Accurate Mouth State Estimation via Convolutional Neural Networks

Jie Cao, Haiqing Li, Zhenan Sun and 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
caojie2016@ia.ac.cn, {hqli, znsun, rhe}@nlpr.ia.ac.cn

***Abstract*—Human mouth is very flexible such that its status (closed or open) is often used as a judgment in the liveness detection of face recognition. However, due to large head pose and illumination variations, accurate mouth status estimation is still challenging in real-world scenarios. In this paper, we propose a deep convolutional neural networks (CNNs) method for mouth status estimation under unconstrained conditions and different types of attacks. Different from previous methods that extract hand-crafted features and then treat the estimation problem as a binary classification task, our method automatically extracts discriminative features via learned convolutional and the pooling layers. To demonstrate the effectiveness of our method and the challenge of mouth status estimation in real-world, we also propose a mouth status estimation dataset that contains 10,714 images in the wild. Experimental results with two types of liveness attacks show that our proposed method outperforms the other traditional methods, especially in the wild condition.**

***Index Terms*—mouth state estimation in the wild; convolutional neural networks; robust feature extracting;**

I. Introduction

Given an arbitrary human face image that contains mouth, to determine the state of the mouth, open or closed, belongs to the problem of mouth state estimation. Mouth state estimation is very beneficial for multiple applications, such as liveness detection, face verification and emotions recognition. For example, the most common detection method in an interactive liveness detection system is to ask the user to open his or her mouth. In practical application, high-accuracy mouth state estimation is a challenging problem due to the large variations in poses, expressions and illumination in the wild conditions, as well as the ambiguity of the definition of an open mouth. Some mouth samples shown in Figure 1 indicate that there is no apparent visible difference between the open and closed mouth sometimes if the pictures are captured in unconstrained conditions. Those factors mentioned above make the prediction very hard even for human being.

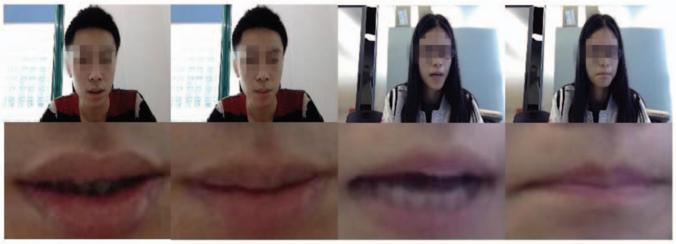

Fig. 1. First row: some samples from the dataset collected by ourselves. Second row: the patches cropped from the images in the first row. The mouthes in the first and the third column are manually labeled as open ones whereas the mouthes in the second and the fourth column are labeled as closed ones.

Mouth state estimation has attracted a great deal of research interest in recent years. In previous traditional methods, extracting hand-crafted features and then solving the problem as a binary classification task have become a standard step in mouth state estimation. Haar-like features and the cascade Adaboost classifier were firstly proposed by Viola and Jones [14] in the field of face detection and the method was soon borrowed to mouth state estimation. A large amount of studies on hand-crafted feature design have made great progress, e.g., Bouvier at al. [2] designed a retina filter to extract desirable features and classify them by a binary SVM classifier, Kumar at al. [10] extracted HOG-like features on various mouth regions to tackle attribute classification, Yuen at al. [18] predicted the mouth state by the shape parameter acquired by detecting the mouth boundary, Wang et al. [15] combined LBP and HOG features as the feature set to deal with partial occlusion situation, Bourdev et al. [1] built a three-level SVM system to extract higher-level information to improve the discriminativeness of hand-crafted features. Although much progress has been achieved in the past decades, the mouth state estimation in real world scene with large variations in poses, expressions and illumination is still not well solved. Nowadays, traditional approaches with hand-crafted features are limited in further improving the estimation accuracy and robustness. How to design robust features in coping with these variations becomes a headache problem. Fortunately, convolutional neural networks have shown its power in many challenging computer vision problems, such as face detection [11], face alignment [20], face recognition [5] and attribute estimation [7]. Due to the deep structure [19], CNNs can learn to extract high-level features which are invariant with poses, expressions, illumination [16].

Considering the advantages of CNNs [3], we employ them to address the mouth state estimation problem. We design the

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

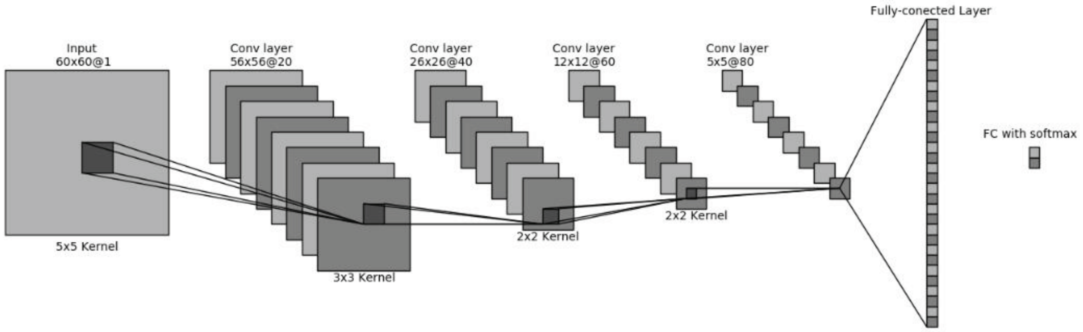

Fig. 2. The structure of our proposed deep convolutional networks. The input is the gray scale image and the outputs are the probabilities of being an open mouth and being a closed mouth. The third and the fourth convolutional layers share weights locally. 2×2 max-pooling layers between the convolutional layers are omitted for simplicity. The numbers above each cuboid, e.g., $(56 \times 56@20)$, denote the size of filter maps(56×56) and the number of filter maps (20), respectively. Local receptive field of neurons in different layers are illustrated by the numbers below each cuboid.

deep convolutional neural networks that are very effective for high-accuracy estimation. Our convolutional networks take the cropped mouth image as input without complex preprocessing to make the best use of texture. The learned convolutional layers extract robust features that are effective for mouth state estimation automatically, then the fully-connected layers generate feature representation for the softmax function to produce the possibility for the mouth to be open and closed. Unlike the cascaded regression approach in [13], we use a single convolutional net to predict the mouth state, and we do not take any pre-train strategies, which rely on a huge amount of data collected for other tasks. We only use the data collected by ourselves and train our model from scratch, so the data preparation for our method is much easier and the training process is much faster.

Sufficient data are an essential part of the model training. However, to the best of our knowledge, there is no publicly available dataset built specifically for mouth state estimation at present. To demonstrate the performance of our model and further research, we establish a dataset with 10,714 labeled frontal face images of 58 subjects. Those images are captured while people are reading the digits 0 to 9 in Chinese in front of the camera. Figure 1 shows some samples from our dataset. The mouth state estimation on our dataset is a challenging problem because the changes among those images are very subtle. We randomly split our dataset into training set and test set and compare our method with traditional methods which use manually designed features and SVM. Experimental results show that our proposed method outperforms the other methods and is more robust to illumination and irregular noise.

II. Our Method

In this section, we first introduce the data preprocessing, then give an overview of our proposed convolutional neural network and introduce the implementation details.

A. Data Preprocessing

We employ SDM method [17] for mouth landmark detection on the frontal face images, then the mouth patches are cropped by locating a bounding box that contains the mouth region of interest. The center of the bounding box of mouth is the center of the mouth, the width and height of the bounding box is 1.1 times and 0.7 times the width of the mouth, respectively. For the convenience of calculation, we resize all the patches to $60 \times 60 \times 1$, where 1 means using gray representation. To further augment the training set, we also randomly flip and rotate the patches.

B. Implementation Details of CNNs

To address the drawbacks of manually designed feature extractors [8], we adopt an deep learning approach. Discriminative features can be extracted by the high-level layers of our deep convolutional structures, which improves the accuracy and robustness of the mouth state estimation. The architecture of our proposed networks is summarized in Figure 2. Our CNNs in this paper contain four convolutional layers, three max-pooling layers, followed by two fully-connected layers. The input of our CNNs is a 60×60 image patches with gray representation as mentioned above. Following the input, the first convolutional layer filters the input patches via 20 kernels of size $5 \times 5 \times 1$ with the stride of 1. The second convolutional layer filters the input of the previous layer with 40 kernels of size $3 \times 3 \times 20$. The third convolutional layer contains 60 kernels of size $2 \times 2 \times 40$. The fourth convolutional layer contains 80 kernels of size $2 \times 2 \times 60$. The first fully-connected layer has 120 neurons, and the last fully-connected layer has 2 neurons. The outputs of the first three convolutional layers in the same kernel map are summarized by a 2×2 max-pooling layer before fed into the next layer. Finally, our CNNs output the possibility for the mouth to be open and closed.

C. Activation function

We adopt the hyperbolic tangent function rectified by absolute value as activation function, which effectively improves the performance of convolutional neural networks on Caltech-101. With the activation function, the convolution operation is formulated as

$$y^{j(r)} = |tanh(b^{j(r)} + \sum_{i=1} w^{ij(r)} \times x^{i(r)})|. \tag{1}$$

Where x^i and y^j are the i-th input map and the j-th output map, respectively. w^{ij} denotes the weight between the i-th input map and the j-th output map. b^j is the bias of the j-th output map, and × denotes the convolutional operation.

D. Locally sharing weights

In general, all the neurons of convolutional networks on the same map are globally shared based on the assumption that the same features may appear everywhere in an image. That means filters useful in one place should also be useful in others. However, for the mouth patches with fixed spatial layout in our experiment, locally sharing weights at high layers is a more beneficial approach for learning different high-level features. Each feature map in the third and the fourth convolutional layers of our model is equally divided into 2 by 2 regions and only the weights in the same region are shared. The convolution operation in the third and the fourth convolutional layers can be modified as:

$$y^{j(r,p,q)} = |tanh(b^{j(r,p,q)} + \sum_{i=1} w^{ij(r,p,q)} \times x^{i(r,p,q)})|. \tag{2}$$

Where $p \in \{0, 1\}$ and $q \in \{0, 1\}$.

E. Training Details

Our CNNs are trained from scratch by back propagation and SGD with cross-entropy loss function. We use a Gaussian distribution with zero mean and a standard deviation of 0.01 to initialize weights. The biases are initialized as 0. In each iteration, the weights are updated after learning the mini-batch with the size of 128. In all layers, we set the momentum as 0.9 and the weight decay as 0.005. This small amount of weight decay is important for the model to learn. The basic learning rate is set to 0.01 and updated as:

$$lr_n = lr_0 \cdot (1 + \gamma \times n)^{-p}. \tag{3}$$

Where lr_0 and lr_n are the basic learning rate and the learning rate at the n-th iteration, respectively. In this paper, we set $\gamma = 0.0001$ and $p = 0.75$.

III. Experiments

A. Dataset

To the best of our knowledge, there is no publicly available dataset built specifically for mouth state estimation at present. To demonstrate the performance of our model and further research, we construct a dataset that contains 10,714 frontal face images of 58 subjects captured from videos of uttering digits 0 to 9 in Chinese at a resolution of 640 × 360. The state of mouths in each image has been manually classified as open or closed. Figure 1 shows the samples from our dataset. The same as [12] , we perform the subject dependent (SD) experiments and the subject independent (SI) experiments to demonstrate the effectiveness of our approach. The training and test data of the SD experiments are from the same set of subjects whereas those data of the SI experiments are from different subjects. In our experiment, we randomly select 8,940 images consisted of 52 subjects as the training set and 916 images from the same subjects as the test set for SD experiments. We select the remaining 858 images of 6 subjects as the test set for SI experiments. The performance of mouth state estimation is measured with the average accuracy in following experiment. It indicates the performance of an algorithm. The average accuracy is measured as

$$accuracy = \frac{TP + TN}{N}. \tag{4}$$

Where TP and TN denote the numbers of correctly predicted images and correctly rejected images, respectively. N is the number of images in the test set.

B. Comparison with other methods

We compare our method with the traditional approaches extracting hand-crafted features and using Support Vector Machine (SVM) classifier to predict the label. Here, two types of feature extractors are compared with our method:

- HOG (Histograms of Oriented Gradients) feature extractor [6]: We use the single window approach and the detection window is divided into cells of size 8×8 pixels and each group of 2×2 cells is integrated into a block in a sliding fashion, each cell consists of a 9-bin Histogram of Oriented Gradients. The input images are the cropped mouth patches rescaled to 48×48 and a 900-dimensional HOG feature vector is extracted from each image.
- LBP (Local Binary Pattern) feature extractor: We equally divide the input image into 16 non-overlapping cells. We use the uniform local binary pattern [9] operator in the neighborhood of size 8×8 pixels. Then we compute and normalize the histogram of the frequency of each pattern occurring over the cell. We concatenate histograms of all cells into a cell-structured 928-dimensional LBP feature vector. The input images are the same as the input images prepared for HOG feature extractor.

TABLE I
Summary of the SD and SI test accuracy of different methods

Setting	HOG+SVM	LBP+SVM	Ours
SD test Accuracy	76.0%	74.1%	90.5%
SI test Accuracy	71.0%	64.5%	84.4%

We classify the extracted feature vectors by employing a linear SVM. In our implementation, we use the libsvm library [4] to train the SVM classifier. The SD and SI test set are divided into 10 subsets by different uttering digit in the comparison experiment. We evaluate the performance of different methods on each subset individually and report the

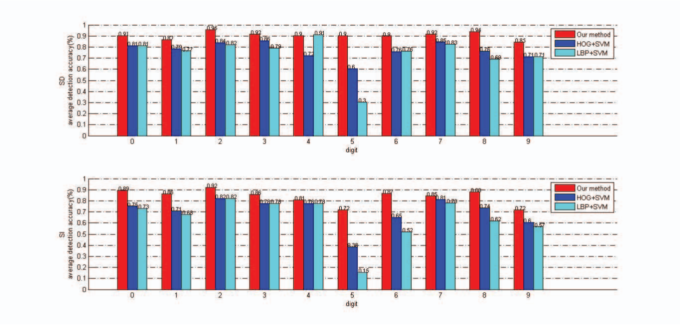

Fig. 3. The SD and SI test accuracy of the different methods on the 10 subsets. The result suggests that our method is apparently better at handling with difficult situations. For example, the mouthes that utters the digit 5 is the hardest to classify in our experiment, and the accuracy of our method is nearly twice the accuracy of HOG+SVM on the SI test set.

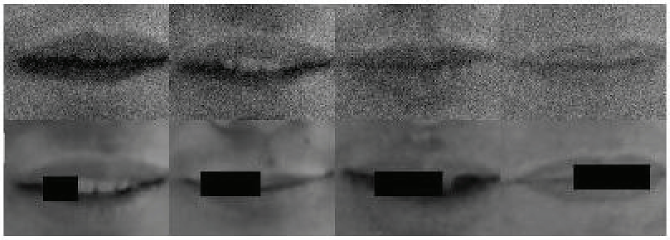

Fig. 4. First row: some samples of the cropped mouth patches corrupted by white Gaussian noise. Second row: some samples of the partially occluded mouth patches.

results in Figure 3. The average accuracy on the whole SD and SI test set are summarized and shown as Table I. According to the experimental results, we can see that our proposed method outperforms previous methods resort to hand-crafted features and SVM.

C. Robustness investigation

To demonstrate the robustness of our method in coping with large illumination variations and occlusion, we perform the liveness attack tests and distort the test images in two ways: (1) We corrupt the SD and SI test sets by adding white Gaussian noise, which arises in digital images caused by poor illumination or transmission, with a zero mean and a standard deviation of 0.01 for each pixel. (2) We randomly shelter the lip region from 20% to 70% with black bars for each mouth patch. Some mouth samples for the liveness attack tests are shown as Figure 4. For simplicity, the SD and SI test sets corrupted by the noise are called the SDN test set and the SIN test set, respectively. The partially occluded SD and SI test sets are called the SDO test set and the SIO test set, respectively.

The experimental results are summarized in Table II. We can see that our model achieve higher accuracy than the other methods even in the noisy situations. The outperforming of our model demonstrates that our learned convolutional layers have extracted robust features in coping with the variations in the wild. The SDO and SIO test accuracy are roughly the same, which indicates that the occluded mouth can be regarded as being from a new subject. In the SDO and SIO test, the average accuracy of our model is about 20% higher than the HOG+SVM and LBP+SVM approaches, which show that our model is able to estimate the state with a high degree of accuracy since our method is less sensitive to occluded inputs. The average accuracy of our method on SIN test set is only 0.8% lower than the accuracy on the SI test set, whereas the average accuracy of LBP+SVM and HOG+SVM on SIN test set are 6.52% and 1.35% lower than the accuracy on the SI test set, respectively. We can see that the performances of SVM+HOG and SVM+LBP methods are much more inferior if the inputs are corrupted by white Gaussian noise or black bars, which indicates that hand-crafted feature extractors fail to represent reliable features to classify the state of mouth in unconstrained conditions. For example, in the worst case as shown in Figure 5, the average accuracy of SVM+HOG method on the SIN uttering digit 5 subset is only 12.3%.

TABLE II
SUMMARY OF THE LIVENESS ATTACK TEST AVERAGE ACCURACY OF DIFFERENT METHODS

Setting	HOG+SVM	LBP+SVM	Ours
SDN test Accuracy	72.9%	73.3%	87.5%
SIN test Accuracy	64.4%	63.1%	83.6%
SDO test Accuracy	59.4%	58.5%	84.1%
SIO test Accuracy	60.2%	60.0%	82.3%

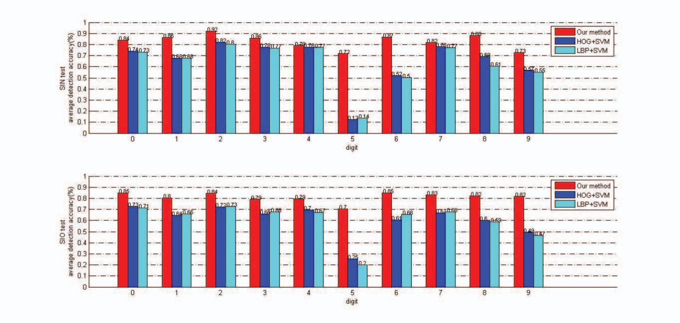

Fig. 5. The SIN and SIO test accuracy of the different methods on the 10 subsets. The experimental results show that our method is more robust to noise attack and the accuracy on the SIN test set almost remains the same as the accuracy on the SI test set.

IV. Conclusions

In this paper, we have proposed a convolutional neural networks method to address the mouth state estimation problem in the wild. Our model takes the cropped mouth patches as an input. The learned convolutional layers automatically extract robust features to deal with illumination variation, irregular noise, and vicious occlusion. We have also established a dataset with challenging images for mouth state estimation. Experimental results have shown that our method outperforms previous methods that resort to manually designed features and SVM, especially in the noisy situations.

V. Acknowledgments

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] Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Describing people: A poselet-based approach to attribute classification. In *2011 IEEE International Conference on Computer Vision (ICCV)*, pages 1543–1550. IEEE, 2011.

[2] Christian Bouvier, Alexandre Benoit, Alice Caplier, and Pierre-Yves Coulon. Open or closed mouth state detection: static supervised classification based on log-polar signature. In *Advanced Concepts for Intelligent Vision Systems*, pages 1093–1102. Springer, 2008.

[3] Linlin Cao, Ran He, and Bao-Gang Hu. Locally imposing function for generalized constraint neural networks-a study on equality constraints. *arXiv preprint arXiv:1604.05198*, 2016.

[4] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[5] Cunjian Chen, Antitza Dantcheva, and Arun Ross. Automatic facial makeup detection with application in face recognition. In *2013 International Conference on Biometrics (ICB)*, pages 1–8. IEEE, 2013.

[6] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 886–893. IEEE, 2005.

[7] Ankur Datta, Rogerio Feris, and Daniel Vaqu. Hierarchical ranking of facial attributes. In *International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011)*, pages 36–42. IEEE, 2011.

[8] Ran He, Yinghao Cai, Tieniu Tan, and Larry Davis. Learning predictable binary codes for face indexing. *Pattern Recognition*, 48(10):3160–3168, 2015.

[9] Marko Heikkila and Matti Pietikainen. A texture-based method for modeling the background and detecting moving objects. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):657–662, 2006.

[10] Neeraj Kumar, Alexander C Berg, Peter N Belhumeur, and Shree K Nayar. Attribute and simile classifiers for face verification. In *12th International Conference on Computer Vision*, pages 365–372. IEEE, 2009.

[11] Haoxiang Li, Zhe Lin, Xiaohui Shen, Jonathan Brandt, and Gang Hua. A convolutional neural network cascade for face detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[12] Yuru Pei, Tae-Kyun Kim, and Hongbin Zha. Unsupervised random forest manifold alignment for lipreading. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 129–136, 2013.

[13] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep convolutional network cascade for facial point detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3476–3483, 2013.

[14] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages I–511. IEEE, 2001.

[15] Xiaoyu Wang, Tony X Han, and Shuicheng Yan. An hog-lbp human detector with partial occlusion handling. In *12th International Conference on Computer Vision*, pages 32–39. IEEE, 2009.

[16] Xiang Wu, Ran He, and Zhenan Sun. A lightened cnn for deep face representation. *arXiv preprint arXiv:1511.02683*, 2015.

[17] Xuehan Xiong and Fernando Torre. Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 532–539, 2013.

[18] Pong C Yuen, Jian-Huang Lai, and QY Huang. Mouth state estimation in mobile computing environment. In *Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 705–710. IEEE, 2004.

[19] Shu Zhang, Ran He, Zhenan Sun, and Tieniu Tan. Multi-task convnet for blind face inpainting with application to face verification. In *2016 international conference on biometrics (ICB)*. IEEE, 2016.

[20] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *Computer Vision–ECCV 2014*, pages 94–108. Springer, 2014.