Neurocomputing 187 (2016) 11–18

Contents lists available at ScienceDirect

Neurocomputing

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

Face identification with second-order pooling in single-layer networks

Fumin Shen [a], Yang Yang [a,*], Xiang Zhou [a], Xianglong Liu [b], Jie Shao [a]

[a] School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, PR China
[b] State Key Lab of Software Development Environment, Beihang University, Beijing 100191, PR China

ARTICLE INFO

Article history:
Received 30 April 2015
Received in revised form
10 June 2015
Accepted 13 July 2015
Available online 12 December 2015

Keywords:
Face recognition
Image classification
Second-order pooling

ABSTRACT

Automatic face recognition has received significant performance improvement by developing specialized facial image representations. On the other hand, spatial pyramid pooling of features encoded by an over-complete dictionary has been the key component of many state-of-the-art generic objective classification systems. Inspired by its success, in this work we develop a new face image representation method under the framework of single-layer networks, where the key component is the second-order pooling layer. The proposed method differs from the previous methods in that, we encode the densely extracted local patches by a small-size dictionary; and the facial image signatures are obtained by pooling the second-order statistics of the encoded features. We show the importance of the encoding procedure, which is bypassed by the original second-order pooling method to avoid the high computational cost. Equipped with a simple linear classifier, the proposed method outperforms the state-of-the-art face identification performance by large margins. For example, on the LFW databases, the proposed method performs better than the previous best by around 13% accuracy.

1. Introduction

Face identification aims to find the subject in the gallery most similar to the probe face image. Despite decades of research effort, it is still an active topic in computer vision due to both its wide applications and technical challenges. The challenges are typically caused by various intra-class variations (e.g., face expressions, poses, ages, and image contaminations), or lack of sufficient training data [1]. One of the key problems is to generate a robust and discriminant representation for facial images. Extensive research effort in the literature has been devoted to projecting the face vectors to a low-dimensional subspace, e.g., as in the method of eigenfaces [2], Fisher-faces [3], and Laplacian faces [4]. However, these holistic feature based methods often are incapable to cope with the aforementioned problems well.

Recently, sparse representation based face classification has achieved promising results [5,6]. Different from previous methods, these methods compute the representation of the probe image to achieve the minimum representation error in terms of a set of training samples or a dictionary learned from training images. Many algorithms have been developed in this category, which achieve state-of-the-art performance on face recognition with image corruptions [5,7], face disguises [6,8–10] and small-size training data [11–13].

To improve the face recognition performance, many local feature based methods have been proposed, which tend to show superior results over those based holistic features. Typical methods in this group include histograms of local binary patterns (LBP) [14], histograms of various Gabor features [15–17] and their fusions [18]. These local feature based methods have been proven to be more robust to mis-alignment and occlusions.

On the other hand, the local feature based image representation—bag-of-visual-words (BOV)—has been shown state-of-the-art recognition accuracy [19]. The typical pipeline of BOV is low-level local feature extraction (raw pixels, SIFT, etc.), feature quantization or encoding against a pre-trained dictionary, and descriptor generation by spatially pooling the encoded local features. This pipeline has been shown to achieve the state-of-the-art performance in generic image classification [20–22]. Despite the success of the BOV model in image classification, it has been rarely applied to face recognition.

The dimensionality of the learned image descriptor through the BOV pipeline is mainly determined by the size of trained dictionary (dimension of the encoded local features) and the pooling pyramid grids. It has been shown that a large dictionary size is critical to achieve a high accuracy for generic image classification [23]. In the meantime, pooling features over a spatial region leads to more compact representations, and also helps to make the representation

* Corresponding author. Tel.: +86 28 61831783.
E-mail addresses: fumin.shen@gmail.com (F. Shen), dlyangyang@gmail.com (Y. Yang), Johnfly2000@gmail.com (X. Zhou), xlliu@nlsde.buaa.edu.cn (X. Liu), shenht@itee.uq.edu.au (J. Shao).

http://dx.doi.org/10.1016/j.neucom.2015.07.133
0925-2312/

invariant to image transformation and more robust [24]. The spatial pyramid pooling model [21] has made a remarkable success, for example, in conjunction with sparse coding techniques [22].

Average pooling and max-pooling are the two most popular pooling methods. The latter method usually leads to superior performance to the former one [22,25]. Most previous methods compute fist-order statistics in the pooling stage. In contrast, recently the average and max-pooling methods that incorporate the second-order information of local features have been proposed in [26] for image segmentation. Without an encoding stage, the second-order pooling strategy of [26] is directly applied to the raw SIFT descriptors. A similar idea has previously been proposed in [27,28] for describing region covariance features in object classification and detection problems.

Inspired by both the BOV model and the second-order pooling method of [26–28], here we propose a new method for facial image representation learning under the framework of single-layer networks, where the key component is the second-order pooling layer. First, local raw patches are densely extracted from the face images. The local patches are then encoded by a *small-size* dictionary, e.g., trained by K-means. The encoded features are finally pooled by employing the second-order statistics over a multi-level pyramid. The efficacy of the learned facial features is verified by the state-of-the-art performance on several public benchmark databases.

The BOV model has been less frequently studied on face recognition problems. The Fisher vectors on densely sampled SIFT features were adopted in [29] for face verification problems. In contrast, we focus on raw intensity features. Also note that no pooling is applied in this method. Combination of sparse coding and spatial max-pooling has been used in [30] for face recognition. However, only the first-order statistics are computed in the pooling stage.

Our contributions mainly include:

1. We propose a new facial representation learning method under the framework of single-layer networks, where the key component is the second-order pooling layer. To our knowledge, this is the first face feature extraction method using the second-order pooling technique.
2. Different from the standard BOV methods which usually involve an over-complete dictionary, we show that a very small number of dictionary basis are sufficient for face identification problems, in conjunction with the second-order pooling. In contrast to the method in [26], which does not apply encoding, we show that feature encoding is critically important for face identification and always improves the recognition accuracy.
3. Coupling with a simple linear classifier, the proposed method outperforms those state-of-the-art by large margins on several benchmark databases, including AR, FERET and LFW. In particular, the proposed method achieves perfect recognitions (100% accuracies) on the 'Fb' subset of the FERET dataset and the 'sunglasses' and 'scarves' subsets of the AR dataset. Our method obtains a higher accuracy than the best previous result by around 13% on LFW.

2. The proposed method

In this section we present the details of the proposed method. We focus on extracting a discriminant representation of an image based on its raw intensity feature other than other specific designed ones like SIFT. We summarize our algorithm in Fig. 1. The frequently used notations are defined in Table 1.

2.1. Dense local patch extraction

Without loss of generality, suppose that a facial image is of $d' \times d'$ pixels. As the first step, we extract overlapped local patches of size $r \times r$ pixels with a step of s pixels. Set $l = \lfloor \frac{d'-r}{s} + 1 \rfloor$, then each image is divided into $l \times l$ patches. Let each local patch be a row vector $\mathbf{x}$.

It has been shown that dense feature extraction and the pre-processing step are critical for achieving better performance [23]. In practice, we extract local patches of 6×6 pixels with a stride of 1 pixel. We then perform normalization on $\mathbf{x}$ as $\hat{x}_i = (x_i - m)/v$, where x_i is the ith element of $\mathbf{x}$, and m and v are the mean and standard deviation of elements of $\mathbf{x}$. This operation contributes to local brightness and contrast normalization as in [23].

2.2. Unsupervised dictionary training

The goal of this step is to learn from the training data a representative "dictionary", i.e., a set of basis, $\mathbf{D} = \mathbf{d}_1, \mathbf{d}_i, \ldots, \mathbf{d}_m \in \mathbf{R}^{d \times m}$. Here $\mathbf{d}_i$ is the i_{th} of the m dictionary basis (atoms), and d is the input patch dimension. With the obtained dictionary, an image is typically represented by a linear combination of the dictionary basis ("word" in the bag-of-visual-words model).

A great deal of unsupervised dictionary learning methods has been developed, for example the K-means clustering, sparse coding, K-SVD [31]. Dictionary can also be trained with the help of category information, such as the supervised sparse coding method [30]. We adopt the K-means algorithm, since it is simple and effective. The dictionary size is a more important factor compared to the dictionary training algorithm.

With a first-order pooling method, it has been shown that the image classification accuracy is consistently improved as the dictionary size increases [23]. However, this is not necessarily true when a second-order pooling technique is applied. Perhaps surprisingly, with second-order pooling, a very small number of dictionary basis are sufficient to obtain high recognition accuracies. We will analysis this in Section 3.1.

As a common pre-processing step in deep learning methods, whitening has been shown to yield sharply localized filters when dictionary are trained by clustering on raw data [23]. We apply the ZCA whitening on each patch [32] before the dictionary learning algorithm are applied.

2.3. Feature encoding

With the leaned dictionary, the pre-processed local patches are then fed into the feature encoder to generate a set of mid-level features. Popular choices of encoding algorithms include the sparse coding [22] and locality-constrained linear coding (LLC

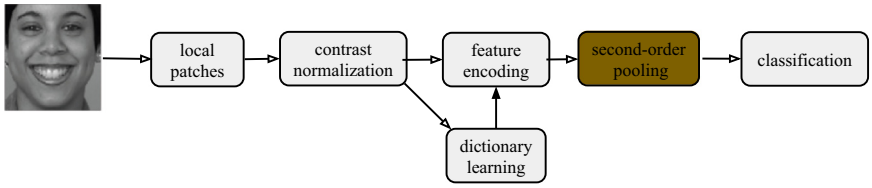

Fig. 1. The pipeline of the proposed algorithm.