Digital recognition from lip texture analysis

Wenkai Dong, Ran He, Shu Zhang
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
dongwenkai2016@ia.ac.cn, {rhe, shu.zhang}@nlpr.ia.ac.cn

***Abstract*—Digital recognition with lip images has become a key step of the interactive liveness detection for Chinese banking systems. However, the problem of the digital recognition is very challenging due to intra class variation of lip images, head pose variations, and uncontrolled illumination. This paper studies a deep learning architecture to model the appearance and the spatial-temporal information of lip texture. The lip texture in still image frames and the spatial-temporal relationship between these frames are jointly modeled by convolutional neural networks and long short-term memory. Two strategies are further exploited to find effective groups of ten digitals for training the deep models. As a result, more information can be utilized for accurate recognition based on lip texture analysis. Besides, two datasets of isolated digits in Chinese are established to simulate real-world liveness detection environments together with various attacks. Extensive experiments have been done to analyze the recognition accuracy of each digit and to provide some clues for determining appropriate digits for interactive liveness detection.**

***Index Terms*—digital recognition, liveness detection, lipreading, deep learning**

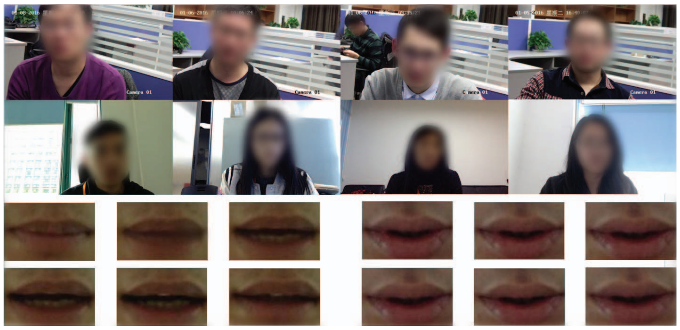

Fig. 1. Subjects in the first and second row are from the first and second dataset respectively. We can see that the environments in the images of the first row are the same while those of the second row are different, which has a significant effect on recognition. And the images below are two sequences of mouth regions when uttering digits. The motions of the lips uttering 8 in Chinese is very clear (below and left) while the montions of the lips uttering 4 in Chinese are tiny (below and right).

I. Introduction

Digital recognition from lip texture analysis is similar to lipreading which has been studied in computer vision for several years. However, it remains as a difficult problem in interactive liveness detection due to the following reasons. First, the intra class variation is rather large due to the appearance discrepancy among different identities and the difference in the range of motion when uttering the same digits (Figure 1 below). Second, the recognition performance is easily affected by the capture environment and the illumination condition in particular. Third, some digits are easily attacked. Namely, some digits are easily confused by other digits, which makes them unsuitable for livenenss detection. Visual representations from video data are crucial for dealing with these issues and designing effective recognition systems. Saenko et al. [7] make use of the HMM to extract dynamic features of visual signals. Zhao et al. [8] propose a spatiotemporal version of LBP features for lipreading. Aharon and Kimmel [9] employ nonlinear dimension reduction methods to analyze visual lip images. Zhou et al. [10] combines LBP-like features and embedding methods for lipreading. Pei et al. [11] integrate multimodal data and fusions of feature channels. However, one limitation of these hand-crafted features is that they lack semantics and discriminative capacity. In contrast, deep-learned features can learn the semantic representation from raw video via deep convolutional networks.

CNN models have proven highly successful in not only static image recognition tasks such as the MNIST, CIFAR, and ImageNet [3, 4, 5, 18, 19, 20, 21], but action recognition in video data, the most successful one of which is probably the Two-Stream ConvNets [6]. It matches the state-of-the-art performance of improved trajectories [12] on UCF101 and HMDB51. However, most of current deep learning based action recognition methods largely ignore the intrinsic difference between temporal domain and spatial domain, and just treat temporal domain dimension as feature channels when adapting the architectures of ConvNets to model videos [13]. To deal with this issue, we apply recurrent neural networks to our framework. LSTMs are widely used in handwriting recognition, speech recognition, emotion detection and evaluating programs. In [14, 15], the authors use CNN to extract features from the video data and feed them into LSTM networks, which has performed well on UCF-101.

Motivated by the analysis above, we propose a deep architecture composed of CNN and LSTMs for digital recognition task, which integrates the advantages of CNN and LSTMs. We utilize CNN to learn powerful image features for the lip appearance and LSTMs to learn the spatial-temporal variation among frames. Lacking of datasets of isolated digits in Chinese, we collect two datasets for our research. All the video data of the first one is collected in the same environment

978-1-5090-4165-7/16/$31.00 ©2016 IEEE

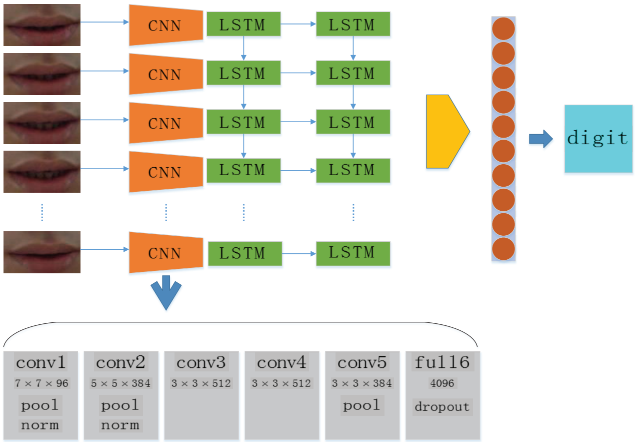

Fig. 2. The architecture of the proposed method for digital recognition. The model combines a deep hierarchical visual feature extractor (CNN) with a model that can learn to recognize and synthesize temporal dynamics for tasks which take sequetial data as inputs.

while the data of the second one is collected in different environments with different illumination conditions (Figure 1 above). We train and evaluate our models on both datasets to analyze the effect of illumination conditions on each digits. Generally, a lipreading system can be considered as subject-dependent (SD) or subject-independent (SI). The former is often used in a private environment and the latter is designed to serve users in a public environment. 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. So we have conducted many SD and SI experiments to show the effect of various speaking characteristics on recognition performance. Moreover, we analyze the accuracy of each digit in all experiments and conclude that 5 and 8 are the most suitable digits for digital recognition task in the interactive liveness detection , because they are robust to attack.

II. Our Work

There are three main difficulties in our task. First, the intra class variation is rather large due to the appearance discrepancy among different identities and the difference in the range of motion when uttering the same digits. Second, the recognition performance is easily affected by the capture environment and the illumination condition in particular. Third, some digits are easily attacked. To deal with the problems, we propose a framework composed of CNN and LSTM neworks and explain how it works in this section. Besides, we introduce the datasets we collect for our task and methods of preprocessing in detail.

A. Overview of our work

- In [15], a Long-term Recurrent Convolutional Network (LRCN) model is proposed to deal with tasks involving sequential data and has achieved good performance. So we apply a similiar method to our digital recognition task.

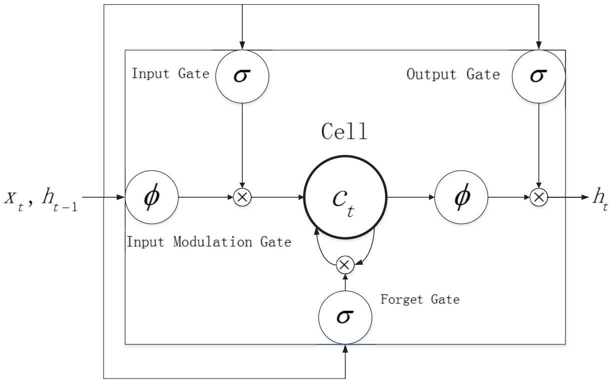

Fig. 3. A diagram of the LSTM memory cell used in this paper

- Our deep architecture is composed of five convolution layers and LSTMs with 256 hidden units, whose outputs are 10 digits.
- As the Figure 2 shows, our model works by passing each visual input (a 227 × 227 × 3 frame) via a feature transformation to produce a 4096-dimention vector representation. Having computed the feature-space representation of the visual input sequence, it is fed into the sequence model, which finally produces a prediction. By both specifically modeling the lip texture in still images and considering the spatial-temporal relationship between frames, our deep model can exploit more information for correct recognition based on lip texture analysis.
- Lacking of datasets of uttering digits in Chinese, we establish two datasets in a controlled or uncontrolled environment for our research containing 960 and 14367 sequences respectively to simulate real-world liveness detection environments together with various attacks, which can be helpful for furthur research.

B. LSTM unit

As we know, though traditional RNNs have proven successful on tasks such as speech recognition and text generation, they have trouble learning over long sequences due to the problem of vanishing and exploding gradients [17]. LSTMs incorporate memory cells, which allows the network to learn when to forget previous hidden states and when to update hidden states given new information. We use the LSTM unit as described in [16] (Figure 3), which is a slight simplification of the one described in [2], to learn the spatial-temporal variation among frames. Letting $\sigma(x) = (1 + e^{-x})^{-1}$ be the sigmoid nonlinearity and letting $\phi(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} = 2\sigma(2x) - 1$ be the hyperbolic tangent nonlinearity, the LSTM updates for timestep t given inputs x_t, h_{t-1} and c_{t-1} are:

$$\begin{aligned}
i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \\
f_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \\
o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \\
g_t &= \phi(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \\
c_t &= f_t \odot c_{t-1} + i_t \odot g_t \\
h_t &= o_t \odot \phi(c_t)
\end{aligned}$$

where h_t, i_t, f_t, o_t, g_t and c_t are the hidden unit, input gate, forget gate, output gate, input modulation gate and memory cell respectively.

C. Details of our framework

We utilize CNN to learn powerful image features for the lip appearance. The architecture of CNN contains five convolutional layers, followed by one fully-connected layer. As shown in Figure 2, the input of our structure is a sequence of $227 \times 227 \times 3$ images. Following the input, the first convolutional layer filters the input images via 96 kernels of size $7 \times 7 \times 3$ with the stride of 2. The second convolutional layer filters the input of the previous layer with 384 kernels of size $5 \times 5 \times 96$ with the stride of 2. The third convolutional layer contains 512 kernels of size $3 \times 3 \times 384$. The fourth convolutional layer contains 512 kernels of size $3 \times 3 \times 512$. The fifth convolutional layer contains 384 kernels of size $3 \times 3 \times 512$. The sixth fully-connected layer have 4096 neurons. Through a deep-learned features extractor, each $227 \times 227 \times 3$ image is turned into a 4096-dimention vector, which is fed into a single-layer LSTM with 256 hidden units later. Finally, a prediction distribution is produced by taking a softmax over the outputs of the sequential model.

D. Datasets collection

Considering that our model is designed for users in China and there are no datasets of uttering isolated digits in Chinese, two datasets are collected for our experiments, the details of which are as follows.

One consists 29 subjects uttering 10 digits six times at a resolution of 960×540. All the data of this dataset is collected in a controlled environment with the same illumination condition.

The other consists 77 subjects uttering 10 digits 15 or 25 times at a resolution of 640×360. The videos of this dataset is collected in an uncontrolled environment with different illumination conditions.

After data collection, the raw video data needs to be annotated due to the following reasons. First, there are many extra motions of lips before or after uttering, which has a significant effect on recognition. Second, some videos fail to capture complete motions of uttering. Therefore, we annotate the raw data manually by deleting extra frames and some videos. After data annotation, controlled and uncontrolled datasets have 960 and 14367 sequences respectively.

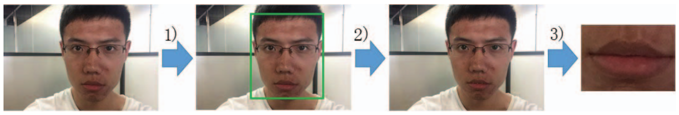

Fig. 4. Three steps of preprocessing: 1) face detection, 2) mouth locating, 3) cropping.

E. Preprocessing

In general, a speech video is preprocessed through the following three steps (Figure 4): 1) detecting faces in its frames, 2) facial feature points location, and 3) cropping off the mouth region. We use the method in [1] to detect faces and locate facial feature points in each frame. We then crop off the mouth region and resize all the images to make the widths of mouths constant. A 100×70 mouth region is cropped off each of video frames. Finally, we resize all the images to 227×227, because the result of feeding the 100×70 images into our model directly is not good.

III. Experiments

In this section, we first train and evaluate our architecture on both datasets to demonstrate that different environments and illuminations can significantly affect the digital recognition performance. Then we conduct SD and SI experiments with or without attack. Finally, we analyze the accuracy of each digit in all experiments and conclude that 5 and 8 are the most suitable digits for digital recognition task in interactive liveness detection, because they are robust to attack.

A. Implementation details

Learning rate. When training a model from scratch, the rate is set to 10^{-2} initially and is changed to 10^{-3} after $10K$ iterations, then to 10^{-4} after $20K$ iterations, and training is stopped after $30K$ iterations. In the fine-tuning scenario, the rate is set to 10^{-3} initially and is changed to 10^{-4} after $10K$ iterations, then to 10^{-5} after $20K$ iterations, and training is stopped after $30K$ iterations.

B. Experiments

Models trained using all digits. First, we train two models with all digits. The first one is trained from scratch on the controlled dataset using 29 subjects and the testing data is from the same subjects (CSD). And for subject dependent and subject independent experiments, which have been mentioned in Section I, the uncontrolled dataset is separated into two parts containing 57 and 20 subjects respectively. The second model is pre-trained on the controlled dataset and fine-tuned on the uncontrolled one using the data from 57 subjects of the first part. Then we evaluate the model using the data from the first part (UCSD) and the second part (UCSI). From the results, presented in Table 1, SI experiment is a more challenging task than SD experiment due to large variations within lip textures, varying speaking speeds and different accents. Generally, it is beneficial to increase the amount of training data in deep learning. However, the accuracies decrease apparently in the UCSD experiment though the second dataset is furthur larger

TABLE I
ACCURACIES OF MODELS TRAINED USING ALL DIGITS IN THREE EXPERIMENTS

Test	Subject-dependent	Subject-independent
Controlled	80.1%	-
Uncontrolled	61.0%	43.1%

TABLE II
ACCURACIES OF MODELS TRAINED USING PARTIAL DIGITS IN SIX EXPERIMENTS

Test	Without attack	With attack
CSD	91.2%	83.7%
UCSD	74.9%	71.2%
UCSI	58.1%	54.2%

TABLE III
MEAN ACCURACY OF 0, 2, 3, 5, 8 AND 9

Test	Using all digits	Using partial digits
CSD	80.1%	83.1%
UCSD	73.8%	74.9%
UCSI	56.4%	58.1%

than the first one. The reason may be that our architecture is not deep enough for the complex environments and illumination conditions.

Models trained using partial digits. Having evaluated models trained using all digits, in order to show the effect of attack, we divide all digits into two parts and take digits 1, 4, 6, 7 as attack. First, we train two models without attack, which means the outputs of models are 0, 2, 3, 5, 8 and 9. Then we evaluate the model in CSDNA, UCSDNA and UCSINA (NA denotes without attack) three experiments, which are similar to CSD, UCSD and UCSI experiments above. Next, two models are trained with attack, which means the outputs of models are 0, 2, 3, 5, 8, 9 and others. The evaluation is also performed in CSDA, UCSDA and UCSIA (A denotes with attack) three experiments. The results are reported in Table 2. As expected, we can conclude that attack has an effect on digital recognition in liveness detection. Moreover, from the results in Table 3, it is beneficial to train models using partial digits.

C. Analysis

Here we analyze the accuracy of each digit in the experiments above to figure out which digits are most suitable for liveness detection.

First, Figure 5 shows the effect of illumination. From the results presented, we can conclude that different illumination conditions hardly affect the accuracies of 2 and 8 while 0, 1, 6, 7 and 9 are sensitive to different light conditions, because their accuracies decrease apparently when light conditions change.

Second, Figure 6 gives the recognition results in six experiments with or without attack. It is clear that 0, 3 and 9 are more sensitive to attack than 2, 5 and 8.

Finally, we compare the results in SD experiments and SI experiments, which are presented in Figure 7. As expected, accuracies of all digits decrease in the SI experiments due to large variations within lip textures, varying speaking speeds and different accents.

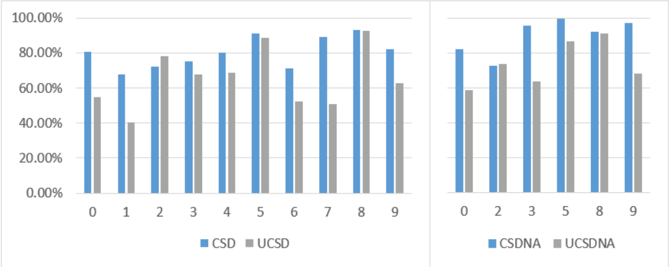

Fig. 5. Recognition rates of each digit in CSD, UCSD, CSDNA and UCSDNA experiments.0, 1, 6, 7 and 9 are more sensitive to different light conditions than 2 and 8.

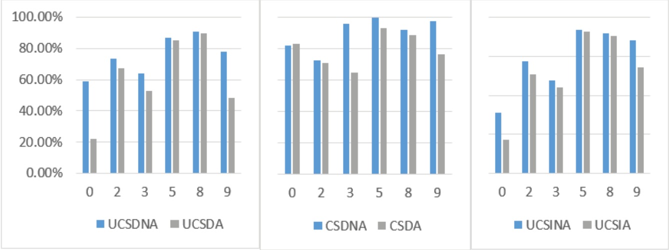

Fig. 6. Recognition rates of each digit in UCSDNA, UCSDA, CSDNA, CSNA, UCSINA, UCSIA experiments.It is clear that 0, 3 and 9 are more sensitive to attack than 2, 5 and 8.

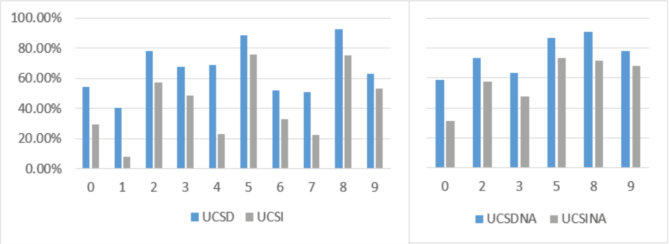

Fig. 7. Recognition rates of each digit in UCSD, UCSI, UCSDNA and UCSINA experiments.Accuracies of all digits decrease in the SI experiments due to the various speaking characteristics.

From the analysis above, we believe that 5 and 8 are the most appropriate digits for liveness detection, because the accuracies of them in all experiments are over 70%.

IV. CONCLUSION

This paper has proposed a deep architecture for digital recognition from lip texture analysis in interactive liveness detection. CNN and LSTMs have been employed to learn powerful image features for the lip appearance and the spatial-temporal variation among images respectively. We also collected two datasets of uttering isolated Chinese digits to simulate various liveness detection environment, which provides a good platform to evaluate different methods and can serve as benchmarks for future research. Moreover, we have conducted comprehensive experiments to evaluate our methods. The experimental results suggest that illumination variations have significant effects on recognition performance. Particularly, some of the 0-9 digits seem to be easily attacked on recognition which makes them inappropriate for liveness detection.

V. 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).

References

[1] X. Xiong, F. D. Torre. Supervised Descent Method and its Applications to Face Alignment. In *CVPR*, 2013.

[2] A. Graves and N. Jaitly. Towards end-to-end speech recognition with recurrent neural networks. In *ICML*, 2014.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *Proc. NIPS*, pages 1097-1105, Lake Tahoe, Nevada, USA, 2012.

[4] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014.

[5] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *Proc. ECCV*, pages 818-833, Zurich, Switzerland, 2014.

[6] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. *NIPS*, 2014.

[7] K. Saenko, K. Livescu, M. Siracusa, K.Wilson, J. Glass, and T. Darrell. Visual speech recognition with loosely synchronized nfeature streams. In *ICCV*, pages 1424-1431, 2005.

[8] G. Zhao, M. Barnard, and M. Pietikainen. Lipreading with local spatiotemporal descriptors. *IEEE Transactions on Multimedia*, 11(7):1254-1265, 2009.

[9] M. Aharon and R. Kimmel. Representation analysis and synthesis of lip images using dimensionality reduction. *IJCV*, 67(3):297-312, 2006.

[10] Z. Zhou, G. Zhao, and M. Pietikainen. Towards a practical lipreading system. In *IEEE Conf. on CVPR*, pages 137-144, 2011.

[11] Y. Pei, T. Kim and H. Zha. Unsupervised Random Forest Manifold Alignment for Lipreading. *ICCV*, 2013.

[12] H. Wang and C. Schmid. Action recognition with improved trajectories. In *ICCV*, 2013.

[13] L. Wang, Y. Qiao and X. Tang. Action Recognition with Trajectory-Pooled Deep-Convolutional Descriptors. In *CVPR*, 2015.

[14] J. Y. Ng, M.Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga and G. Toderici. Beyond Short Snippets: Deep Networks for Video Classification. In *CVPR*, 2015.

[15] J. Donahue, L. A. Hendricks, and S. Guadarrama. Long-term Recurrent Convolutional Networks for Visual Recognition and Description. In *CVPR*, 2015.

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

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

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

[19] Shu Zhang, Ran He, Zhenan Sun, Tieniu Tan. Multi-task ConvNet for Blind Face Inpainting with Application to Face Verification. International Conference on Biometrics (ICB), 2016.

[20] 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.

[21] Xiang Wu, Ran He, Zhenan Sun. A Lightened CNN for Deep Face Representation. CoRR abs/1511.02683 (2015).