Pattern Recognition Letters 90 (2017) 36–42

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Large Age-Gap face verification by feature injection in deep networks

Simone Bianco

University of Milano-Bicocca, Viale Sarca 336, Ed. U14, Milan, 20126, Italy

A R T I C L E   I N F O

Article history:
Received 7 June 2016
Available online 14 March 2017

Keywords:
Face verification
Deep learning
Cross-age face verification

A B S T R A C T

This paper introduces a new method for face verification across large age gaps and also a dataset containing variations of age in the wild, the Large Age-Gap (LAG) dataset, with images ranging from child/young to adult/old. The proposed method exploits a deep convolutional neural network (DCNN) pre-trained for the face recognition task on a large dataset and then fine-tuned for the large age-gap face verification task. Fine-tuning is performed in a Siamese architecture using a contrastive loss function. A feature injection layer is introduced to boost verification accuracy, showing the ability of the DCNN to learn a similarity metric leveraging external features. Experimental results on the LAG dataset show that our method is able to outperform the face verification solutions in the state of the art considered.

1. Introduction

Face verification is an important topic in both computer vision, imaging and multimedia. Verification accuracy mainly depends on four elements: face pose, facial expression, illumination, and aging [17]. The greatest part of the works in the state of the art studied the face verification problem in constrained scenarios, controlling and fixing one or more of these four elements.

Recently many researchers achieved or even surpassed human-level performance [9,35] on face verification benchmark taken in unconstrained environments such as the Labeled Faces in the Wild dataset (LFW) [16]. These results have been made possible thanks to the improvement in facial landmark detection and to the increase of the computational power available to train deep models. However, the LFW dataset fixes the aging element: it contains large variations in pose, facial expression, and illumination, but contains very little variation in aging. As people grow, face appearance can be very different, which makes it difficult to recognize people across age. The problem is even harder when large age gaps are considered. In this work large age gap is interpreted in two ways: it refers both to the cases with extreme difference in age (e.g. young vs old) and to the cases with large difference in appearance due to the aging process (e.g. baby vs teenager/adult). Being able to recognize people across large age gaps could be beneficial in many applications: on photo sharing websites such as Facebook and Flickr that are providing services for over ten years; in all the personal photo management applications such as Google Photos and Apple Photo where albums can likely span decades; for the identification of long-lost and found persons.

To address the problem of face verification across large age gaps, in this work a new approach is proposed. Differently from other approaches in the state of the art, the proposed method does not rely on parametric models nor tries to model age progression. The idea is to use deep learning to jointly learn face features that matching faces share, and a similarity metric on top of these features. This is done coupling two deep convolutional neural networks (DCNN) with shared parameters in a Siamese network [4,10] ended with a contrastive loss function. The discriminative power of the network is further improved including a feature injection layer, which fuses externally computed features with the activations of the deepest layers of the DCNN.

The contributions of this work are summarized as follows:

- A new large-scale Large Age-Gap (LAG) dataset is collected, that includes images in the wild of 1010 international celebrities spanning large age gaps.
- A new DCNN architecture is proposed, including a feature injection layer that fuses external features with the activations of the deepest DCNN layers.
- Extensive experiments are conducted on LAG and show that the proposed DCNN architecture can outperform state-of-the-art methods. In particular, experimental results show that it is possible to increase the performance of a DCNN by injecting external features.

The remaining sections are organized as follows: Section 2 reviews the related works on face recognition, age-invariant face recognition and existing face datasets. Section 3 describes the proposed method, while Section 4 introduces the Large Age-Gap (LAG) dataset. Experiments are presented in Section 5. Finally, Section 6 draws the conclusions and discusses future works.

E-mail address: simone.bianco@disco.unimib.it

http://dx.doi.org/10.1016/j.patrec.2017.03.006
0167-8655/

2. Related works

2.1. Deep feature fusion

The idea of deep feature fusion has been mainly explored in the video categorization task. One of the earliest work is from Simonyan and Zisserman [32] where they proposed a two-stream ConvNet architecture which incorporated a spatial and a temporal network. Perhaps the most similar work is [18] where multimodal video features are combined (e.g. frame-based features computed by a convolutional neural network, trajectory-based motion descriptors and audio descriptors). Wang et al. [38] integrate the advantages of hand-crafted and deep-learned features: they utilize deep architectures to learn multi-scale convolutional feature maps, and introduce the strategies of trajectory-constrained sampling and pooling to encode deep features into effective descriptors. Zha et al.[46] propose a late fusion approach between CNN features (taken at different layers) and Fisher Vectors [29]. The features are fused using an external classifier and thus not in an end-to-end training, excluding the possibility of backward feedbacks on feature extraction. Ng et al. [23] investigated the combination of Long Short Term Memory (LSTM) networks [13] with optical flow. Park et al.[26] propose a multiplicative fusion method for combining multiple CNNs trained on different sources.

2.2. Face recognition

Face recognition has been investigated for a long time in many different works. One of the earliest works is that of Turk and Pentland where they introduced the idea of eigenface [36]. Ahonen et al. [1] explored the use of a texture descriptor, i.e. local binary pattern (LBP), for the face recognition task. Wright et al. [42] cast face recognition problem as one of classifying among multiple linear regression models via sparse signal representation, showing a high degree of robustness against face occlusions. Chen et al. [9] proposed a high dimensional version of LBP (HDLBP) and studied the performance of face feature as a function of dimensionality, showing that high dimensionality is critical to achieve high performance.

Recently there have been many works exploiting deep learning for face recognition. Results obtained by Taigman et al. [35] and by Sun et al. [33,34] using deep convolutional neural networks (DCNNs) reach or even surpass human-level performance on the widely used labeled face in the wild dataset (LFW) [16]. Their performance has been further improved by Schroff et al. [31] introducing FaceNet, that directly learns a mapping from face images to a compact Euclidean space where distances directly correspond to a measure of face similarity. Very recently Parkhi et al. [28] achieved even higher performance by using a very deep network that learns a face embedding using a triplet loss similar in spirit to that of Schroff et al. [31]. Although these methods achieve very high performance on face recognition, they do not work well when in presence of age variation, since this information is not used.

2.3. Age-Invariant face recognition

The largest part of existing works related to age in face image analysis focus on age estimation and simulation. Only recently researchers have started to work on cross age face recognition. Existing works can be grouped into generative and discriminative methods. Among the first group, some of the approaches build 2D [12] or 3D [27] aging models. These methods rely on parametric models and accurate age annotation or estimation, and thus do not work well in unconstrained scenarios. Wu et al. [43] propose a relative craniofacial growth model to model the facial shape change, which is based on the science of craniofacial anthropometry. Their method needs age information to predict the new shapes, limiting its applicability since this information is not always available.

Among the works based on a discriminative approach Li et al. [21] use multi-feature discriminant analysis (MFDA) to process in a unified framework the two local feature spaces generated by the two different local descriptors used, i.e. SIFT and multi-scale LBP. Gong et al. [14] proposed a method separating the HOG local feature descriptor into two latent factors using hidden factor analysis: an identity factor that is age-invariant and an age factor affected by the aging process. Chen et al. [6,7] use a data-driven method that leveraging a large-scale image dataset freely available on the internet as a reference set, encodes the low-level feature of a face image with an age-invariant reference space. Liu et al. [22] propose a generative-discriminative approach based on two modules: the aging pattern synthesis module and the aging face verification module. In the aging pattern synthesis module, an aging-aware denoising auto-encoder is used to synthesize the faces of all the four age groups considered. In the aging face verification module, parallel CNNs are trained based on the synthesized faces and the original faces to predict the verification score.

2.4. Face datasets

Existing face datasets can be divided into two main groups: the former consists of datasets acquired in controlled environments, the latter datasets in unconstrained environments. Most of the older datasets belong to the first group, such as FERET [30], Yale, and CMU PIE. The most popular dataset in uncontrolled environment is the LFW [16], with a total of 13,233 images of 5749 people extracted from news programs. Pubfig [20] has been collected with the aim of providing a larger number of images for each individual, and it contains 58,797 images of 200 identities. The largest dataset available is the CasiaWebFace dataset [44] with a total of 986,912 images of 10,575 people. All the above datasets can be used only for face recognition and verification tasks, since there is almost no age variation. Concerning age estimation and face recognition across age, the most used datasets are FGNet [11] and MORPH [30]. The former is composed of a total of 1002 images of 82 people with age range from 0 to 69 and an age gap up to 45 years. The latter contains 55,134 images of 13,618 people with age range from 16 to 77 and an age gap up to 5 years. Recently the CACD dataset has been collected [6,7] crawling the web suing as query 2000 celebrities names for a total of 163,446 images. For a subset of 200 identities images are manually checked and the noisy ones have been removed. Age ranges from 14 to 62 and age gap is up to 10 years. Very recently the CAFE dataset has been collected [22]. It is the first permitting a study on face verification with large age gaps. It is composed of 4659 images of 901 people and, due to the way images are collected it does not contain precise age information. A summary of the comparison between existing datasets is reported in Table 1.

3. The proposed method

Fig. 1 gives an overview of the proposed method. First face and landmark detection are performed on CASIA-WebFace and Large Age-Gap (LAG) database to localize and align each face to a reference position. Next, a DCNN is trained on the CASIAWebFace [44] for the face recognition task. The knowledge learned by the DCNN is then transferred [25] from the source task of face recognition to the target task of large age gap face verification. Knowledge transfer is performed by fine-tuning [2,45] the DCNN using a contrastive loss in a Siamese architecture in which pre-computed external features are injected in the fully connected layer.