Expert Systems With Applications 108 (2018) 170–182

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Unconstrained face identification using maximum likelihood of distances between deep off-the-shelf features

Andrey V. Savchenko[a,*], Natalya S. Belova[b]

[a] National Research University Higher School of Economics, Laboratory of Algorithms and Technologies for Network Analysis, 36 Rodionova St., Nizhny Novgorod, Russia

[b] National Research University Higher School of Economics, 20 Myasnitskaya St., Moscow, Russia

ARTICLE INFO

Article history:
Received 13 October 2017
Revised 2 April 2018
Accepted 29 April 2018
Available online 9 May 2018

Keywords:
Statistical pattern recognition
Unconstrained face recognition
Maximum likelihood estimation
CNN (Convolution neural network)
Kullback–Leibler divergence
Off-the-shelf deep features

ABSTRACT

The paper deals with unconstrained face recognition task for the small sample size problem based on computation of distances between high-dimensional off-the-shelf features extracted by deep convolution neural network. We present the novel statistical recognition method, which maximizes the likelihood (joint probabilistic density) of the distances to all reference images from the gallery set. This likelihood is estimated with the known asymptotically normal distribution of the Kullback–Leibler discrimination between nonnegative features. Our approach penalizes the individuals if their feature vectors do not behave like the features of observed image in the space of dissimilarities of the gallery images. We provide the experimental study with the LFW (Labeled Faces in the Wild), YTF (YouTube Faces) and IJB-A (IARPA Janus Benchmark A) datasets and the state-of-the-art deep learning-based feature extractors (VGG-Face, VGGFace2, ResFace-101, CenterFace and Light CNN). It is demonstrated, that the proposed approach can be applied with traditional distances in order to increase accuracy in 0.3–5.5% when compared to known methods, especially if the training and testing images are significantly different.

1. Introduction

Face identification is widely applied in various intelligent systems, such as video surveillance, border processing, virtual reality, search for a person in a social network, preventing voter fraud, conference socializing, driver license identification, law enforcement, etc. (Li, Wang, You, Li, & Li, 2013; Zhao, Liu, Liu, Zhong, & Hua, 2015). Though the face recognition has been thoroughly studied for several decades (Prince, 2012; Zhang, Yan, & Lades, 1997), it is still a challenging problem, which appears in many intelligent systems (Savchenko, 2016) due to the variability of individuals presented in images (Learned-Miller, Huang, RoyChowdhury, Li, & Hua, 2016). The latter task is usually solved with modern deep learning techniques (Goodfellow, Bengio, & Courville, 2016), which have recently reached a certain level of maturity. The most promising results are achieved with deep convolutional neural networks (CNN) (LeCun, Bengio, & Hinton, 2015). The CNNs have recently made it possible to achieve a near human-level performance in various computer vision contests including facial verification (Schroff, Kalenichenko, & Philbin, 2015; Taigman, Yang, Ranzato, & Wolf, 2014), especially when the training set contains plenty of reference images.

The situation becomes more complicated if the gallery set contains a small number of reference instances per each class. This small sample size problem (Chen, Liao, Ko, Lin, & Yu, 2000; Raudys, Jain et al., 1991) is especially crucial in face recognition, when sometimes only one reference image of each person is available (Tan, Chen, Zhou, & Zhang, 2006; Zhao et al., 2015). Such tasks with single reference photo per person of interest are typical for still-to-video face recognition (Dewan, Granger, Marcialis, Sabourin, & Roli, 2016; Parchami, Bashbaghi, & Granger, 2017; Savchenko, Belova, & Savchenko, 2018). In such case the transfer learning or domain adaptation methods can be applied (Cao, Wipf, Wen, Duan, & Sun, 2013; Goodfellow et al., 2016). In these methods the CNN is

Abbreviations: AUC, Area Under Curve; CNN, Convolution Neural Network; HOG, Histogram of oriented gradients; IJB-A, IARPA Janus Benchmark A; KL, Kullback-Leibler divergence; LBP, Local Binary Patterns; LDA, Linear Discriminant Analysis; LFW, Labeled Faces in the Wild; MAP, Maximum A Posteriori; MLP, Multi-Layered Perceptron; NN, Nearest Neighbor; PCA, Principal Component Analysis; SVM, Support Vector Machine; YTF, YouTube Faces.

* Corresponding author.

E-mail addresses: avsavchenko@hse.ru (A.V. Savchenko), nbelova@hse.ru (N.S. Belova).

URL: http://www.hse.ru/en/staff/avsavchenko (A.V. Savchenko), http://www.hse.ru/en/staff/belova (N.S. Belova)

https://doi.org/10.1016/j.eswa.2018.04.039
0957-4174/

used to extract facial features, which can be processed using the known classifiers. It is known that most of existing face recognition algorithms suitable for application with small training samples are typically based on the similarity comparison between the image features in the training set and an observed image (Guo & Zhang, 2017; Taigman et al., 2014), which is essentially the nearest neighbor (NN) method.

In order to increase the accuracy of the NN classifier, in this paper we propose to exploit the idea of several approximate NN methods (Micó, Oncina, & Vidal, 1994; Savchenko, 2012): if there exists a *reliable* decision $\mathbf{x}^*$, for which the distance to the input feature vector $\mathbf{x}$ is low ($\rho(\mathbf{x}, \mathbf{x}^*) \ll 1$), then $\rho(\mathbf{x}, \mathbf{x}_r) \approx \rho(\mathbf{x}^*, \mathbf{x}_r)$ with high probability for an arbitrary rth reference point. This assumption is known to be asymptotically correct for the KL divergence and probabilistic model of each class (Kullback, 1997). Following this idea, we introduce the novel face recognition method, which is based on the probabilistic interpretation of recognition task (Savchenko, 2016). At first, several nearest reference images (instances from the gallery set) are obtained for an observed facial image. Next, computed distances to all instances are used to weight the recognition results based on the estimation of their reliability. The more is the likelihood of the computed vector of distances for particular individual, the more is the weight corresponding to this subject. The likelihoods (joint probabilistic densities) are computed using the idea of the maximum-likelihood approximate NN algorithm (Savchenko, 2017a; 2017c) by assuming that the Kullback-Leibler (KL) divergence (Kullback, 1997) is used to compare distances between deep high-dimensional off-the-shelf features. However, we demonstrate that our approach can be successfully applied with traditional Euclidean distance.

The rest of the paper is organized as follows. In Section 2 we briefly overview existing image recognition methods suitable for small training samples. In Section 3 we present a simple statistical formulation of the face recognition task (Shakhnarovich, Fisher, & Darrell, 2002) using the KL minimum discrimination principle and introduce the novel approach, in which the maximum a-posteriori (MAP) rule is regularized using the computation of the joint probability densities of distances based on the asymptotic properties of the KL divergence. Section 4 presents the experimental results in recognition of images from either IJB-A (IARPA Janus Benchmark A) (Klare et al., 2015) or YTF (YouTube Faces) datasets (Wolf, Hassner, & Maoz, 2011) with the still images from the LFW (Labeled Faces in the Wild) dataset (Learned-Miller et al., 2016). Finally, concluding comments are given in Section 5.

2. Literature survey

One possible research direction in face recognition with small training samples is the usage of traditional computer vision methods (Szeliski, 2010), i.e., classification of either local features, e.g., SIFT, or a global descriptor, e.g., HOG (histogram of oriented gradients) (Dalal & Triggs, 2005). These methods usually partition each face image into several patches/blocks, and then perform feature extraction on them. Hence, they are sometimes called patch/block based approach (Zhu, Yang, Zhang, & Lee, 2014). After that, the NN methods are used with an appropriate similarity measure (Savchenko, 2016) between such features extracted from the observed image and all reference instances. A variant of such approach was proposed by Zhang, Yang, and Qian (2012), who performed PCA (principal component analysis) for feature extraction, chose k nearest neighbors of a given testing sample globally, and then used these neighbors to represent the testing sample via ridge regression. Similar ideas were introduced in the discriminative multi-manifold analysis method Lu, Tan, and Wang (2013), in which discriminative features are learned by maximizing the manifold margins of different persons. In this approach the facial images are segmented into disjoint sub-images, which form an image set for each sample per person. Another extension of these methods is the binary classification of the distances between images for intra/extra-personal classification (Zhang, Huang, Li, Wang, & Wu, 2004). It is suitable for the case when several reference facial images are available for each individual. Given the impossibility to train complex classifiers, the vector of distances between corresponding parts is assigned to one of two classes, depending on whether these distances are calculated between objects of the same or different classes. An observed image is segmented into a grid of blocks, then the distance vector is estimated for each instance from the training set. This distance vector is classified by a trained AdaBoost classifier, and the decision is made in favor of the class corresponding to the model with the highest confidence. In fact, it is the same NN rule, but the similarity measure is the AdaBoost's confidence.

The second group of methods includes enlarging of the training set by modifying images from the gallery set. After that statistical subspace-based approach can be implemented (Prince, 2012). These methods were most widely studied in literature devoted to face recognition from a single image per person (Tan et al., 2006). For instance, Zhao et al. (2015) proposed to automatically detect important local feature points by template matching and use a statistical model to learn the discriminative feature in the hidden space for each individual. Adaptive appearance model in still-to-video face recognition was considered by Dewan et al. (2016). Another well known technique is the singular value decomposition perturbations (Zhang, Chen, & Zhou, 2005) for the linear discriminant analysis (LDA), which enriches the eigenspace learned by the single training image. A linear generative model that creates a one-to-many mapping from an idealized identity space to the observed data space was introduced in Prince, Elder, Warrell, and Felisberti (2008) in order to deal with large pose variations.

One variant of such approach expands the training set using synthetic face generation (Mokhayeri, Granger, & Bilodeau, 2015), in which multiple virtual face images are generated from each single reference (Dewan et al., 2016). Image synthesis aims to estimate the intra-class face variations by simulating extra samples for each subject and, hence, increasing the number of samples. For instance, Li et al. (2013) proposed to enlarge the training set based on inter-class relationship and extended LDA in order to extract features from the enlarged training set. Zeng et al. (2017) used similar ideas to introduce intra-class facial variations, which are assumed to be shared across different persons.

Another variant is the multiple face representations (Bashbaghi, Granger, Sabourin, & Bilodeau, 2014), in which different discriminant features are extracted from a reference image to enhance the face models (Dewan et al., 2016). Leonidou, Tsapatsoulis, and Kollias (1999) proposed to use multiple representations of the gallery images, including variation in scaling, content and luminance. Multiple representations in Zhang, Hu, Xiang, and Zhao (2017) were created by cropping the original image into a number of non-overlapping blocks, applying of certain operations to each block, and merging all the modified blocks.

Unfortunately, most of these methods were successfully applied only in *constrained* face recognition task (Phillips et al., 2003). However, modern intelligent systems require identification of faces observed in *unconstrained* conditions, i.e., various illumination, pose, presence of noise, etc. (Best-Rowden, Han, Otto, Klare, & Jain, 2014; Learned-Miller et al., 2016). Hence, nowadays enlarging the training set (Dewan et al., 2016) (or generic learning) becomes all the more popular. In these methods the face intra-class variation information is gathered from an external generic training set.