Facial Smile Detection Based on Deep Learning Features

Kaihao Zhang[1,2], Yongzhen Huang[2], Hong Wu[1], Liang Wang[2]
[1]University of Electronic Science and Technology of China, Chengdu, China
[2]National Laboratory of Pattern Recognition (NLPR),
Institute of Automation, Chinese Academy of Sciences, Beijing, China
{kaihao.zhang, yzhuang, wangliang}@nlpr.ia.ac.cn, hwu@uestc.edu.cn

Abstract

Smile detection from facial images is a specialized task in facial expression analysis with many potential applications such as smiling payment, patient monitoring and photo selection. The current methods on this study are to represent face with low-level features, followed by a strong classifier. However, these manual features cannot well discover information implied in facial images for smile detection. In this paper, we propose to extract high-level features by a well-designed deep convolutional networks (CNN). A key contribution of this work is that we use both recognition and verification signals as supervision to learn expression features, which is helpful to reduce same-expression variations and enlarge different-expression differences. Our method is end-to-end, without complex pre-processing often used in traditional methods. High-level features are taken from the last hidden layer neuron activations of deep CNN, and fed into a soft-max classifier to estimate. Experimental results show that our proposed method is very effective, which outperforms the state-of-the-art methods. On the GENKI smile detection dataset, our method reduces the error rate by 21% compared with the previous best method.

1. Introduction

Facial expression analysis play an important role in understanding human emotions and behaviors. The varying facial expression is an important cue for psychology of emotions. As one of the most common facial expressions that occurs in people's daily life, smile often indicates agreement, satisfaction, happiness, etc. Detecting smiles can be used to estimate people's mental state which have broad applications in many areas such as smiling payment, patient monitoring and photo selection. For example, when people want to pay, they can smile at the camera instead of inputting complex passwords. In this way, they can increase convenience of operation as well as avoid leaking password to the people around them.

Figure 1. Examples of smile faces (top two rows) and neutral faces (bottom two rows) on the GENKI smile detection dataset.

Current state of the art approaches to smile detection typically represent faces with traditional handcrafted features, such as: Local binary patterns (LBP) features [11], Histograms of oriented gradients (HOG) features [3], or Scale invariant features transform (SIFT) features [12]. Shan et al. [11] built face representation based on histograms of LBP features. They formulate Boosted-LBP to extract the most discriminant LBP features, and using Support vector machine (SVM) classifiers to obtain the satisfactory recognition performance. Their experiments illustrate that LBP features are effective and efficient for low resolution images. Dahmane et al. [3] study histograms of HOG features from dense grids for facial expression recognition. Their representation followed by nonlinear SVM outperforms the method based on LBP. Sikka et al. [12] use SIFT features which also yield competitive performance.

More recently, a number of researchers propose some new features or classifiers to smile detection. Shan [10] uses the intensity differences between pixels in the grayscale

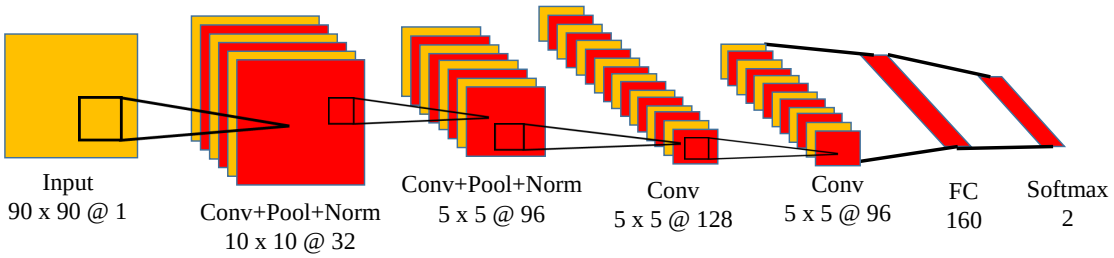

Figure 2. The architecture of the our basic deep CNN for smile detection. The number below the each layers demotes the dimension of each map and the map number. The output is a two-value label which is smile or non-smile.

face images as features. They adopt AdaBoost to choose and combine weak classifiers to form a strong classifiers for smile detection. Jain et al. [5] employ multi-scale gaussian derivatives combined with a SVM which is an effective approach to smile detection. Kahou et al. [6] build high dimensional binary features composed of dense census transformed vectors to smile detection which further improve the accuracy of smile detection. However, the handcrafted descriptors are low-level features which cannot well discover information implied in facial images for smile detection. Meanwhile, deep convolutional neural networks has achieved surprisingly encouraging performance in many vision tasks, such as image classification [7, 14], object recognition [4] and face verification [13], which motivates us to exploit deep convolutional networks for smile detection.

In this paper, we propose to extract high-level features by a well-designed deep CNN using the recognition signal (an image labeled as smile or non-smile) as supervision. One image is put into our CNN without complicated preprocessing, and low-level features are extracted in the bottom layers. The features are fused to be the input of higher layers and high-level features are generated in the top layers. Finally, the high-level features can be fed into a two-way softmax classifier to predict. Thanks to the deep architecture's large learning capacity, we can extracted high-level features which have large different-expression differences. Experimental results demonstrate that our method is very effective, which outperforms the state-of-the-art methods. Beside, we find that most traditional methods only try to increase the different-expression differences, but they do not pay attention to reducing the same-expression variations. In order to extracted more powerful features, we designed a new deep CNN uses both recognition signal and verification signal (a pair of images labeled as same-expression or different-expression) as supervision. The recognition signal can pull apart the features of different expressions, while the verification signal can reduce the same-expression variations. In this way, we can extracted more powerful features. The results show that this method can further improve the accuracy of smile detection. On the GENKI smile detection dataset which is shown in Figure 1, our method reduces the error rate by 21% compared with the previous best method.

2. Our Method

In our method, we first carefully design a basic structure of CNN. In order to reduce same-expression variations and enlarge different-expression differences, we modify the basic structure to a new structure of CNN which uses both recognition and verification signals as supervision.

2.1. Basic Structure of CNN

Overall architecture. Now we are ready to describe the structure of our CNN. As depicted in Figure 2, our CNN architecture has 6 stages containing different layers. The first two stages include a convolutional layer followed by a pooling layer and a local response normalization layer. The third and fourth stages contain only a convolutional layer. The fifth stage includes a fully-connected layer. Finally, this networks is trained via a two-way soft-max classifier at the last layer to predict smile or non-smile. The first convolutional layer uses 32 filters with a stride of 4 and each filter is with the size of $11\times11\times1$. The second convolutional layer increases the number of filter up to 96 filters with a stride of 1 and each filter is $5\times5\times32$. The third convolutional layer use 128 filters, and the fourth convolutional layer 96, all of size 5×5 pixels with a stride of 1. The fifth fully-connected layer has 160 neurons. The high-level features are taken from the fifth fully-connected layer neuron activations, and fed into a soft-max classifier to estimate.

The widely used activation functions are sigmoid function and rectified linear unit (ReLU) function. In this paper, we use ReLU as the activation function in the convolutional layers and fully-connected layers. The reason we choose it is considering the training time with the gradient descent algorithm, the non-saturating nonlinearity is faster than these non-saturating nonlinearities [7]. In addition, ReLU function has better fitting abilities than the sigmoid function. In order to increase the translation invariance and avoid overfitting, we choose max-pooling with a neighboring region size of 3×3 and a stride of 2.

Implementation details. we detect the face with a facial detection algorithm. Each face image is cropped and

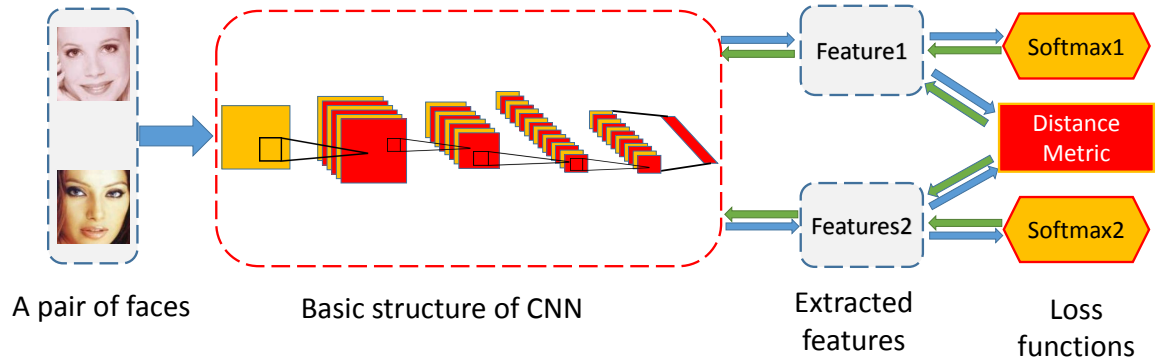

Figure 3. The two-signal guided structure of CNN. A pair of images are sent into the structure, with our proposed basic of CNN to generate their feature representations. Recognition and verification signals are used as supervision to update all the weights of network according to two kinds of loss function.

aligned according to the face location. To eliminate the impact of image sizes, we resize the images to the 96×96×1. And then randomly crop a 90×90 patch at the image which is put into our CNN. In this way, we can greatly expand the training samples. To further enlarge the training dataset, the patches are also randomly flipped. Our model is trained by back-propagation with logistic loss over the predicted scores using the 2-way soft-max function. we initialize weights using a Gaussian distribution with zero mean and a standard deviation of 0.01. Each layer has some same parameters: the bias is initialized as 1, the momentum is set as 0.9 and the weight decay is set as 0.005. The dropout rate for the fifth fully-connected layers is set to be 0.5 to reduce overfitting. We update all the weights after learning the mini-batch with the size of 256 in each iteration.

2.2. Two-signal Guided Structure of CNN

The basic structure of CNN achieves satisfactory performance which outperforms the state-of-the-art methods. The method uses the recognition signal as supervisory to extract features. Since the features have to be classified into different classes, this recognition signal is useful to pull apart the features of different expressions. In order to extract more powerful features, we employ an additional expression verification signal which is not only helpful to enlarge different-expression differences, but also can reduce same-expression variations. We modify the basic structure of CNN to a new structure which is shown in Figure 3.

Namely, our new CNN learns features with two supervisory signals. The first is expression recognition signal, which can classifiers each face image into smile or non-smile. The network is trained to minimize the cross-entropy loss, which is defined as

$$\mathrm{Re}Loss(p,q) = -\sum_{x} p(x) log\, q(x), \qquad (1)$$

where x is the expression feature vector, $p(x)$ is the "true" distribution, and $q(x)$ is the predicted probability distribution. We can learn the features with large different-expression variations.

The second is expression verification signal, which is effective to reduce the variations of features extracted from same-expression image. We adopt the loss function based on the L2 norm [9], which is formulated as:

$$VeLoss(x_i, x_j, c_{ij}) = \qquad (2)$$
$$\begin{cases} \frac{1}{2}\left\| f(x_i) - f(x_j) \right\|_2^2 & if\ c_{ij} = 1 \\ \frac{1}{2}\max\left(0, \delta - \left\| f(x_i) - f(x_j) \right\|_2\right)^2 & if\ c_{ij} = 0 \end{cases},$$

where x_i and y_j are two input images, and their features are f_i and f_j extracted from fifth fully-connected layer. $c_{ij} = 1$ means that x_i and y_j have same facial expression. In this case, we enforce f_i and f_j to be close (in the L2 norm). $c_{ij} = 0$ means the two input images are different expressions, and we push their features apart. δ is the size of the margin. We hope the distance of different-expression images is larger than δ. In the training phase, two signals are weighted by a hyperparameter k. We will search a satisfying value of k in the next section.

3. Experiments

3.1. GENKI-4K Database

We carry our experiments on the publicly available GENKI-4K database [1], which is shown in Figure 1. The database contains 4000 face images of a wide range of subjects with different ages and races, as well as variability in pose, illumination and imaging conditions. Among these pictures, 2162 images are labeled as smile and 1838 images are labeled as non-smile. Unlike detecting smiles in laboratory-controlled databases, this database represents the real-world scenarios for detection which is more challenging.

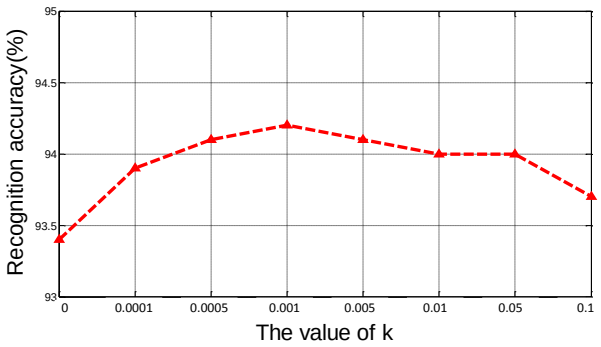

Figure 4. Smile detection accuracy of our proposed CNN versus different values of k.

In our experiments, the images are converted to grayscale. The database is equally divided into four folds. We perform four-fold validation experiments on the database. In each time, 3000 images are selected as training images and the rest 1000 images are used for testing. In the following experiments, we report the average detection accuracy and the standard deviation.

3.2. Balancing the Two Supervisory Signals

As mentioned before, our method uses recognition and verification signals as supervision. The two signals are weighted by a hyperparameter k. In this section, we investigate the weight between the two supervisory signals on learning feature by varying k from 0 to 0.1. When $k = 0$, only the recognition signal at work. With the increasing of k, the verification signal plays a more and more important role in our method. Figure 4 shows the recognition accuracy on the test set as the change of k. It denotes that verification signal is helpful for smile detection. We can observe from the Figure 4 that the performance of our model becomes better as the k increases, and achieves the highest accuracy when $k = 0.005$. When the value of k becomes larger than 0.005, the accuracy will decline. So we set k to 0.005 in our next experiments.

Thanks to the deep architecture's learning capacity and the two supervisory signals, we can extracted high-level features which is the key to detect smile faces. Examples of the features learned from the training set and extracted from the test set are shown in Figure 5. We can clearly find that even though faces have large variations in age, ethnicity and gender, face features of the same expression are very similar and different expressions have large variations.

3.3. Methods Comparison

Several researchers have explored the problem of smile detection. In this section, we compare our method with five recent methods. Table 1 shows the smile detection results of different smile detection algorithms on the GENKI-4K database. An et al. [2] report 88.5% accuracy using

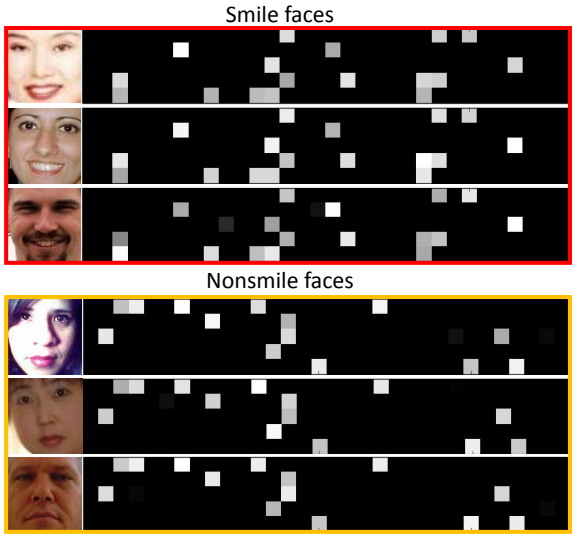

Figure 5. Samples of the learned expression features. The first three images are faces with smile, and the latter three are neutral faces. Left column shows test images and their features extracted from our model are shown in the right. The features are in 160-dimensional.

method	feature	classifier	accuracy
An et al. [2]	HOG	ELM	88.5%
Shan et al. [10]	Pixel	AdaBoost	89.7±0.45%
Liu et al. [8]	HOG	SVM	92.29±0.81%
Jain et al. [5]	Guassian	SVM	92.97%
Kahou et al. [6]	LBP	SVM	93.2±0.92%
CNN-Basic	**CNN**	**Softmax**	**93.6±0.47%**
CNN-2Loss	**CNN**	**Softmax**	**94.6±0.29%**

Table 1. The experimental results on the GENKI-4K database. The previous best method is proposed by Kahou et al. [6].

HOG features and their proposed Extreme Learning Machine (ELM) classifier. Shan et al. [10] use pixel difference as the feature descriptor and use AdaBoost for both feature selection and classification, which achieve 89.7±0.45 accuracy. Liu et al. [8] increase to the 93.2±0.92% accuracy with the help of unlabeled reference data. Jain et al. [5] obtain 92.97% accuracy using multi-scale gaussian derivatives combines with a SVM. However, they achieve this accuracy by removing some images which are ambiguous or of serious illumination problems. The previous best method used for smile detection is proposed by Kahou et al. [6], which achieves the 93.2±0.92% accuracy. They extracted high dimensional binary features and classify the faces by a SVM. Results listed in Table 1 indicate that our method outperforms the state-of-the-art method which achieves an even higher 94.6±0.29% smile detection accuracy.

According to the results shown in Table 1, we can draw two clear conclusions as follows. Firstly, due to our care-

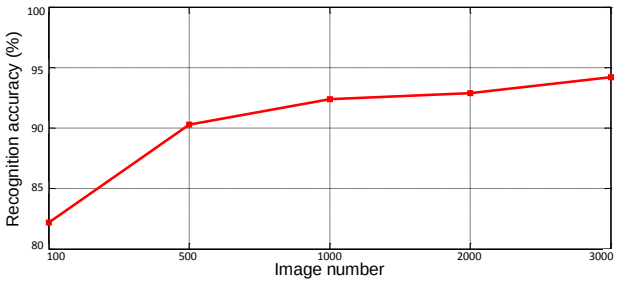

Figure 6. smile detection accuracy of the two-signal guided structure of CNN versus different number of images on the GENKI-4K database.

fully design, the proposed basic structure of CNN outperforms the other five algorithms including the previous best method, which implies that deep convolution neural networks is an effective method to solve the problem of smile detection. Secondly, the two-signal guided structure of CNN improves smile detection of CNN-Basic. The reason is that the new structure of CNN uses both recognition and verification signals as supervision which are helpful to reduce same-expression variations and enlarge different-expression differences. The results show that the two signals are effective to extracted more powerful feature descriptors for smile detection.

We also investigate the influence of the data size on our proposed CNN-2Loss. In particular, we train our model with an increasing number of face images from 100 to 3000. Figure 6 shows the variation of accuracy rate with the number of training set. We can observe from the figure that the performance becomes better with the data size increasing. The experimental results imply that the data size is a key for face detection, and we are likely to get better results if the size of training set becomes larger.

4. Conclusion and Future Work

In this paper, we proposed two deep CNNs to address smile detection. Firstly, we carefully design a basic structure of CNN for smile detection which achieves comparable results to the state-of-the-art algorithm. In order to learn more powerful expression features, we modify the basic model to a new structure of CNN using both the recognition and verification signals as supervision which reduce same-expression variations and enlarge different-expression differences. Our model have shown remarkable results on the GENKI-4K database which reduces the error rate by 21% comparing with the previous best method. Some readers may find that our current experiments are conducted on a smile face detection database. However, our method can be easily extended for other face expressions, e.g., replacing the two-way soft-max classifier with a n-way soft-max classifier. This is not the focus of this paper, and we leave it to feature work.

5. Acknowledgement

This work is jointly supported by National Basic Research Program of China (2012CB316300), National Natural Science Foundation of China (61135002, 61420106015), CCF-Tencent Open Fund, SAMSUNG GRO Program and 360 OpenLab Program.

References

[1] The MPLab GENKI-4K Database: http: //mplab.ucsd.edu/. 3

[2] L. An, S. Yang, and B. Bhanu. Efficient smile detection by extreme learning machine. *Neurocomputing*, 2015. 4

[3] M. Dahmane and J. Meunier. Emotion recognition using dynamic grid-based hog features. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, 2011. 1

[4] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, 2014. 2

[5] V. Jain and J. Crowley. Smile detection using multi-scale gaussian derivatives. In *12th WSEAS International Conference on Signal Processing, Robotics and Automation*, 2013. 2, 4

[6] S. E. Kahou, P. Froumenty, and C. Pal. Facial expression analysis based on high dimensional binary features. In *Computer Vision-ECCV 2014 Workshops*, 2014. 2, 4

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012. 2

[8] M. Liu, S. Li, S. Shan, and X. Chen. Enhancing expression recognition in the wild with unlabeled reference data. In *Computer Vision–ACCV 2012*. 2013. 4

[9] H. Mobahi, R. Collobert, and J. Weston. Deep learning from temporal coherence in video. In *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009. 3

[10] C. Shan. Smile detection by boosting pixel differences. *Image Processing, IEEE Transactions on*, 2012. 1, 4

[11] C. Shan, S. Gong, and P. W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 2009. 1

[12] K. Sikka, T. Wu, J. Susskind, and M. Bartlett. Exploring bag of words architectures in the facial expression domain. In *Computer Vision–ECCV 2012. Workshops and Demonstrations*, 2012. 1

[13] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, 2014. 2

[14] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *arXiv preprint arXiv:1409.4842*, 2014. 2