Study on CNN in the Recognition of Emotion in Audio and Images

Bin Zhang
Hefei University of Technology
Hefei, China
zhangbinlh@yeah.net

Changqin QUAN
Kobe University 1-1 Rokkodai
Kobe, Japan
quanchqin@gmail.com

Fuji Ren
Hefei University of Technology
Hefei, China
ren@is.tokushima-u.ac.jp

Abstract—in this paper, the performance of Convolution Neural Network (CNN) in image recognition and emotion recognition in speech will be compared and presented. Feature extraction and selection in pattern recognition is an important issue and have been frequently discussed. Moreover, two-dimensional signals such as image and voice are hard to be modelled well by traditional models like SVM. The ability of CNN to characterize two-dimensional signals is prominent. And CNN can adaptively extract feature to eliminate the dependence on human subjectivity or experience. It mimics the effect of local filtering in visual cortex cells to dig local correlation in natural dimensional space. In this work, for the problems of the image recognition and emotion recognition in speech, CNN and SVM which is used as baseline for comparison of the recognition effect. Different kernel functions in SVM have been experimented for image recognition with, the best accuracy is 94.17%. However, the accuracy of using CNN is 95.5% (7291 pictures for train and 2007 pictures for test) with less time consuming. In the emotion recognition of speech, the accuracy of CNN is 97.6% corresponds to 55.5% by baseline model (4000 utterances for training, 1500 for validation, 500 for test). The experimental results showed that CNN can effectively extract features and its modeling capability for two-dimensional signals is prominent.

Keywords: Convolutional Neural Network, Emotional speech recognition, Image recognition.

I. INTRODUCTION

Image and voice are the most direct and most natural channels that people acquire information. If achievements in these two fields are applied on robots that can greatly improve the intelligence of the machine. In practice, in image recognition and speech recognition we will encounter the feature selection problem [1] [2]. Common image features are composed of color feature, texture feature, shape feature, spatial relations characteristics [3]. The commonly used features in speech recognition are MFCC (Mel Frequency Cepstrum Coefficient), prosodic features, sound quality characteristics and acoustic features [4] [5]. Sometimes in order to acquire a better final result, these characteristics also be integrated appropriately [6]-[8].

Selection and integration of these features require human experience and subjective judgment so the result on these feature sets are not good enough. Therefore, in order to better identify emotions or image a large number of features are added. In emotion recognition in speech, to identify the emotion in the voice, the linguistic information [9] or emotional point information [7] are also added, such as context [10]-[12], keywords, etc. Typically, this will cause improvements of the recognition rate. However, there is also the phenomenon that the information of voice and text interfere with each other to influence the judgment. Besides, emotion is an important aspect of intelligence. The problem that we want to make the system to distinguish emotion leads scholars have to resort to our own emotional cognitive system. Bionics, biology have been used to detect emotion in the voice [13]. They utilize the physical structure of we human ear to generate suitable features for human ears perceived characteristics, such as MFCC, Lyon cochlear model [14], or to improve the model to enhance recognition performance. These methods are often easier for people to trust and adopt for they can be better understood. Due to more fully take into account the physical structure of the human ear, mimicking the physiological activity, the adaptability and stability of the system built by this method are better than general system. Such a large feature set leads to the use of dimensionality reduction methods, such as PCA [15], Fisher and the like. But the final result is not satisfactory [4].

And for multi-class emotional speech signal, no classification tools is especially suitable for multi-class classification. Thus, optimized classification tools and perfected classification strategies [16] may improve the accuracy of emotion recognition. But experiments show that this strategy is not robust and there is limited room for improvement [4] [16].

In this study, traditional model SVM in the image recognition and speech emotion recognition is as a baseline system.

CNN is a specially designed multi-layer perceptron to identify two-dimension shapes. Therefore dimensional information retained in waveform points is effectively utilized by CNN. CNN model due to its characteristics of adaptive feature extraction, it is applied for image recognition and emotion recognition in voice signals. In the emotional speech recognition, based on the test of two classic characteristics of the speech signal, we propose that directly use waveform points to characterize the emotional speech signals. It neither loss information, but also take advantage of the natural correlation information between the waveform to identify emotion. In image recognition, SVM and CNN models are used for image

978-1-5090-0806-3/16/$31.00 copyright 2016 IEEE
ICIS 2016, June 26-29, 2016, Okayama, Japan

recognition. And we compare the recognition result before and after PCA

The rest of the paper is organized as follows. In Section II, we present the traditional model SVM and the newly test model CNN. Experiments and the results are provided in Section III. Finally, we summarize our findings in Section IV.

II. Tradition Model and CNN

Fig.1. Pattern Recognition Systems.

Fig.1 shows the flowchart of Pattern recognition. The feature represent the signals in abstract. And the classifier will map the feature and the output. So the choice of models affect a lot on the output.

A. SVM

The baseline model chosen in this experiment is SVM which act well in many cases in machine learning or pattern recognition. The input vectors which are hard to distinguish in original space are projected to high space to distinguish. And Kernel function type are linear kernel, Polynomial kernel, sigmoid kernel, RBF (Radial Basis Function). SVM is designed for classification of two types but its performance of the multi-classification is poor in actual problems.

When choose C-SVC model, the decision function is

$$\text{Plabel} = \text{sgn}(\sum_{i=1}^{n} \omega_i \cdot K(x_i, x) + b) \tag{1}$$

Where n is the number of support vector, $K(x_i, x)$ is the kernel function and b is a constant.

When RBF kernel function is selected, the kernel function $K(x_i, x)$ is

$$K(x_i, x) = \exp(-gamma \| x_i - x \|^2) \tag{2}$$

So the decision function in our problem is

$$\text{Plabel} = \text{sgn}(\sum_{i=1}^{n} \omega_i \cdot \exp(-gamma \| x_i - x \|^2) + b) \tag{3}$$

Similarly, the linear kernel $K(x_i, x)$ is

$$K(x_i, x) = x_i \bullet x \tag{4}$$

Polynomial kernel $K(x_i, x)$ is

$$K(x_i, x) = ((x_i \bullet x) + 1)^d \tag{5}$$

When selecting kernel function to solve practical problems, commonly used methods are as follows: First using a priori knowledge of experts to preselect kernel; and then using the Cross-Validation method, namely carrying a nuclear function selection and trying different kernels respectively. The kernel with the smallest induction error would be the best kernel.

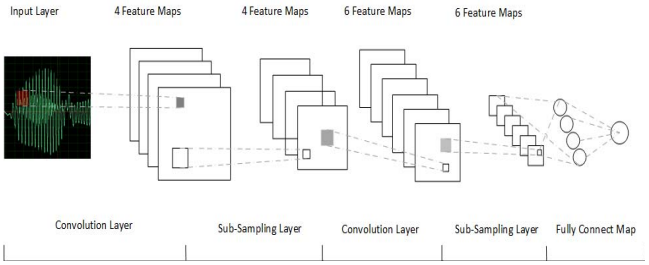

Fig.2.The structure of LeNet-5.

B. CNN

Convolutional Neural Network [17] is a multi-layer neural network, each layer is composed by a plurality of two-dimensional plane, and each plane is composed by a plurality of individual neurons. The structure of LeNet-5 is shown in Fig. 2[18].

Sparse, convolutional layers and max-pooling are the key of LeNet models [19]. From Fig.2 there are five layers in LeNet5: 2 convolution layers, 2 sub-sampling layers and 1 fully connected MLP layer. Alternating convolution and max-pooling layers are composed of the lower-layers. The upper-layers however are fully-connected traditional MLP (hidden layer + logistic regression).

B.1 Sparse Connectivity

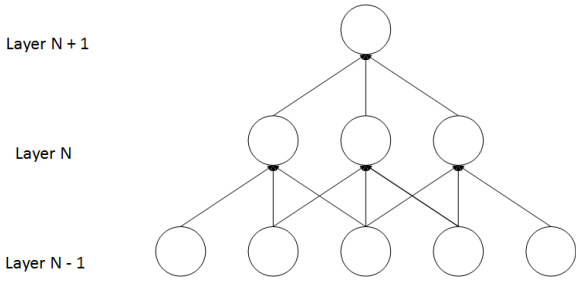

Fig.3. the schematic diagram of sparse connectivity.

The schematic diagram of sparse connectivity is shown in Fig.3.

The input of N-th layer is a subset of (N-1)-th layer. By strengthening a local connection between neurons in adjacent layers CNNs exploit spatially-local correlation. Such a structure is similar to a local filter which is able to generate the strongest response to the input pattern. However, if the increasing layers in the figure above it will lead to a non-linear filtering which responds a bigger space.

C. Image Recognition and Emotion Recognition in Speech

In pattern recognition perspective image recognition and emotion recognition in speech can be abstracted into the form in fig. 4.

Specific to the field of image recognition and emotion recognition in speech, the process can be summarized in fig.5.

As shown in Figure 4 and Figure 5, their methodological

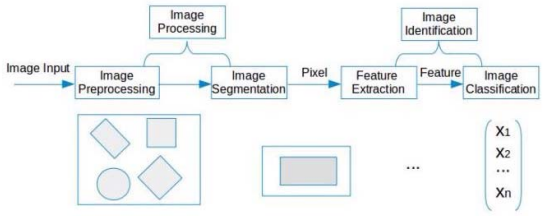

Fig.4. Four steps of image recognition.

framework are similar. They both go through feature extraction and screening, and these steps are critical to the entire recognition. Another key issue is that the ability of model to reconstruct signals. CNN used in this paper is able to solve the two key issues at the same time. The details will be shown in Section III.

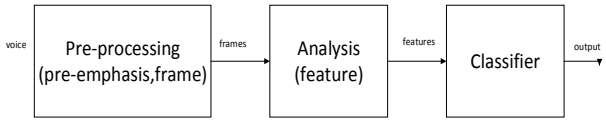

Fig.5.Traditional method of emotion recognition in speech.

III. EXPERIMENT

A. Data preparation

All experiments on images were conducted on USPS dataset (The US Postal handwritten digit dataset). It is a frequently used data set in machine learning, artificial intelligence, and data mining. This dataset (scaled to [-1:1] instead of [0:2]) also appears in the book "The elements of statistical learning"[20] [21]. The dataset consists of 7291 pictures for train and 2007 pictures for test.

All experiments on speech were conducted on the emotion speech database recorded by Chinese Academy of Social Sciences (CASS). There are four speakers including two females and two males. Five emotions were included, they are angry, fear, happy, sad, and surprise. Each emotion corresponds to 1200 different utterances, a total of 1200*5 utterances. The signals are sampled at 16 kHz and transcribed in mono. Each sampling point is represented with 16bit. The utterances are between one or two seconds to reserve prominent parts in emotions.

B. Experiment environment

CNN is set as Table 1 shown in emotion recognition in speech. Table 1 can be explained or understood in the following way.

Layer 0 (convolution layer + pooling layer): wave file is read and framed. After that, the sequence acquired by framing is reshaped to (100,100). CNN network is set as follows: wave_shape = (100,100), poolsize = (2, 2), and filter shape = (5, 5), filtering reduces the wave size to (100-5+1, 100-5+1) = (96, 96), maxpooling reduces this further to (96/2, 96/2) = (48, 48).

Layer 1 (convolution layer + pooling layer): the output of layer 0 is the input of layer 1.And wave_shape = (48, 48), filter_shape = (5, 5), poolsize = (2, 2). With the same method of calculation as layer 0, filtering reduces the wave size to (48-5+1, 48-5+1) = (44, 44), maxpooling reduces this further to (44/2, 44/2) = (22, 22).

TABLE I. SETUP OF CNN IN EMKOTION RECOGNITION IN SPEECH

Layer	Input
Input level	**100*100**
Pool size	**200 feature map, convolution window size : 5*5, pool window size: 2*2**
Full connection layer 1	**Hidden neurons:500**
Full connection layer 2	**Hidden neurons:500**
Output layer	**6 classification output**

Layer 2(Hidden Layer): the Hidden Layer is fully-connected, it operates on 2D matrices of shape. It is set as following: input = 50 * 22 * 22, output = 500.

Layer 3 (Logistic Regression Layer): it classify the values of the fully-connected sigmoidal layer. And the input of layer 3 is the output of layer2 so input=500, output=5. The output of layer 3 is the prediction labels.

CNN used in image recognition is set in table II.

And the baseline model SVM are applied with different kernel types: linear kernel, Polynomial kernel, sigmoid kernel, RBF (Radial Basis Function). With different kernels, the results are different and they are be shown in part C. Grid parameter optimization is applied by the baseline system. Pick out the best model parameters in the ten-fold cross-validation and used in the test set. In the process of parameter optimization, the ability of modelling varies a lot. And the parameters we got is especially for train data. We desire for a model with learning skills and it act well on any train data. Hence, CNN is applied for comparison in this experiments.

TABLE II. SETUP OF CNN IN IMAGE RECOGNITION

Layer	Input
Input level	**16*16**
Pool size	**200 feature map, convolution window size : 3*3 , pool window size: 2*2**

Full connection layer 1	**Hidden neurons:50**
Full connection layer 2	**Hidden neurons:100**
Output layer	**10 classification output**

C. Recognition results

As for image recognition, the results on baseline model SVM and newly applied model CNN are shown in table III and table IV.

TABLE III. RESULT ON IMAGE RECOGNITION WITH SVM AND CNN

Kernel type	**Results (%)**	
	SVM	**CNN**
linear kernel	**92.53**	**95.5** [a]
Polynomial kernel	**92.87**	
sigmoid kernel	**87.79**	
RBF	**94.17**	

a. kernel types means the kernel type of SVM, and they have no influence on CNN.

TABLE IV. RESULT ON EMOTION RECOGNITION IN SPEECH WITH SVM AND CNN

Feature type	**Results (%)**	
	CNN	**SVM**
MFCC	**36.6**	**46.6**
Prosodic Feature	**20.1**	
Wave points	**97.6**	

From the above table III, the effect of the baseline system SVM with RBF kernel function is the best. However, its results is 94.17% and not better than CNN with accuracy 95.5%. In addition, CNN avoids manual selection process for feature sets so that the result is more general in practical.

As for emotion recognition in speech, the result is even more amazing. It can be seen from the table that the waveform points directly applied as input for CNN which rely on adaptive modeling the results are surprisingly good (accuracy: 97.6%). This is in line with the characteristics of bionics in neural network: construct connection according to the phenomenon and learn connection weights in their own without human decisions, in other word the neural network determines what they can see in their own instead of us. It also explains why MFCC and prosodic feature are ineffective in recognizing emotions. If we cannot acquire features which is able to characterize the emotions in speech signals by priori knowledge or experiments, and directly apply the existing features regardless of its characterization capability for emotions, it is hard to say these features can effectively identify emotions.

IV. CONCLUSION

The traditional model in machine learning is weak in practice. In this paper, the CNN model is applied in image recognition and emotion recognition in speech. The speech waveform points are directly applied for optimized Convolutional Neural Network (CNN) to improve the performance in recognizing emotion in speech. And CNN is applied to identify images. For image recognition, recognition results on CNN are better than them on SVM baseline system with better robustness. As for emotion recognition in speech, by contrasting MFCC feature and prosodic feature, the result of waveform points is surprisingly good and reaches 97.6%.

ACKNOWLEDGMENT

This research has been partially supported by National Natural Science Foundation of China under Grant No.61472117, No. 61203312 and the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry.

REFERENCES

[1] Arimura, K., Hagita, N., "Feature space design for image recognition with image screening," in *Pattern recognition*, vol. 2, 1996, pp. 261–265.

[2] Lijiang Chen a, Xia Maoa, Yuli Xue, Lee Lung Cheng, "Speech emotion recognition: Features and classification models", in Digital Signal Processing, vol.22, 2012, pp.1154-1160.

[3] Dai Fang, He Haimei, Han Wei, "Integrating multi-feature of image based on correspondence analysis", in 2010 the 5th IEEE Conference on Industrial Electronics and Applications (ICIEA), 2010, pp.15-17.

[4] Changqin QUAN，Bin ZHANG, and Fuji REN, "Joined cepstral distance features two-stage multi-class classification for emotional speech", Presented at the 9th International Conference on Natural Language Processing and Knowledge Engineering, pp.91-96, 2014.

[5] Christer Gobl, Ailbhe Ni Chasaide:" The role of voice quality in communicating mood and attitude". Speech Communication 40, 2003, pp. 189–212.

[6] Chung-Hsien Wu, and Wei-Bin Liang, "Emotion Recognition of Affective Speech Based on Multiple Classifiers Using Acoustic-Prosodic Information and Semantic Labels", in IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, vol.2, No.1, 2011.

[7] Lijiang Chen, Xia Mao, PengfeiWei, Yuli Xue: "Mandarin emotion recognition combining acoustic and emotional point information", Appl Intell 37, 2012, pp.602–612.

[8] Jaekyong Jeong; Hyeongyong Jeon; Chijung Hwang; Byeungwoo Jeon, "Efficient Image Feature Combination with Hierarchical Scheme for Content-Based Image Management System", in Third International Conference on Multimedia and Ubiquitous Engineering, 2009, pp.539-545.

[9] Wen-Li Wei, Chung-Hsien Wu, Chung-Hsien Wu, Jen-Chun Lin and Han Li: "Exploiting Psychological Factors for Interaction Style Recognition in Spoken Conversation", IEEE/ACM Transaction on Audio, Speech, and Language Processing, 2014,Vol.22No.3, pp.659-670.

[10] Ashish Tawari and Mohan Manubhai Trivedi:"Speech Emotion Analysis: Exploring the Role of Context", IEEE Transaction on Multimedia, 2010, Vol.12No.6, pp.502-509.

[11] Martin Wöllmer, Björn Schuller, Florian Eyben, and Gerhard Rigoll, "Combining Long Short-Term Memory and Dynamic Bayesian Networks for Incremental Emotion-Sensitive Artificial Listening", IEEE Journal of Selected Topics in Signal Processing, 2010, Vol.4No.5, pp.867-879.

[12] Swati Johar, "Paralinguistic profiling using speech recognition", Int J Speech Technol, 2013, DOI 10.1007/s10772-013-9222-4.

[13] Matthis Drolet, Ricarda I. Schubotz,Julia Fischer:"Explicit authenticity and stimulus features interact to modulate BOLD response induced by emotional speech", Cogn Affect Behav Neurosci 13,2013, pp.318–329.

[14] Laura Caponetti, Cosimo Alessandro Buscicchio and Giovanna Castellano: "Biologically inspired emotion recognition from speech",

EURASIP Journal on Advances in Signal Processing, 2011, available: http://asp.eurasipjournals.com/content/2011/1/24.

[15] Changqin QUAN, Dongyu WAN,Bin ZHANG, Fuji REN, "Reduce the Dimensions of Emotional Features by Principal Component Analysis for Speech Emotion Recognition", Presented at the sixth symposium on system integration proceedings,2013, pp. 222 - 226.

[16] B. Yang, M. Lugger, "Psychological motivated multi-stage emotion classification exploiting voice quality feature." *F. Mihelic, J. Zibert, Speech Recognition, In-Tech*, 2008, chapter 22.

[17] Hinton, G. E. and Salakhutdinov, R.R, "Reducing the dimensionality of data with neural networks", Science, 2006, Vol. 313. no. 5786, pp. 504 - 507.

[18] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. "Gradient-based learning applied to document recognition." Presented at Proceedings of the IEEE, 1998.

[19] Theano 0.7 documentation [online], website: http://www.deeplearning.net/tutorial/lenet.html#lenet, 2015.

[20] USPS Dataset, available: http://www-i6.informatik.rwth-aachen.de/~keysers/usps.html, 2015.10.28.

[21] Hastie, Tibshirani and Friedman, book: "The elements of statistical learning", Springer, 2001.