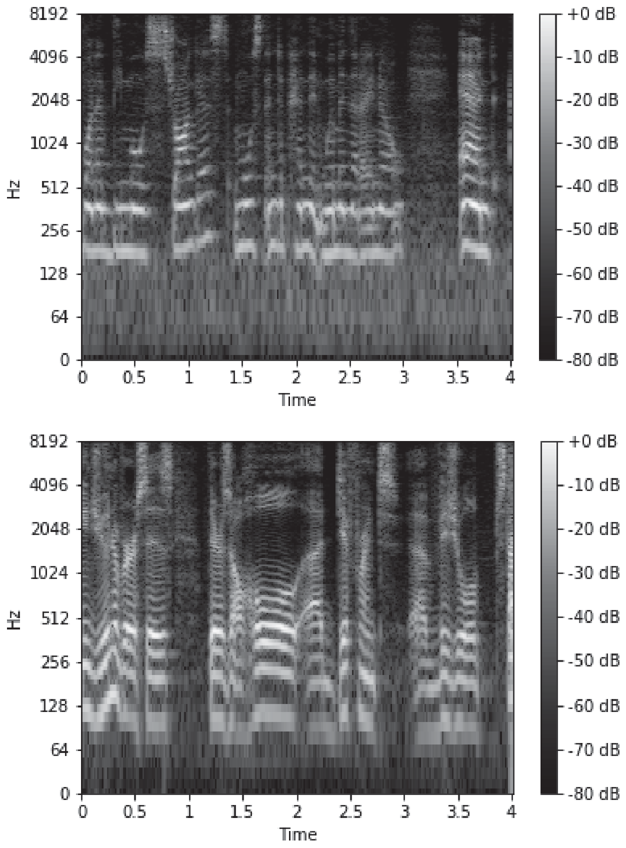

Fig. 5. Examples of the logarithm power spectrogram

The transition from time to frequency area translates the audio signal into a matrix representation. Spectrograms are often used in speaker classification and verification as an appropriate method for presenting an audio signal [15-17]. Also, the two-dimensional object type is well suited for working with convolutional neural networks. The CNN-VGGM was chosen to solve the speaker recognition problem

[16, 18, 19]. The Adam with a learning rate of 0.001 was used as the optimizer. The optimizer speed dynamically decreased by 0.9 times when a local minimum was detected. A dropout with a probability of 0.4 in fully connected layers was used for the regularization of the model during training. The batch size was 32. The number of training epochs was 200. In Fig. 6, the architecture of the convolutional neural network is represented.

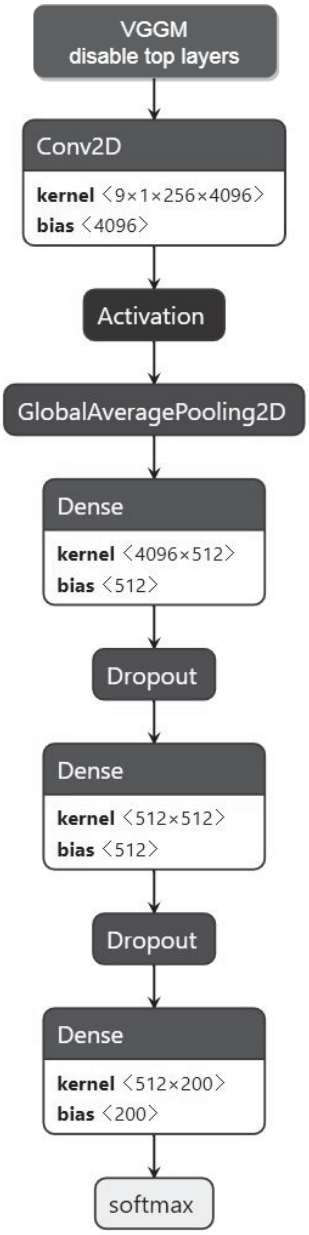

Fig. 6. The architecture of the CNN for speaker recognition from the database VoxCeleb1

For the processing of audio and video data and the training of neural networks, NVIDIA DGX-1 VOLTA, which is a supercomputer based in the AI-center of P.G. Demidov Yaroslavl State University, was used. This supercomputer delivers up to 960 TFLOPs. The internal TensorFlow implementation of the Keras was chosen as a framework. Netron was used to analyze the weights. It is a viewer for neural networks, deep learning, and machine learning models.

IV. Experiments and results

The metrics as accuracy (acc), precision (P), recall (R) and F1-score (F-measure, F-score) were used to analyze the learning process of convolutional neural networks. In Fig. 7, there is the CNN-VGGFace network learning process based on a set of face images VoxCeleb1.

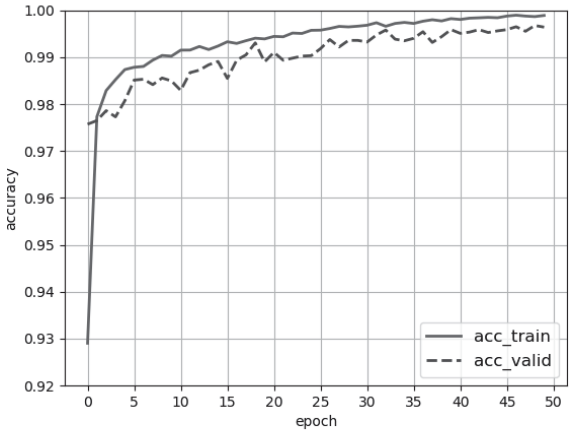

Fig. 7. Accuracy analysis in the CNN-VGGFace learning process

According to the results, the accuracy on the validation set is more than 97% at the first epoch of training. After training during 50 epochs, the classifier has an accuracy of 99.64%. The data accuracy on the test set is 99.57%. It is important to mention that there are no signs of overfitting or underfitting. The high accuracy on the test data is a great result. The results show that the neural network has a good generalization ability.

For the qualitative analysis of convolutional neural network performance by P, R and F1-score metrics, it is important to check the cross-validation set of images for shift within classes. Researchers developed a histogram for the cross-validation set, which shows the number of examples within each of the classes (Fig. 8).

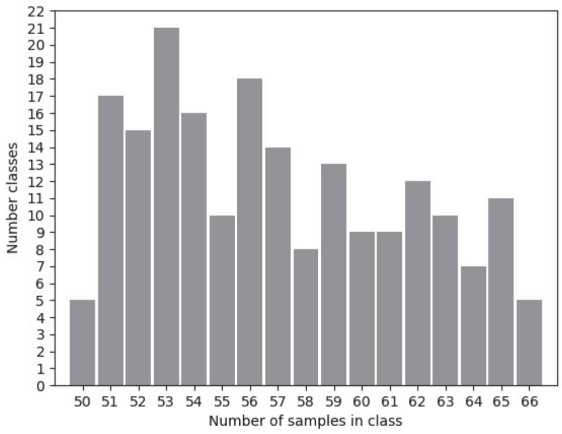

Fig. 8. A histogram for the number of examples in each of 200 classes

The results show that the number of examples in each of the classes varies from 50 to 66, and the average number is 57. There is no shift in cross-validation data. Therefore, researchers can describe P, R and F1-score metrics using a micro- or macro-averaging approach for the multiclass classification. In the current paper, the researchers used macro-averaging. Macro-averaging implies that the calculation of parameters P, R and F1-score within each class is done with subsequent averaging. The results for the above-mentioned metrics were the following: P=99.90%, R=99.87% and F1-score=99.88% [24].

The CNN-VGGFace network showed high results on the test set, but it was important to consider several factors. First, the VoxCeleb1 database contained images of subjects from the popular VGG-Face database. Second, we used a small part of the VoxCeleb1 database of 114,062 images for 200 subjects. We took these into account and compared our method with solutions base on a database with similar characteristics the IARPA Janus Benchmark A (IJB-A). The IJB-A set contains 5,712 images and 2,085 videos for 500 subjects. So in [25], neural networks ResNet50 and SENet have a top-1 in the range from 92.5% to 98.2%. These networks were pre-trained on large sets of face images VGG-Face, MSIM, and VGG-Face2.

It was previously mentioned that the researchers worked with a sample of only 200 classes. In the speaker identification research, the data samples were divided into three groups of training, cross-validation and test sets. The training set contained 12,599 audio recordings, the validation and test datasets contained 1,123 and 1,343 audio recordings. The categorical loss function (J(Q), Categorical Cross-Entropy Loss) in Fig. 9 was used to control the convergence of the model and exclude the chance of overfitting [26].

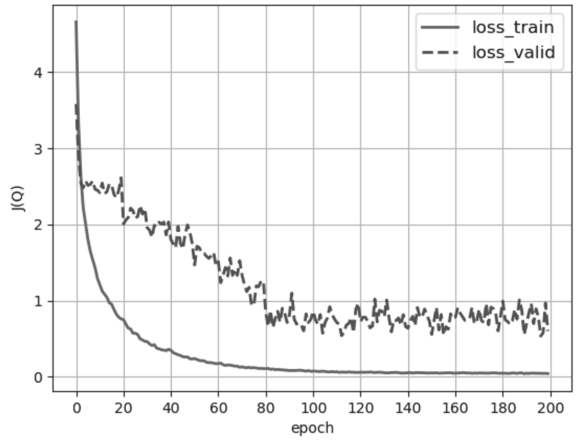

Fig. 9. Loss function analysis in the CNN-VGGM learning process

In Fig. 9, there is the learning process of a convolutional neural network where $\text{loss}_{\text{train}}$ and $\text{loss}_{\text{valid}}$ are losses on the training and validation set, respectively. The implementation of the model is correct as the J(Q) has a decreasing trend ending on a plateau of productivity. Moreover, during the training, an accuracy analysis of the cross-validation and training set was conducted (Fig. 10).

Fig. 9 and Fig. 10 show that the network recognizes training audio data with 98.91% accuracy. However, on a cross-validation set, this parameter is 78.87%. During the additional analysis of the validation samples, it was discovered that as the Voxceleb1 database was created, many videos were selected for each class, but fragments from a particular video clip were sorted into only one of the three sets. As a result, during the process of analyzing the validation and test sets the network analyzed examples containing noise and acoustic interference, which it could not meet in the learning process.

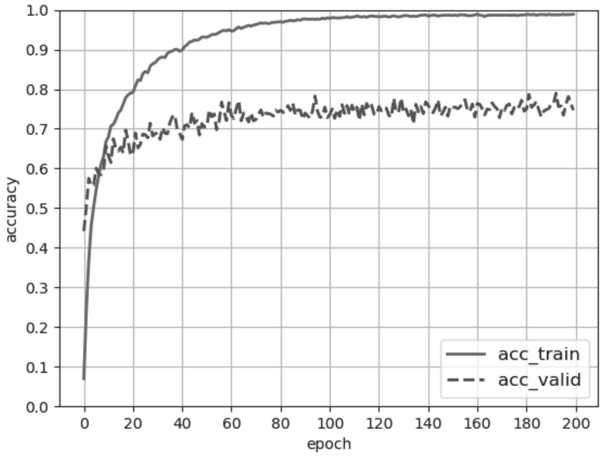

Fig. 10. Accuracy analysis in the CNN-VGGM learning process

In Table II, there are the results of the CNN-VGGM network for the training, cross-validation and test datasets. For the quality analysis of the algorithm, various types of accuracy are used: top-1, top-3, top-5.

TABLE II. SPEAKERS' IDENTIFICATION RESULTS FOR THE VOXCELEB1 BASE OF SPEECH SIGNALS

Accuracy	**Top-1 (%)**	**Top-3 (%)**	**Top-5 (%)**
Train	93.10	96.73	98.52
Valid	77.27	83.17	88.32
Test	74.65	82.40	86.97

CNN-VGGM shows acc-top-5 86.97% on the test set for the problem of classification of 200 classes from the VoxCeleb1 database. The score is quite high, and it can be compared with the results of state-of-the-art solutions for a VoxCeleb1 signal database. For example, in [16] a CNN shows acc-top-1 80.50% and acc-top-5 92.10% on the test set. Certainly, we took into account that we used a small part of the Voxceleb1 database.

To improve the quality of the CNN-VGGM model, researchers recommend to increase the dropout of fully connected layers or apply augmentation of audio data using various transformations and background noises. It is also possible to enhance the training dataset by using the full dataset VoxCeleb1 or a larger voice database VoxCeleb2. The VoxCeleb2 consists of over a million utterances from over 6,000 speakers. These recommendations will be taken into account when upgrading this CNN-VGGM network at the next research stage [16, 17].

In the present paper, audio signals were represented in the frequency field. This method is a classic solution. But currently, algorithms that can work with the raw waveform are gaining popularity. For example, the SincNet convolutional neural network was presented in [27] showed high results in speaker identification and verification tasks. The network analyzed raw audio data. SincNet's convolutional layers are represented as bandpass filters. This approach will also be considered in the next research stage.

V. Conclusion

The paper analyzes the personality classification with use different biometric parameters, such as voice and face. Convolutional neural networks of VGG-type architecture of various modifications were used to solve the task. The audiovisual database Voxceleb1 was chosen as a base of digital images of faces and speech signals. The number of defined classes was reduced from 1,251 to 200 classes for the purpose of decreasing the computational complexity and time of experimental research. The research was conducted quickly, and the researchers achieved great results in the classification of digital images and audio signals.

The CNN-VGGFace neural network, pre-trained on the large digital image database VGG-Face, was used to identify a person by face. During an experiment, the researchers used the technique of transfer learning with fine-tuning. Eventually, the new model showed great results on the test set: accuracy=99.57%, P=99.90%, R=99.87% , F1-score=99.88%.

The audio data set was represented by voice activity recordings. The CNN-VGGM network was used as the speaker recognition algorithm. The result for the acc-top-5 metric was 86.97% for audio signals classification from the test set. The results of the experiment represent high performance for database VoxCeleb1.

The next research stage is to improve the precision of audio signal recognition by using the following techniques: mel-frequency cepstral coefficients; modernization of the topology trained networks; changes in the regularization parameters; the use of a full dataset VoxCeleb1 or a larger speech signal database VoxCeleb2. In addition, the researchers will develop a multimodal solution based on speech and facial modality as well as the algorithms for fusion biometric systems.

Acknowledgments

The reported study was funded by Russian Foundation for Basic Research (RFBR), project number 19-37-90158. We would like to thank AI-center of P.G. Demidov Yaroslavl State University for providing access to the supercomputer NVIDIA DGX-1.

References

[1] O.M. Parkhi, A. Vedaldi, A. Zisserman, “Deep face recognition”, *In Proceedings British Machine Vision Conference*, 1, 41.1-41.12 10.5244, 2015, pp. 29-41.

[2] A. Lebedev, V. Khryashchev, A. Priorov, O. Stepanova, “Face verification based on convoluional neural network and deep learning”, *In Proceedings of 15-th IEEE East-West Design and Test Symposium (EWDTS 2017),* Novi Sad, Serbia, 2017, pp. 261-265.

[3] V. Khryashchev, A. Topnikov, A. Stefanidi, A.Priorov, “Bimodal person identification using voice data and face images”, *In Proceedings SPIE 11041, Eleventh International Conference on Machine Vision*, WEB: https://doi.org/10.1117/12.2523138.

[4] L.L. Stoll, *Finding difficult speakers in automatic speaker recognition*. Technical Report No. UCB/EECS-2011-152, 2011.

[5] D.A. Reynolds, T.F. Quatieri, R.B. Dunn, “Speaker verification using adapted Gaussian mixture models”, *Digital Signal Processing*, Vol.10, 2000, pp. 19-41.

[6] G. Tupitsin, A. Topnikov, A. Priorov, “Two-step noise reduction based on soft mask for robust speaker identification”, *In Proceedings 18th Conference of Open Innovations Association FRUCT*, 2016, pp. 351-356.

[7] T. May, S. van de Par, A. Kohlrausch, “Noise-Robust speaker recognition combining missing data techniques and universal background modeling”, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 20, No. 1, 2012, pp. 108-121.

[8] P. Kenny, “Joint factor analysis of speaker and session variability: Theory and algorithms”, *CRIM, Montreal, (Report) CRIM-06/08-13*, 2005.

[9] N. Dehak, R. Dehak, P. Kenny, N. Brümmer, P. Ouellet, P. Dumouchel, “Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification”, *In Proceedings INTERSPEECH*, 2009, pp. 1559-1562.

[10] S.J.D. Prince, J.H. Elder, “Probabilistic Linear Discriminant Analysis for Inferences About Identity”, *In Proceedings IEEE 11th International Conference on Computer Vision ICCV*, 2007, pp. 1-8.

[11] D. Garcia-Romero, C.Y. Espy-Wilson, “Analysis of i-vector Length Normalization in Speaker Recognition Systems”, *In Proceedings INTERSPEECH*, 2011, pp. 249-252.

[12] P. Kenny, “Bayesian Speaker Verification with Heavy-Tailed Priors”, *In Proceedings Odyssey,* 2010.

[13] O. Ghahabi, J. Hernando, “Deep Learning Backend for Single and Multi-session i-Vector Speaker Recognition”, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 25, no. 4, 2017, pp. 807-817.

[14] S.V. Ault, R.J. Perez, C.A. Kimble, J. Wang, “On Speech Recognition Algorithms”, *International Journal of Machine Learning and Computing*, Vol. 8, no. 6, 2018, pp. 518-523.

[15] S. Bunrit, T. Inkian, N. Kerdprasop, “Text-Independent Speaker Identification Using Deep Learning Model of Convolution Neural Network”, *International Journal of Machine Learning and Computing,* Vol. 9, no. 2, 2019, pp. 143-148.

[16] A. Nagrani, J.S. Chung, A. Zisserman, “VoxCeleb: a large-scale speaker identification dataset”, 2017, Web: https://arxiv.org/abs/1706.08612v2.

[17] J.S. Chung, A. Nagrani, A. Zisserman, “VoxCeleb2: Deep Spea-ker Recognition”, *In Proceedings Interspeech*, 2018, pp. 1086-1090.

[18] K. Simonyan, A. Zisserman, “Very deep convolutional networks for large-scale image recognition”, *In International Conference on Learning Representations*, 2015, Web: https://arxiv.org/abs/1409.1556v6.

[19] K. Chatfield, K. Simonyan, A. Vedaldi, A. Zisserman, “Return of the Devil in the Details: Delving Deep into Convolutional Nets”, *In Proceedings British Machine Vision Conference*, 2014, pp. 1–11.

[20] X. Xiang, S. Wang, H. Huang, Y. Qian, K. Yu, “Margin Matters: Towards More Dicsriminative Deep Neural Network Embeddings for Specker Recognition”, 2019, Web: https://arxiv.org/abs/1906.07317v1.

[21] Y. Sun, L. Ding, X. Wang, X. Tang, “DeepID3: Face recognition with very deep neural networks”, 2015, Web: https://arxiv.org/abs/1502.00873.

[22] Y. Taigman, M. Yang, M. Ranzato, L. Wolf, “Deepface: Closing the gap to human-level performance in face verification”, *In IEEE Conf. on CVPR,* 2014.

[23] Diederik P. Kingma, Jimmy Ba, “Adam: A Method for Stochastic Optimization”, 2017, Web: https://arxiv.org/abs/1412.6980v9.

[24] M. SokolovaN. Japkowicz, S. Szpakowicz, “Beyond Accuracy, F-score and ROC: a Family of Discriminant Measures for Performance Evaluation”, *In Proceeding of National Conference on Artificial Intelligence*, 2016, pp. 1-6.

[25] Q. Cao, L. Shen, W. Xie, O.M. Parkhi, A. Zisserman, “VGGFace2: A dataset for recognising faces across pose and age”, 2018, Web: https://arxiv.org/abs/1710.08092.

[26] W. Liu, Y. Wen, Z. Yu, and M. Yang, “Large-margin softmax loss for convolutional neural networks”, *In ICML*, 2016, pp. 507-516.

[27] M. Ravanelli, Y. Bengio, “Speaker Recognition from Raw Waveform with SincNet”, 2018, Web: https://arxiv.org/abs/1808.00158.