Person Identification from Lip Texture Analysis

Zhihe Lu, Xiang Wu, Ran He
National Laboratory of Pattern Recognition, CASIA
Center for Research on Intelligent Perception and Computing, CASIA
Center for Excellence in Brain Science and Intelligence Technology, CAS
University of Chinese Academy of Sciences, Beijing, 100049, China
zhitongdao_he@sina.com, alfredxiangwu@gmail.com, rhe@nlpr.ia.ac.cn

***Abstract*—The interactive liveness detection for face recognition often requires users to read some digits from 0 to 9. The movement and variation of lip texture during reading potentially provide discriminative information for human identification. This paper firstly addressed the issue of whether the lip texture during reading can serve as a soft-biometric for person identification. Different from the traditional lip recognition methods that are based on color statistics and lip shapes, we develop a deep architecture that incorporates both CNN and LSTM to jointly model the appearance and the spatial-temporal information of lip texture. We also build a new lip recognition database that contains 11,123 videos for the number 0∼9 in Chinese from 57 people. Experimental results show that the proposed method can achieve 96.01% on close-set protocols, suggesting the usage of lip texture as soft-biometrics for facilitating face recognition.**

***Index Terms*—lip movement recognition; liveness detection; recurrent convolutional networks;**

I. INTRODUCTION

Video-based biometric recognition [1], [2], [3] is a primary challenge of computer vision. Compared with image-based recognition, more information of subjects can be used from videos. For example, when speaking, the lip movement has particular structures and characteristics such as the way of speaking, the size of upper and lower lips, furrows, grooves and the distance between lines and edges. The visible characteristics are unique to each subject. Compared with other biometric features, for example face [22],[23],[25] and fingerprint recognition, lip recognition is more secure due to the inimitability. Moreover, the lip recognition can combined with voice recognition to obtain the better performance and safety.

Traditional biometric recognition methods are usually based on hand-craft features for images rather than videos. These can lead to lose some information among frames. The key point for video-based recognition is how to build an appropriate model to obtain the visual representation which can integrate the information across different frames together. Graphical model [4] is a good way to model the relations accross frames but the computation is huge. Recently, with the developments of convolution neural network (CNN) [5] and recurrent neural network (RNN) [6], numerous biometric tasks [7], [8], [9] have benefited from the robust representation and the performance have made great progress. While CNN has been successful to obtain local spatial information, RNN has become an efficient method for modeling sequential data. In this paper, we introduce long-term recurrent convolutional networks (LRCN) [10] to video-based lip recognition.

Firstly, we employ CNN to extract high-level spatial features from each frame. Secondly, the long-short term memory (LSTM) [11] networks are used to build the relations among the CNN features of different frames. The framework is shown in Fig.1. The contributions are summarized as follows:

1. We collect a new database which contains 11,123 videos from 57 people, which each person speaks number from 0 to 9 in Chinese. In the first experiment, the training set has 10,011 videos and testing set has 1,112 videos, which both contain all the digits. Then we do other tests with each number to find which digit has better performance on distinguishing different individuals. The videos of training set and testing set can be seen in the section 4. We evaluate the effectiveness of proposed method in the case of database above.
2. We introduce LRCN to video-based lip recognition. CNN is a better feature extractor than other hand-craft features such as SIFT [12], HOG [13] and LBP [14]. And then the deep LSTM networks are able to capture complex temporal state dependencies. As is shown in Fig. 1, the LRCN [10] is an end-to-end framework from the trainable video to the confidence of each subject.
3. The experimental results are comparable. We obtain 96.01% $\pm$ 2.73% on subject-dependent(SD) testing set and 86.61% $\pm$ 2.62% on the subject-independent(SI) testing set.

The paper is organized as follows. In section 2, we briefly review the CNN and LSTM. Section 3 describes the proposed LRCN framework for video-based lip movement recognition. The experimental results are given in section 4 and the conclusion is presented in section 5.

II. PRELIMINARY

A. Convolution Neural Networks(CNN)

Convolutional neural network [24] is widely used deep learning structures for computer vision tasks. It is inspired by biological processes [5] and are variations of multilayer perceptrons [15], which are designed to use minimal amounts of preprocessing. The convolutional networks contains convolutional layer, pooling layer, ReLU layer, and fully connected layer.

Convolutional layer is the core part of a CNN and it has

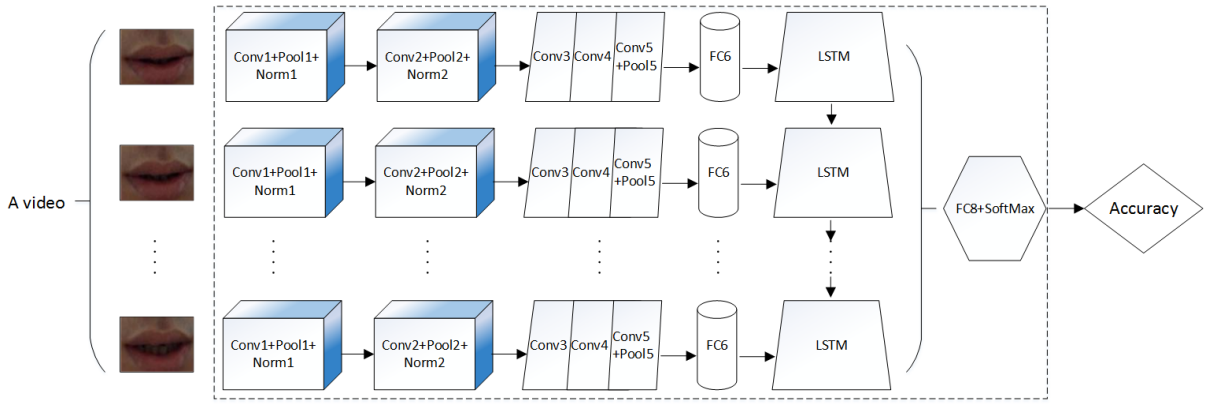

Fig. 1. The framework proposed consists of convolutional neural networks (CNN) and recurrent neural networks(RNN). The inputs of the architecture are videos. Then the videos have to go through five convolutional layers, one recurrent layer, two fully-connected layers, and one softmax classifier. Finally, we can obtain the correct recognition rates of each number in all the digits. The detailed results are in Section 4.

a set of learnable kernels. When inputs pass the kernels, each kernel is convolved across the width and height of the input volume, computing the dot product between the entries of the kernel and the input and producing a 2-dimensional activation map of that kernel. Hence, some specific type of feature in the input can be learned. Then Rectified Linear Units (ReLU) [16] layer applies the non-saturating activation function $f(x) = max(0, x)$, which increases the nonlinear properties of the receptive fields of the convolutional layer. The function of the pooling layer is to progressively reduce the spatial size of the representation to reduce the amount of parameters and computation in the network. After several convolutional and pooling layers, the fully connected layer can do complex reasoning tasks. The last fully connected layer holds the output, such as the class scores. The loss layer specifies how the network training penalizes the deviation between the predicted and true labels and is normally the last layer in the network.

B. Long short-term memory(LSTM)

Long short-term memory (LSTM) [11] is a recurrent neural network (RNN) architecture proposed by Sepp Hochreiter and Jrgen Schmidhuber. LSTM layer is capable of learning long-term dependencies from sequential data. It can overcome the gradient vanishing for traditional RNN layer, which leads to that LSTM outperforms alternative RNNs and Hidden Markov Models [17] and other sequence learning methods in numerous applications. For example, LSTM achieved the best results in unsegmented connected handwriting recognition[6], and won the ICDAR handwriting competition in 2009. LSTM networks have also been used for automatic speech recognition, and were a major component of a network that achieved a record 17.7% phoneme error rate on the classic TIMIT natural speech dataset [18]. Fig.2 has shown a simple LSTM architecture. A Cell consists of four parts such as input gate, out gate, forget gate, and cell unit. Some formulas related are as follows.

- Gates:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \quad (1)$$
$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \quad (2)$$
$$o_t = \sigma(W_{xo}x_t + W_{ho}ht - 1 + b_o) \quad (3)$$

- Input transformation:

$$c\_in_t = tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_{c\_in_t}) \quad (4)$$

- Status update:

$$c_t = f_t \cdot c_{t-1} + i_t \cdot c\_in_t \quad (5)$$
$$h_t = o_t \cdot tanh(c_t) \quad (6)$$

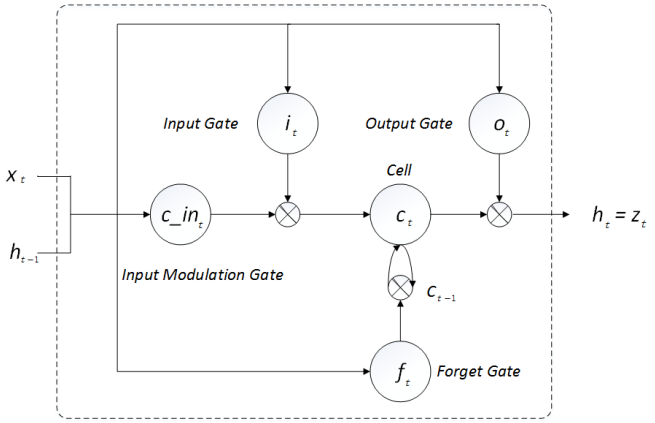

Fig. 2. A LSTM memory cell used in this paper (from [19], a slight simplification of the architecture described in [20], which was derived from the LSTM initially proposed in [11]).

III. OUR METHOD

A. Framework

The proposed architecture is explicitly shown in Fig.1. The critical ideas are listed as follows.

- Our deep recurrent convolutional networks are composed of five convolutional layers, three pooling layers, two

fully connected layers, one LSTM layer, and one softmax classifer.

- A video including several 227×227 images with RGB representations first passes convolutional layers, which can extract high-level features of the inputs. Then the pooling layers can reduce the dimension of the signals. All the signals are feeded into fully connected layers that have strong reasoning ability.
- The LSTM layer, which is the significant unit of this structure, obtain the inputs from the fully connected layer. With its intelligent ability, it determines when to remember the value or forget it and when to output the value.
- Then the last fully connected layer holds the output of LSTM. We can gain the final class scores through softmax function. The detailed data of inputs and outputs of five convolutional layers are shown in Table 1. The sixth fully-connected layer has 4096 neurons. The LSTM layer has 256 outputs. And the last fully-connected layer has 57 outputs.

TABLE I
INPUTS AND OUTPUTS OF CONVOLUTIONAL LAYERS

Layer	Input Size	Output Size
Conv1	227×227×3	111×111×96
Conv2	55×55×96	26×26×384
Conv3	13×13×384	13×13×512
Conv4	13×13×512	13×13×512
Conv5	13×13×512	13×13×384

B. Lip Movement Recognition

In our research, a lip movement recognition algorithm includes four basic modules: image quality assessment and selection, preprocessing, feature extraction, and matching. Fig.3 shows how it works.

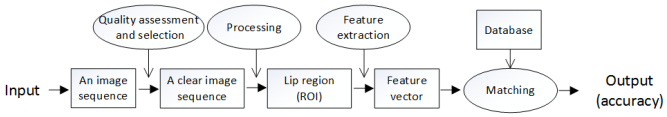

Fig. 3. The flowchart of our method.

1) Image Quality Assessment and Selection: When we collect image data, one subject can obtain many videos including 40 frame images of only one number, where includes ten numbers. However, a majority of these images are just useless for our purpose so that we should remove these bad images. As is known to all the researchers, good samples have an unexpected effect on the result. So we really need clear and sharp images for this lip movement recognition.The differences between good lip images and bad lip images that we do not need are shown in Fig.4. By comparison, you should already know why we must conduct image quality assessment and selection.

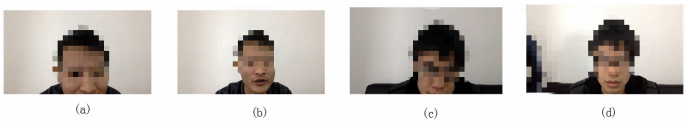

Fig. 4. Differences between good samples and bad samples. (a)Bad image. (b)Good image. (c)Bad image. (d)Good image. You can select the bad ones immediately.

2) Image Sequence processing: The images that we obtain contain the whole face, but what we really need is the lip, which is usually called region of interest (ROI) in an experiment. First, removing bad images is our beginning of the task. Second, we should intercept the lip from the images, which sounds difficult. Fortunately, [21] has made much efforts for this job. Then we gain the pictures we need with the method [21] proposed after our own modification. However, the rest of the images by our seriously selecting are not so many as before and the number of the images of one number of each subject is different from each other. This becomes a realistic question that needs our solution because of the need of our architecture. In order to give a good answer to it, we write a program with Matlab, which can make every video representing one number have the same number of images. Fig.5 has given some samples.

Fig. 5. After the steps above, we obtain these images.

3) Feature extraction: In contrast to video recognition, the static image description task only requires a single convolutional network since the input consist of only the single images. As we all know, videos always have much more features than only images. But there are also more difficulties for video recognition. With the development of deep learning, [10] has come up with a simple but practical method to video recognition. Then we set up a novel architecture with our own design, which is inspired by the approach they proposed. By this method, we can contact videos effectively and utilize much more feature information to recognize the identity of the person.

4) Lip Matching: Once we train a lip model, we find the lip in a given video using template matching. After several convolutional layers, fully-connected layers, a recurrent layer and a softmax classifier, we can predict the labels of all pixels in a video, which is the key to match videos respectively from training set and testing set. Then we make use of predictive class scores of different categories to obtain the final accuracy.

IV. EXPERIMENT

This paper presents a new method based on deep learning for identifying individuals from lip videos. We thus set up a

new database to test the performance of our approach.

A. New Database

We first select some people as our subjects in Shenzhen, China. Then we arrange our own men, who know what information we need clearly, to teach them how to use the advanced system, which is made use of to obtain the data we wish to have. Although the individuals in this research perform very well, there are still many bad images, which are useless to us. So we remove the bad images with high concentration of energy. By now, we have acquired initial image data. Because what we care about mostly is the lip, which is usually called region of interest, finding a good method to intercept lips from the images that contain the whole faces is our next task. As described in Section 3, we carry out next steps smoothly. Finally, we have set up our own database.

The database we established includes 11,123 videos from 57 subjects who speak numbers from 0 to 9 in Chinese. In order to produce a more comprehensive performance, we divide database into ten subsets, which each one contains the information of only one number. The profile of the database is shown in Table 2.

TABLE II
DATABASE

Number	Subjects	Video (frames)	Training set (videos)	Testing set (videos)
0	57	20	936	116
1			966	110
2			998	108
3			1031	112
4			977	98
5			1019	116
6			1017	105
7			997	128
8			1038	110
9			1032	109
0∼9			10011	1112

B. Experimental results

In order to achieve our purpose, we divide the database into several subsets and do both subject-dependent(SD) experiment and subject-independent(SI) experiment.In the SD experiments, the training and the testing data are from the same set of subjects and in the SI experiments, the training and the testing data are from different subjects.

1) Subject-dependent experiment: At first, we divide database into two parts, which are respectively used as training set and testing set. The training set has 10,011 videos and testing set has 1,112 videos (here people speak numbers from 0 to 9), which the latter is approximately accounting for 0.1 of the former. Then we conduct the first experiment. However, only this experiment can not explain which number is more distinguished to identify from one another. So we do other experiments as follows.

We utilize the classification of different digits in Table 2 and do experiments with these subsets respectively. All the SD experiment results are shown in Fig.6.

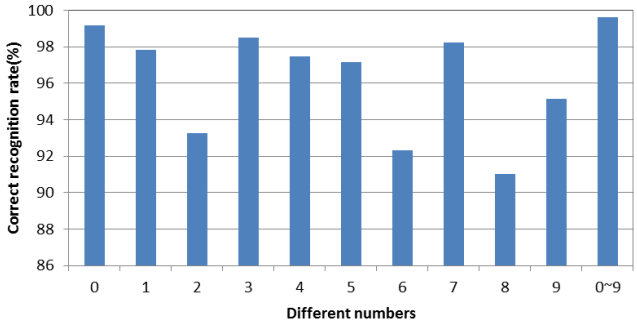

Fig. 6. In this figure,you can see the number '0' has the most distinguished ability to recognize different people. The number '8' has the worst performance in these experiments although its correct recognition rate has achieved ninety-one percent, which is usually regarded as a high rate in other experiments.

2) Subject-independent experiment: In order to verify the robustness of our method, we conduct more experiments to observe the performance of it. Therefore, we add a new class ,which has 20 videos, in the testing set of the experiments aboved and do not change anything of training set. After several tests, the results have shown that this method still need to be improved to enhance robustness. This is also the work we have to do next stage. All the detailed information is listed in Table 3.

TABLE III
RESULTS COMPARISON

Number	0	1	2	3	4
Accuracy(%) (SD)	99.19	97.83	93.25	98.50	97.46
Accuracy(%) (SI)	88.71	88.38	89.17	87.96	84.88
Number	5	6	7	8	9
Accuracy(%) (SD)	97.17	92.33	98.23	91.01	95.17
Accuracy(%) (SI)	88.42	82.71	88.56	81.42	85.92

V. CONCLUSION

In this paper, we have described an efficient method based on deep learning for personal identification from lip movement videos in interactive liveness detection. In order to verify our idea, we set up a bran-new database, which includes a large amount of information. This database can benefit researchers who want to study correlative topics. As you can see, this new soft biometrics recognition approach has better performance on person identification based on lip movement. We have leveraged the recurrent convolutional neural networks, a class of models that has the flexibility to be applied to a variety of vision tasks involving sequential inputs and outputs. Moreover, our results consistently demonstrate that by learning sequential dynamics with a deep sequence model, we can improve the correct recognition rate on identification task.

ACKNOWLEDGMENT

This work is supported by National Natural Science Foundation of China (Grant No. 61473289), Beijing Municipal Science and Technology Project (Grant No. Z141100003714131, Z161100000216144), the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDB02070000). We thank the anonymous reviewers for their helpful comments.

REFERENCES

[1] Kuang-Chih Lee, Jeffrey Ho, Ming-Hsuan Yang, and David Kriegman, *Video-Based Face Recognition Using Probabilistic Appearance Manifolds*. CVPR,2003.

[2] Dmitry O. Gorodnichy, *Video-based framework for face recognition in video*. CRV,2005.

[3] Guang-yu Xu and Rong-yi Cui, *Video-Based Recognition of Walking States*. ICAIC,2011.

[4] Jordan and M. I. , *Graphical Models*. Statistical Science,2004.

[5] Matusugu, Masakazu, and Katsuhiko Mori, *Subject independent facial expression recognition with robust face detection using a convolutional neural network*. Neural Networks,2003.

[6] A. Graves, M. Liwicki, S. Fernandez, R. Bertolami, H. Bunke, and J. Schmidhuber, *A Novel Connectionist System for Improved Unconstrained Handwriting Recognition*. IEEE Transactions on Pattern Analysis and Machine Intelligence,2009..

[7] Li Ma,Tieniu tan,Yunhong Wang,and Dexin Zhang, *Personal Identification Based on Iris Texture Analysis*. IEEE Computer Society,2003.

[8] A. Jain, R. Bolle and S. Pankanti, *Biometrics: Personal Identification in a Networked Society*. Kluwer,1999.

[9] D. Zhang, *Automated Biometrics: Technologies and Systems*. Kluwer,2000.

[10] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama,and Marcus Rohrbach, *Long-term Recurrent Convolutional Networks for Visual Recognition and Description*. Technical Report No. UCB/EECS-2014-180,2014.

[11] S. Hochreiter and J. Schmidhuber, *Long short-term memory*. Neural Computation,1997.

[12] Lowe, and David G, *Object recognition from local scale-invariant features*. ICCV,1999.

[13] Navneet Dalal and Bill Triggs, *Histograms of Oriented Gradients for Human Detection*. CVPR,2005.

[14] T. Ojala, *Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns*. TPAMI,2002.

[15] LeCun, and Yann, *LeNet-5, convolutional neural networks*. 2003.

[16] Xavier Glorot, Antoine Bordes and Yoshua Bengio, *Deep sparse rectifier neural networks*. AISTATS,2011.

[17] Baum, L. E. and Petrie, *Statistical Inference for Probabilistic Functions of Finite State Markov Chains*. The Annals of Mathematical Statistics,2011.

[18] Graves, Alex; Mohamed, Abdel-rahman; Hinton,and Geoffrey, *Speech Recognition with Deep Recurrent Neural Networks*. ICASSP,2013.

[19] W. Zaremba and I. Sutskever, *Learning to execute*. arXiv preprint arXiv:1410.4615,2014.

[20] A. Graves, *Generating sequences with recurrent neural networks*. arXiv preprint arXiv:1308.0850, 2013.

[21] Xuehan Xiong and Fernando de la Torre, *Supervised Descent Method and Its Application to Face Alignment*. CVPR, 2013.

[22] Ran He, Yinghao Cai, Tieniu Tan and Larry Davis, *Learning Predictable Binary Codes for Face Indexing. Elsevier Pattern Recognition*. 48(10): 3160-3168, 2015.

[23] Shu Zhang, Ran He, Zhenan Sun and Tieniu Tan, *Multi-task ConvNet for Blind Face Inpainting with Application to Face Verification*. ICB, 2016.

[24] Linlin Cao, Ran He and Baogang Hu, *Locally Imposing Function for Generalized Constraint Neural Networks - A Study on Equality Constraints*. International Joint Conference on Neural Networks, 2016.

[25] Xiang Wu, Ran He and Zhenan Sun, *A Lightened CNN for Deep Face Representation*. CoRR abs/1511.02683, 2015.