Hybrid Deep Learning for Face Verification

Cheng Guan

Abstract

This paper proposes a hybrid convolutional network (ConvNet)-Restricted Boltzmann Machine(RBM) model for face verification in wild conditions. A key contribution of this work is to directly learn relational visual features, which indicate identity similarities, from raw pixels of face pairs with a hybrid deep network. The deep ConvNets in our model mimic the primary visual cortex to jointly extract local relational visual features from two face images compared with the learned filter pairs. These relational features are further processed through multiple layers to extract high-level and global features.

1. Introduction

Face recognition has been extensively studied in recent decades [2–4]. This paper addresses the key challenge of computing the similarity of two face images given their large intra-personal variations in poses, illuminations, expressions, ages, makeups, and occlusions. It becomes more difficult when faces to be compared are acquired in the wild. We focus on the task of face verification, which aims to determine whether two face images belong to the same identity.

Existing methods generally address the problem in two steps: feature extraction and recognition. In the feature extraction stage, a variety of hand-crafted features are used. Although some learning-based feature extraction approaches are proposed, their optimization targets are not directly related to face identity [1]. Therefore,the features extracted encode intra-personal variations. More importantly,existing approaches extract features from each image separately and compare them at later stages . Some important correlations between the two compared images have been lost at the feature extraction stage.

All of the issues discussed above motivate us to learn a hybrid deep network to compute face similarities. A high-level illustration of our model is shown in Figure 1. Our model has several unique features, as outlined below.

2. Related Work

All existing methods for face verification start by extracting features from two faces in comparison separately. A variety of low-level features are commonly used , including the hand-crafted features like LBP and its variants [5], SIFT , Gabor and the learned LE features. Some methods generated midlevel features with variants of convolutional deep belief networks (CDBN) or ConvNets . They are not learned with the supervision of identity matching. Thus variations other than identity are encoded in the features, such as poses, illumination, and expressions, which constitute the main impediment to face recognition.

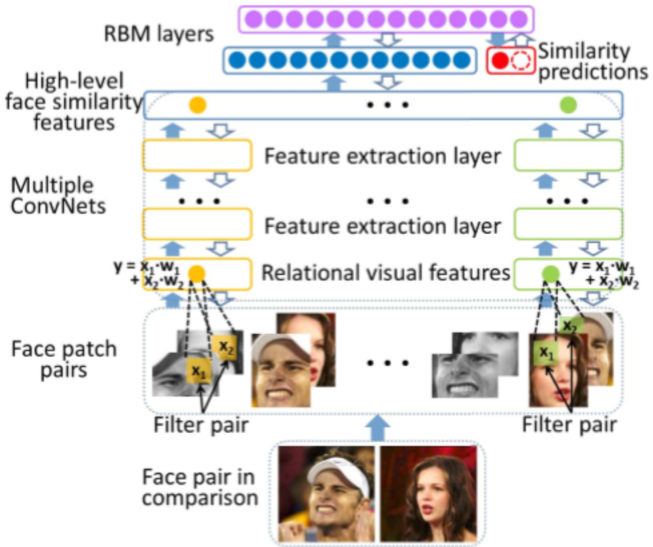

Figure 1. The hybrid ConvNet-RBM model. Solid and hollow arrows show forward and back propagation directions

Many face recognition models are shallow structures, and need high-dimensional over-completed feature representations to learn the complex mappings from pairs of noisy features to face similarities ; otherwise, the models may suffer from inferior performance. Many methods used linear SVM to make the same-ordifferent verification decisions.

3. The hybrid ConvNet-RBM model

3.1. Architecture overview

We detect the two eye centers and mouth center with the facial point detection method proposed . Faces are aligned by similarity transformation according to the three points. Figure 2 is an overview of our hybrid ConvNet-RBM model,which is a cascade of deep ConvNet groups, two levels of average pooling, and Classification RBM.The lower part of our hybrid model contains 12 groups, each of which contains five ConvNets.The lower part of our hybrid model contains 12 groups, each of which contains five ConvNets. Figure 3 shows the structure of one ConvNet.

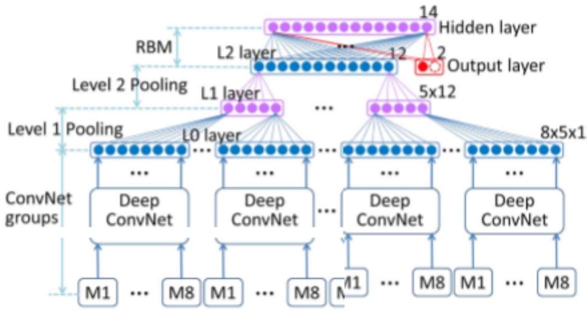

Figure 2. Architecture of the hybrid ConvNet-RBM model. Neuron (or feature) number is marked beside each layer.

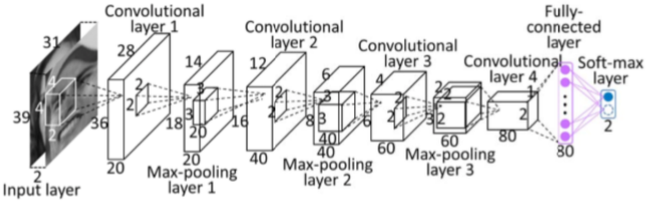

Figure 3. The structure of one ConvNet. The map numbers and dimensions of the input layer and all the convolutional and max-pooling layers are illustrated as the length, width, and height of cuboids

3.2. Deep ConvNets

Our deep ConvNets contain four convolutional layers (followed by max-pooling). The operation in each convolutional layer can be expressed as

$$y_j^r = max\left(0, b_j^r + \sum_i k_{ij}^r * x_i^r\right) \quad (1)$$

where $*$ denotes convolution, x_i and y_i are the *i*-th input map and the *j*-th output map respectively, k_{ij} is the convolution kernel (filter) connecting the *i*-th input map and the *j*-th output map, and b_j is the bias for the *j*-th output map. $max\,(0, \cdot)$ is the non-linear activation function, and is operated element-wise.

References

[1] Z. Cao, Q. Yin, X. Tang, and J. Sun. Face recognition with learning-based descriptor. In *CVPR*, 2010.

[2] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *ICCV*, 2009.

[3] X. Wang and X. Tang. Dual-space linear discriminant analysis for face recognition. In *CVPR*, 2004.

[4] X. Wang and X. Tang. A unified framework for subspace face recognition. *IEEE TPAMI*, 2004.

[5] L. Wolf, T. Hassner, and Y. Taigman. An associate-predict model for face recognition. In *ECCV*.