Research on Voiceprint Recognition Based on Convolutional Neural Network

Zhang Yipeng, Wang Zihao, Shang Yang

China University of Mining and Technology, Xuzhou, Jiangsu, 221116

Keywords: voiceprint recognition; deep learning; neural network; feature extraction

Abstract: With the advent of the information age, network security has become an issue that cannot be ignored at present, and identity authentication plays an indispensable role in it. Voiceprint recognition has attracted the attention of many scholars with its security and convenience, and has gradually become current research hotspots. Based on this, this paper researches voiceprint recognition based on deep learning convolutional neural network. Firstly introduces the structure of convolutional neural network, deepens its understanding, and then explains the related content of voiceprint recognition, including its definition, connotation, principle and development process. Finally, a voiceprint recognition scheme is designed based on the convolutional neural network. After verification, the scheme has high accuracy and can meet the current actual needs.

1. Introduction

In a rapidly developing information society, the need for identification is becoming more and more widespread, such as in the financial field and the investigation field. Traditional identification is mainly through personal items such as identity cards and passwords. Information leakage is very risky. With the development of Internet technology, people's lifestyles have also changed. The popularity of these online payments and social networks tied to personal information not only brings convenience, but also has hidden dangers [1]. The number of information leaks in the network is increasing, making traditional authentication methods no longer reliable. Therefore, identity authentication technology is also continuously progressing, and biometric technology has begun to enter the focus of researchers. Biometric technology is a technology that uses the biological characteristics of humans to identify individuals, including physiological characteristics and behavior characteristics. Physiological characteristics such as DNA, fingerprints, iris, facial recognition, etc; behavior characteristics include personal signatures, sounds, etc. These physiological characteristics are determined by human's innate physiological structure, because everyone's physiological structure is almost not exactly the same, so the biological characteristics extremely difficult to counterfeit [2].

2. Convolutional neural network structure

In recent years, with the development of computer hardware technology and the popularity of high-performance processors such as GPUs, ASICs, and FPGAs, deep learning technologies have achieved remarkable results in many fields. Convolutional neural networks are a multi-layer neural network. The layer is composed of multiple two-dimensional planes, and each plane is composed of multiple independent neurons. The network contains some simple and complex elements, which are denoted as S-elements and C-elements. S-elements are aggregated to form S-planes, S-planes are aggregated to form the S-layer, which is represented by us. There is a similar relationship between the C-element, C-plane, and C-layer (Us). Any intermediate level of the network is composed of the S-layer and the C-layers are concatenated, and the input stage contains only one layer, which directly accepts the two-dimensional visual mode. The sample feature extraction step has been embedded in the interconnected structure of the convolutional neural network model. Generally, Us is the feature extraction layer. The input of each neuron is connected to the local experience of the previous layer, and the local features are extracted. Once the local feature is extracted, the positional relationship between it and other features is also determined; Uc is the feature mapping layer. Each

computing layer of the network consists of multiple feature maps, each feature maps to a plane. The feature map structure uses a sigmoid function with a small influence function kernel as the activation function of the convolution network, so that the feature map has displacement invariance. In addition, because neurons on a mapping surface share weights, so the number of free parameters of the network is reduced, and the complexity of network parameter selection is reduced. Each feature extraction layer (S-layer) in the convolutional neural network is followed by a calculation for local average and secondary extraction layer (C-layer). This unique feature extraction structure enables the network to have higher distortion tolerance for input samples during recognition.

The output connection value of neurons in the network conforms to the "maximum detection hypothesis", that is, in a set of neurons existing in a small area, only the output neuron will strengthen the output connection value. So if there is a neuron near the neuron, when outputting a stronger neuron, its output connection value will not be strengthened. According to the above hypothesis, only one neuron will be strengthened. The seed of the convolutional neural network is the largest S-element output on a S-plane, which not only strengthens itself, but also controls the reinforcement results of neighboring elements. Therefore, all S-elements gradually extract the same features in almost all positions. It dominated the early research of convolutional neural networks. In unsupervised learning, it takes quite a long time to train a model to automatically search for the seed with the largest output among all the elements on a layer, and in the current supervised learning method, the training patterns and their seed all set by the teacher.

3. Voiceprint recognition

3.1 Definition of Voiceprint Recognition

A voiceprint is a spectral pattern drawn by a special electro-acoustic conversion instrument into sound wave characteristics. It is a collection of various acoustic characteristic maps. A voiceprint is a human body's "identity card" and a long-term stable characteristic signal. The pattern recognition is to draw the speech pattern of unknown person's speech material and known person's speech material by electro-acoustic conversion instrument, and then compare and comprehensively analyze the speech acoustic characteristics on the graph to obtain the judgment process of whether they are the same [3]. Voiceprint recognition has a very broad application prospect, and is being widely used in the fields of finance, securities, social security, public security, military, and other civil security certifications worldwide. At present, the Chinese market is still in its infancy and its development space is more expansive.

3.2 Voiceprint Recognition Connotation

Voiceprint recognition is broadly divided into speech recognition and speaker recognition. Speech recognition is based on the speaker's pronunciation to identify the spoken speech, syllable, word or single sentence, which requires excluding the personal characteristics of different speakers to find out the common characteristics of each speech unit. Speaker recognition is based on speech to identify the speaker, without regard to the content and meaning of the sound, which needs to separate the characteristics of each individual. At present, the concept of voiceprint recognition in a general sense refers to speaker recognition.

Speaker recognition includes speaker recognition and speaker confirmation. Speaker recognition is a one-to-many analysis process, that is, to determine which one of several people speaks a certain voice, mainly used in criminal investigation, criminal tracking, national defense monitoring, personalized applications, etc. Speaker confirmation is a one-to-one determination process, that is, confirming whether a certain piece of voice belongs to a specified person, mainly used in securities transactions, bank transactions, personal computer sound locks, car sound locks. The core of the recognition is to pre-record voice samples, extract the unique features of each sample, build a feature database, match the sound to be detected with the features in the database when using, and realize speaker recognition through analysis and calculation.

3.3 Principle of Voiceprint Recognition

3.3.1 Feature extraction

Feature extraction is the extraction of basic features that can reflect individual information. These basic features must be able to accurately and effectively distinguish different vocal individuals, and for the same individual, these basic features should have stability [4].

Current voiceprint recognition systems rely on lower-level acoustic features for identification. These acoustic features mainly have the following aspects:

(1) The speech information is output through the filter bank, and the output is sampled at an appropriate rate to obtain the spectral envelope characteristic parameters;

(2) Feature parameters extracted based on the physiological structures of vocal organs such as the glottis, vocal tract, and nasal cavity, such as the pitch contour, formant frequency bandwidth and its trajectory;

(3) Feature parameters derived from linear prediction, such as linear prediction coefficients, autocorrelation coefficients, reflection coefficients, etc .;

(4) Hearing characteristic parameters obtained by simulating the human ear's perception of sound frequency, such as Mel cepstrum coefficient, perceptual linear prediction, etc.

As the application scope of voiceprint recognition continues to expand, and the requirements for system accuracy continue to increase, considering only lower-level acoustic features cannot meet the requirements, it is necessary to consider high-level feature information such as speed of speech, grammar, prosody, language, dialect, characteristic pronunciation, characteristic words, channels (channels for sound signal acquisition), etc. For these high-level information, the most critical issue is selection, and at the same time, it should be determined for specific situations. For example, for channel characteristic is that in the criminal investigation case, that is, the channel will not affect the identification, so that the sound obtained by indirect means such as recording can be used as evidence to help solve the case; and in bank transactions, the channel is expected to have an impact on recognition, so that the harm caused by malicious acts such as recording can be eliminated. Therefore, in the voiceprint recognition process, a combination of different characteristic parameters must be arranged according to the actual situation to improve the actual system performance, when the correlation between the combined parameters is not large, a better recognition effect will be obtained.

3.3.2 Pattern matching

The key to voiceprint recognition technology is to process various acoustic feature parameters and determine the pattern matching method. The main pattern matching methods are:

(1) Probabilistic statistics method: Sound information is relatively stable in a short time. Through statistical analysis of steady-state characteristics such as pitch, glottic gain, and low-order reflection coefficient, statistics such as mean and variance and probability density functions can be used for classification. This method does not need to regularize the feature parameters in the time domain, and is suitable for text-independent voiceprint recognition [6].

(2) Dynamic time regularization method: speaker information not only has stability factors (structure of vocal organs and vocal habits, etc.), but also sometimes variable factors (speech rate, intonation, stress, rhythm, etc.). For comparison, the similarity between the two templates can be determined according to a certain distance [7].

(3) Vector quantization method: encode the specific text of each person into a codebook, and encode the test text according to this codebook during recognition. The distortion caused by quantization is used as the judgment standard. It has the characteristics of high recognition accuracy and fast judgment speed [8].

(4) Hidden Markov model method: Hidden Markov model is a random model based on transition probability and transmission probability. It regards speech as a random process composed of observable symbol sequences, and the symbol sequence is the the output of the sequence state of the vocalization system. During recognition, a vocalization model is established for each uttered individual, and the state transition probability matrix and symbol output probability matrix are

obtained through training [9]. During recognition, the maximum probability of unknown speech in the state transition process is calculated, and the corresponding maximum probability method does not require time regularization, which can save the calculation time and storage amount during judgment, and has been widely used; the disadvantage is that the calculation amount is large during training.

(5) Artificial neural network method: Artificial neural network simulates the perceptual characteristics of living things to some extent. It is a network model of distributed parallel processing structure. It has self-organizing and self-learning capabilities, and has a strong classification boundary distinction ability and robustness to incomplete information, its performance approximates an ideal classifier [10]; the disadvantages are long training time, weak dynamic time warping ability, and the network size may be too large to train when the number of speakers increases.

3.4 Development Process of Voiceprint Recognition

In the early stages of voiceprint recognition, human ear hearing recognition experiments were mainly performed. In the 1930s, researchers began to study the relevant fields of voiceprint. Initially, Bell Labs observed the speech spectrum of the sound to distinguish different speeches, and first proposed the concept of voiceprints. After that, researchers in the scientific field realized that there is a lot of space for identifying speakers using speech features, and began to use various methods to extract voiceprint features that help recognition. The process of distinguishing different speakers by comparing spectrograms is too complicated. Later, S. Pruzansky of Bell Labs proposed a recognition method based on probabilistic statistical analysis of variance and pattern matching. This scheme greatly improved the efficiency of voiceprint recognition. Since then, other researchers have begun to realize the possibility of automatic speaker recognition using machines, which has raised the research peaks of various research teams in the field of voiceprint recognition and gradually reflected in practical applications. In the 1960s, an American court used the voiceprint recognition technology to identify the suspect for the first time as a sentencing reference. The voiceprint was first used in the judicial field in 1660. Identification is the key evidence of the case, and the criminals in the death of Charles were convicted because of the voice of the speaker. Feature extraction is one of the key technologies of voiceprint recognition. After about two decades, most research focused on the extraction of feature parameters, Bogert et al. proposed the use of cepstrum for speaker recognition, improving accuracy. The concept of fast Fourier transform is proposed by Tukey and Cooly. So far, extractable parameters that can characterize speech features such as linear prediction coding coefficients, linear prediction cepstrum coefficients, Mel frequency cepstrum coefficients, etc. are widely used. Among them, MFCC parameters is the most effective and widely used feature parameter now, because its research is based on the human ear's auditory characteristics of speech, which is the same as the non-linearly perceived speech signal of the human ear. The principle of the MFCC parameter is to first convert the speech to non-based within the Mel scale of the linear scale, and then transform it into the cepstrum domain. On top of this result, the researchers found that using a mixture of multiple features for identification may get better results, such as using the Mel frequency cepstrum coefficient. In 2010, Hossan, Memon, Gregory, etc. proposed improved MFCC features, which combined three features of MFCC, first-order differential MFCC, and second-order differential MFCC. Later, researchers are no longer focusing on the extraction of feature parameters, but have gradually turned to the research of matching algorithms and model construction. After feature extraction, technical means are needed. The speaker builds a voiceprint model. The Gaussian mixture model proposed by MIT researchers Reynolds and Rose uses multiple Gaussian probability density functions to characterize the speaker model and obtains good results. Based on this, Reynolds proposed a more robust general background model greatly reduces the dependence of the training model on the target speaker data. Dehak and Kenny of the Montreal Institute of Canada found in the study that the channel factor contains speaker information, in order to reduce the influence on speaker's by channel pair, it uses joint factor analysis technology to deal with voiceprint features. Later, based on Gaussian mixture

model and joint factor analysis, a model that can eliminate the influence of the channel is called an identity vector. This model replaces other technologies and becomes one of the mainstream technologies with the best effect. Although China's research on voiceprint recognition is relatively late, it has developed rapidly and is now widely used in products. The voiceprint recognition system established by HKUST for the security field is in June 2008, he participated in the Speaker Recognition Contest SRE held by the National Institute of Standards and Technology NIST. The first score in the combined evaluation; companies such as Ali, Tencent, and Baidu have developed voiceprint recognition systems for their own business, such as WeChat, which can use voice to judge personal identity to log in; Alipay provides a voice lock, which can be protected by voice in 2015, the general manager of Baifubao demonstrated the research results of using voiceprints at the Global Mobile Finance Summit; the voiceprint recognition system demonstrated by Baidu in "The Strongest Brain" defeated "Listening Prodigy" competitors; In addition, the People's Bank of China has published voiceprint recognition application standards, which are widely used in mobile banking and third-party payment services. It can be seen that voiceprint recognition has broad development prospects and application prospects.

4. Voiceprint recognition based on convolutional neural network

In this paper, CNN-LSTM network combining CNN first and LSTM network is used to verify the voiceprint recognition. Here the network model is a whole, and the convolution pooling layer of CNN is added to learn the local spatial features of the spectrogram, and connected the LSTM network later to learn timing characteristics. The network structure consists of four layers, including the first layer of the convolution layer, the second layer of the pooling layer, the third layer of the LSTM network layer, and the fourth layer of the softmax classification layer. It can reflect the change of the speaker's speech spectrum with time at various moments. The spectrograms of different people contain speaker personality information, and the spectrograms are in the form of pictures. They are input to the deep learning network. The CNN network can better extract high characteristics of the form, so after the original speech is framed and windowed and fast Fourier transformed, a spectrogram that can be sent to the network is obtained. It has a 3-channel color image with a size of 106 * 80, which is the data dimension of the input network. It is 106 * 80 * 3. In this experiment, the number of speakers is 10, and the speaker labels are processed in the form of one-hot coding and input to the network in the form of a matrix. The model built in this paper is randomly disconnected 20% before training. Neurons are connected to prevent overfitting due to the large number of data dimensions and the small number of network layers. The number of convolution kernels of a CNN network is 20, and the size of the convolution kernel is 3 * 3, usually smaller convolution kernels can better make feature recognition. The activation function of the convolution layer is relu. The size of the pooling layer is 4 * 4, and the largest pooling method is selected, that is, the largest value in the range of 4 * 4 is selected as the new pooling layer features. The data after convolution pooling is sent to the LSTM network for time-series feature extraction. Because LSTM is a recurrent neural network, in order to prevent over-learning of the data in the network, it is divided between the recurrent neural units and between the loops. Dropout and Recurrent_dropout are added to temporarily disconnect a certain proportion of the connections between neurons in the same LTSM unit and the connections between different cyclic LSTM units. The proportion of disconnection in this article is 0.2, that is, 20% of internal nerves. The unit is disconnected from the external circulation unit. The softmax fully connected layer is finally connected to the network to identify the speaker. The number of people in the experiment is 10, so the number of classifications is 10.

In the actual application of voiceprint recognition in criminal investigation scenes, the data required for the identification of the suspect's voiceprint in each case is usually one or more voices, and there may be one or several suspects in the corresponding case. How to use the suspect voiceprint in the suspect determining identity among people is the focus of research. In this paper, the CNN-LSTM network with the best voiceprint recognition effect is selected for verification in actual noisy speech. The actual solution is as follows, collecting several suspects' noisy speech in

the real environment. First, the criminal voiceprint and the collected suspect's noisy voice are segmented into phrase sounds with a length of 4s, and then they are uniformly subjected to adaptive Wiener filtering to denoise, and then the noise-reduced voice is subjected to spectrogram feature extraction. After encoding the feature vectors of identity tags and spectrograms, the input words are embedded in the dimensionality reduction layer to reduce the dimensionality of the spectrogram. After the processing is completed, the CNN-LSTM network is used for training. After the network model is stable, the identified criminal's voice as a separate test set and obtain the identity of the person who belongs to the voice, then the corresponding identity person is the suspect. Of course, there are cases where the suspect does not include the real criminal. This requires the model to return the probability of the probability of each identity in practical applications. When the probability range is below a certain threshold, it is considered that there is no criminal in this group of suspects. In order to simulate voiceprint recognition applications in real scenarios, this section uses voice recording in a noisy environment, and confirms the accuracy of model recognition in this case through training tests. It also simulates the correctness of identity verification when the number of suspects is 2, 3, 5, and 10, and the above number settings are in line with the real identification requirements of criminal investigation suspects under the circumstances. The accuracy and robustness of the LSTM network based on word embedding proposed in this paper in voiceprint recognition research have been verified. On the basis of the standard data set, in order to verify the opinions presented in this paper, this article collects noisy speech. The collected voice samples are classmates of the author's laboratory. In order to simulate the impact of noise scenarios and channel differences under actual conditions, the recording device uses ordinary Android phones. The built-in recording software has collected the effective voice information of 10 classmates, each of which is usually about 9 minutes. The collection scene includes the natural environment in order to eliminate the change of voiceprint characteristics caused by the emotional differences in semantic content in speech samples, the content of different chapters in a dissertation is specifically selected to eliminate the impact of emotional differences. It also guarantees the differentiation of speech content. The speech data obtained after collection is processed in the same way as the standard data set. The speech material is divided into 4s fragments, each of which contains 120 phrases. Each speech sample needs to be carried out in the actual application scenario. The number of samples of voiceprint recognition is not fixed at 10, so when this happens, the parameter value of the final softmax classification layer in the network structure needs to be adjusted to meet the number of new recognition processes, and then the voiceprint verification process is performed. Through verification findings, the experimental results of the spatial and temporal characteristics of the voiceprint personality information are better than the traditional voiceprint recognition methods, and the accuracy is greatly improved. Therefore, it is decided to use the combined structure of CNN-LSTM network for the final practical data set verification: The actual data set assumes that the number of suspects is 2, 3, 5, 10, which is inconsistent with the standard. It is the last kind of identity is not exactly the same set of training, in order to make the network structure can adapt to the actual application requirements, there are two ways for reference.

Method 1: Modify the parameters of the deep learning network classification category in this method to be the same as the number of voiceprint recognition samples. This method does not need to add other voiceprint samples outside the verification, reducing the reason that the original voiceprint may be similar during the recognition process. The probability of recognition errors ensures that the original sample is pure and pollution-free, which increases the accuracy of recognition.

Method 2: Considering that it is impossible to equip professional computer personnel in practical applications, that is, it is impossible to change network parameters at any time as needed. To ensure the consistency of the number of participants participating in voiceprint verification, except for the actual participants, when the number of personnel is less than the number of network parameter settings. At the time, the method of increasing the number of samples is used to maintain consistency, that is, adding and supplementing the number of voiceprint samples is the same as the number of voiceprint samples. When the number of personnel exceeds the network parameter

setting, the sample of personnel who need voiceprint verification will be divided. The batch is performed, and the sample addition operation is performed when there is a mismatch in the number of the last batch. This method virtually increases the sample size, also consumes a large amount of computing storage capacity, and theoretically reduces the probability of accurate verification. But considering the professional staffing situation, this method is suitable for general operators, and can be applied to many aspects of promotion and application.

Through experiments in the standard data set, we can know that the recognition performance of the CNN-LSTM network is higher than that of the LSTM-CNN network, and further verify that the spatial feature signals in the voiceprint signal have a greater impact than the time-series feature signals. It can be guessed that with the number of suspects, the increase of the recognition accuracy rate after noise reduction will be significantly higher than the recognition accuracy of noisy speech. At the same time, after the adaptive Wiener filtering and noise reduction, the CNN-LSTM network based on word embedding spectrum reduction is analyzed to have high recognition accuracy which can meet the actual application requirements.

5. Conclusion

Modern society has higher and higher requirements for the intelligence of various industries, and more and more advanced artificial intelligence technologies have emerged as the times require. In the life of the rapid development of the Internet, voiceprint has a wide range of applications as a unique biological feature of human beings. It also has a wide range of applications in security and other aspects. With the advent of the deep learning era, convolutional neural networks have made great progress in the field of image recognition. Compared to traditional methods, CNN avoids the lack of manual feature extraction capabilities. In recent years, CNN has also been introduced in the field of speech recognition. Compared with the traditional method of extracting speech features, the spectrum map reduces the loss of information in the time and frequency domains. At the same time, due to the local connection and weight sharing of CNN, the CNN has translation invariance, so it can overcome the question of speech signal diversity.

References

[1] Zheng Yonghong. Development and application strategies of voiceprint recognition technology [J]. Science and Technology. 2017 (21): 103-105.

[2] Liu Su, Xu Lu.The application of voiceprint recognition technology in hearing people[J]. Communication World. 2017 (01): 159-161.

[3] Wang Yan. Talking about voiceprint recognition technology and security[J]. Network Security Technology and Application. 2017 (01): 78-81.

[4] Lu Yinan, Shan Baoyu, Guan Chao. Status and development of voiceprint recognition technology [J]. Information System Engineering. 2017 (02): 202-204.

[5] Zheng Fang, Li Lantian, Zhang Hui, Escal Rouzi. Voiceprint recognition technology and its application status [J]. Research on Information Security. 2016 (01): 134-136.

[6] Hu Qing. Research on application of convolutional neural network in voiceprint recognition[D]. Guizhou University. 2016.

[7] Zhang Zhizheng. Research and application of voiceprint recognition related technology[D]. Nanjing University of Aeronautics and Astronautics. 2016.

[8] Wang Zhengchuang. Research on voiceprint recognition system based on MFCC[D]. Jiangnan University. 2014.

[9] Yang Nan. Research and implementation of speaker recognition based on deep learning[D]. Zhengzhou University. 2019.

[10] Ding Dongbing. Research on intelligent voiceprint recognition[D]. Yangtze University. 2019.