Pattern Recognition (xxxx) xxxx–xxxx

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

A weakly supervised method for makeup-invariant face verification

Yao Sun[a], Lejian Ren[b], Zhen Wei[a], Bin Liu[c], Yanlong Zhai[b], Si Liu[a,*]

[a] State Key Laboratory of Information Security (SKLOIS), Institute of Information Engineering, Chinese Academy of Sciences, No. A89, Minzhuang Road, Haidian District, Beijing 100093, China
[b] School of Computer Science & Technology, Beijing Institute of Technology, Beijing, China
[c] Moshanghua Tech Co. Ltd., Beijing, China

ARTICLE INFO

Keywords:
Face verification
Makeup-invariant
Weakly supervised method
Video context
Triplet loss function

ABSTRACT

Face verification, which aims to determine whether two face images belong to the same identity, is an important task in multimedia area. Face verification becomes more challenging when the person is wearing makeup. However, collecting sufficient makeup and non-makeup image pairs are tedious, which brings great challenges for deep learning methods of face verification. In this paper, we propose a new weakly supervised method for face verification. Our method takes advantages of the plentiful video resources available from the Internet. Our face verification model is pre-trained on the free videos and fine-tuned on small makeup and non-makeup datasets. To fully exploit the video contexts and the limited makeup and non-makeup datasets, many techniques are used to improve the performance. A novel loss function with a triplet term and two pairwise terms is defined, and multiple facial parts are combined by the proposed voting strategy to generate better verification results. Experiments on a benchmark dataset (Guo et al., 2014) [1] and a newly collected face dataset show the priority of the proposed method.

1. Introduction

Human face verification [2,3] has been extensively studied and has various practical applications. In human perception and psychology studies [4], it has been revealed that heavy makeup can significantly decrease the human ability of recognizing faces. As shown in Fig. 1, significant appearance changes can be observed for individuals with and without makeup.

Convolutional Neural Network (CNN) extracts features from the raw image data, and has achieved great success in many tasks such as image recognition [5–8], fashion [9–11] and general face recognition [12]. Although methods based on deep learning have achieved competitive results, there is a great difficulty to use CNN for our task, which needs to do face verification to those with heavy makeups. The difficulty is that it is hard to find or collect enough makeup and non-makeup face pairs for training the network.

In this paper, due to the lack of large amount makeup and non-makeup datasets, we propose a weakly supervised makeup-invariant method for face verification. Our method takes advantages of large amount of free video contexts from the Internet, and only needs small sets of labelled makeup and non-makeup images to fine-tune. The pipeline of our method is demonstrated in Fig. 1. First, a large amount of free videos are downloaded from the web. Next, face detection and tracking are used to generate the *positive pairs*. By saying positive pairs, we mean that the faces belong to the same persons, while the persons appearing in the successive frames are usually regarded as identical. Negative pairs can be selected by some strategies or simple random selected. Then we can get triplet pairs from the video contexts. These triplets are used to pre-train a model. With the pre-trained weights, we only need to collect small amount of before–after makeup faces in the next step. By fine-tuning the model on these small amounts of makeup and non-makeup images, we can obtain a face verification model robust to makeups. To avoid overfiting in the fine-tuning stage, we use many techniques in our network, including using larger features, extra pairwise losses, mirror images, and local facial parts. These techniques will be detailed in Section 3. Experiments on a benchmark dataset [1] and a newly collected face dataset show the advantage of the proposed method.

Our major contributions can be concluded as follows.

(1) We propose a weakly supervised method for face verification that is robust to cosmetic changes and achieves state-of-the-art performance.
(2) We propose a deep framework for makeup-invariant face verifica-

* Corresponding author.
E-mail addresses: sunyao@iie.ac.cn (Y. Sun), renlejian@outlook.com (L. Ren), zhen.wei@hotmail.com (Z. Wei), liubin@dress-plus.com (B. Liu), ylzhai@bit.edu.cn (Y. Zhai), liusi@iie.ac.cn (S. Liu).

http://dx.doi.org/10.1016/j.patcog.2017.01.011
Received 15 July 2016; Received in revised form 6 January 2017; Accepted 7 January 2017

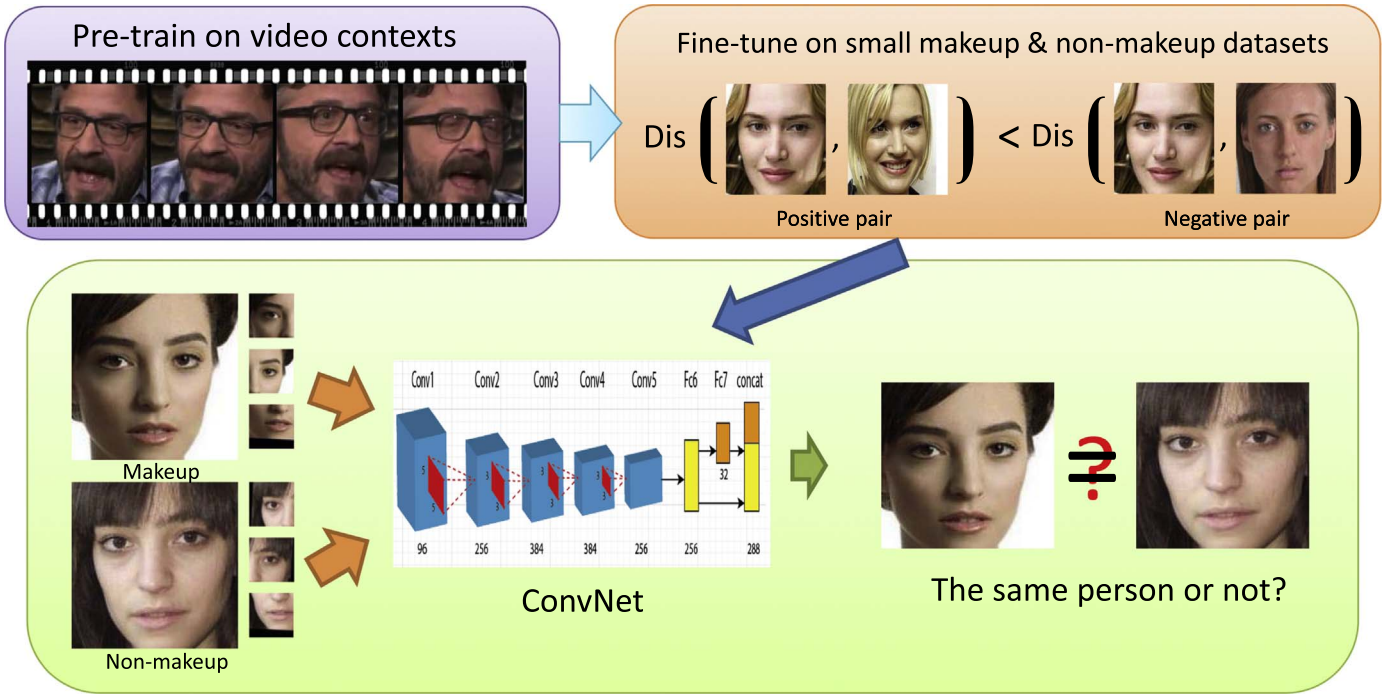

Fig. 1. A weakly supervised method for face verification.

tion, which has two distinguishing properties, i.e. (a) it utilizes freely available videos for pre-training and (b) multiple facial parts are combined to achieve better performance.

(3) We collect a large scale video face dataset and a before–after makeup pair dataset, which can be used as the benchmark in the further studies.

2. Related works

Face verification: Recently, there are increasing interest and great progress in the face verification task. Taigman et al. [13] proposed DeepFace which carefully designed both the alignment step and the representation step by employing explicit 3D face modelling in order to apply a piecewise affine transformation, and derived a face representation from a nine-layer deep neural network. Sun et al. [3] proposed a hybrid convolutional network (ConvNet)-Restricted Boltzmann Machine (RBM) model for face verification in wild conditions. Sun et al. [14] proposed DeepID which can be effectively learned through challenging multi-class face identification tasks, whilst they can be generalized to other tasks (such as verification). Sun et al. [15] proposed a Deep IDentification-verification features (DeepID2). The face identification task increases the inter-personal variations by drawing DeepID2 features extracted from different identities apart, while the face verification task reduces the intra-personal variations by pulling DeepID2 features extracted from the same identity together. In [2], Schroff et al. proposed a Facenet system for face recognition and clustering. A triplet network was used to train the network and several sample selecting strategies were discussed as well. There are some other works that address face verification tasks under specific conditions, such as with occlusions [16,17] or in videos [18]. However, these works are not specifically designed for makeup invariant face verification task.

Weakly supervised learning for face recognition: Face data in images and videos are of great volume on the Internet. They can be easily obtained through video websites and social websites. Previous work [14] has proved that getting more face data and more identities involved helps to improve recognition performance.

However, manually labelling such great number of face data is laborious and impractical. Recent researches begin to address more importance on weakly supervised labels. Rim et. al. [19] leveraged weak labelled data for face recognition based on probabilistic graphical models. Chen and Deng [20] built up a challenging unlabelled database and proposed an efficient Self-Learning DCNN structure (SL-DCNN) to handle weakly supervised training for face recognition. In this paper, we make use of unlabelled video context data to generate weakly supervised labels for model pre-training.

Makeup studies: Recently, there are more works focusing on the makeup related studies, such as makeup transfer [21–23] and makeup recommendation [24].

A dual attributes approach [25] was proposed to learn facial attributes in makeup and non-makeup faces separately, and face matching uses the semantic-level attributes to reduce the influence of makeup on low-level features. Another approach [26] is to preprocess face images with a self-quotient image technique to reduce makeup effects before face matching. However, these methods cannot reduce the makeup influence significantly. Guo et al. [1] proposed performing correlation mapping between makeup and non-makeup faces on features extracted from local patches.

3. Methods

A triplet network is presented in our methods and is illustrated in Fig. 2. Based on this network, we take three stages to obtain the final face verification results. Firstly, we pre-train the network on video contexts which are easy to get from public available videos. Next, we fine-tune the pre-trained models on makeup and non-makeup images, and we also fine-tune on several parts of these images in the second stage. In the last stage, by using a *voting* approach, we summarize the verification results obtained both from the whole face images and the parts of the faces, and give a final decision on whether the input two images belong to the same identity.

Details of the method are given in the following.

3.1. The triplet network

Our goal is to learn a discriminative feature representation so that