Kinship Verification with Deep Convolutional Neural Networks

Kaihao Zhang[12]
kaihao.zhang@nlpr.ia.ac.cn
Yongzhen Huang[2]
yzhuang@nlpr.ia.ac.cn
Hong Wu[1]
hwu@uestc.edu.cn
Liang Wang[2]
wangliang@nlpr.ia.ac.cn

[1] Statistical Machine Intelligence & LEarning (SMILE)
Big Data Research Center
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

Kinship verification from facial images is an interesting and challenging problem. The current algorithms on this topic typically represent faces with multiple low-level features, followed by a shallow learning model. In this paper, we propose to extract high-level features for kinship verification based on deep convolutional neural networks. Our method is end-to-end, without complex pre-processing often used in traditional methods. The high-level features are produced from the neuron activations of the last hidden layer, and then fed into a soft-max classifier to verify the kinship of two persons.

We first propose a basic structure of CNN (CNN-Basic) contains three convolutional layers, followed by a fully-connected layer and a soft-max layer. As shown in Figure 1, the input is a pair of 64×64 images with three channels (RGB). Following the input, the first convolutional layer is generated after convolving the input via 16 filters with a stride of 1. Each filter is with the size $5 \times 5 \times 6$. The second convolutional layer filters the input of the previous layer with 64 kernels of size $5 \times 5 \times 16$. The third convolutional layer contains 128 kernels of the size $5 \times 5 \times 64$. After the convolutional layers, a fully-connected layer projects the extracted features into a subspace with 640 neurons. Max-pooling layers follow the first and second convolutional layers. Finally, this network is trained via a two-way soft-max classifier at the top layer.

We adopt the ReLU function [1] as the activation function of the convolution layers, which has been shown to achieve better performance than the sigmoid function. With ReLU, the convolution operation is formulated as

$$y^{j(r)} = \max\left(0, b^{j(r)} + \sum_i w^{ij(r)} * x^{i(r)}\right), \quad (1)$$

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 $\times$ denotes the convolutional operation.

We choose max-pooling with a neighboring region size of 2×2. Max pooling is helpful to increase the translation invariance and avoid overfitting, which is defined as

$$y^i_{j,k} = \max_{0 \le m,n \le s} \left\{ x^i_{j \cdot s+m, k \cdot s+n} \right\}, \quad (2)$$

where $y^i_{j,k}$ denotes the outputs of the i-th feature map in the location of (j,k). Similarly, $x^i_{j,k}$ denotes the value of location (j,k) in the i-th feature map.

The CNN is trained by back-propagation with logistic loss over the predicted scores using the soft-max function. To initialize weights, we use a Gaussian distribution with zero mean and a standard deviation of 0.01. The biases are initialized as zeros. In each iteration, we update all the weights after learning the mini-batch with the size of 128. In all layers, the momentum is set as 0.9 and the weight decay is set as 0.005. To expand the training set, we also randomly flip images during training.

When a subject is demanded to verify the kinship from two face images, it is highly possible that the key-points are focused, such as their eyes, mouth and nose. We consider that the facial key-points have a significant impact on kinship analysis, and thus design a key-points-based feature representation for kinship verification. In particular, we detect the centers of two eyes, the corners of the mouth and the nose with a facial point detection algorithm [2]. Then each face image is cropped and aligned according to the five key-points. To extract more complementary information, we also crop other five face regions without key-points detection. The five images are the original image and its four local regions,

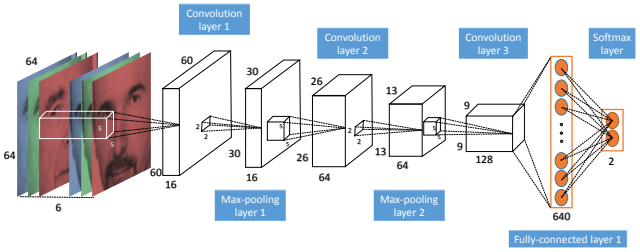

Figure 1: The proposed architecture of basic CNN for kinship verification. For all layers, the length of each cuboid is the map number, and the width and height of each cuboid are the dimension of each map.

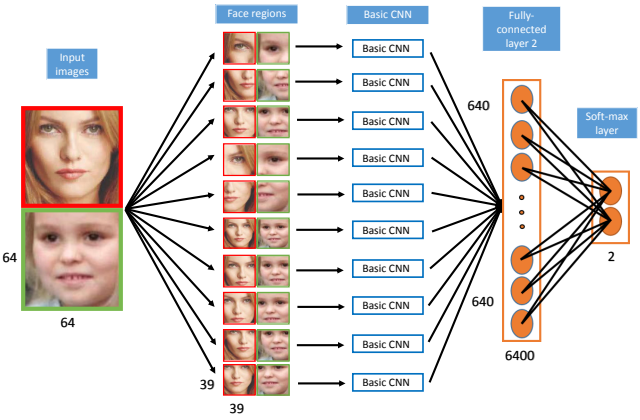

Figure 2: Overview of the proposed CNN-Points structure for kinship verification. The input is a pair of RGB images, which are cropped into ten face regions and fed into different basic CNNs.

i.e., the top-left corner, the top-right corner, bottom-left corner, bottom-right corner.

In order to improve kinship verification with these face regions, we propose a new structure (CNN-Points) which is shown in Figure 2. The new structure contains 10 basic CNNs (see Figure 1), each of which receives a pair of face regions. Ten sets of 640-dimensional features are produced from the last hidden layer of the basic CNNs. The last hidden layer of the CNN-Points is fully-connected to the ten basic CNNs, which is defined as

$$y^i = f\left(\sum_{k=1}^{10} \sum_{j=1}^{640} w^i_{j,k} * x_{j,k} + b^i\right), \quad (3)$$

where y^j is the output of the i-th neuron activation, $w^i_{j,k}$ denotes the weight between the input features and the i-th neuron, and $f(\cdot)$ is chosen to be the sigmoid function. The final representation is fed into a soft-max classifier to predict the kinship of two persons.

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[2] Y. Sun, X. L. Wang, and X. O. Tang. Deep convolutional network cascade for facial point detection. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3476–3483. IEEE, 2013.