Neurocomputing 147 (2015) 403–411

Contents lists available at ScienceDirect

Neurocomputing

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

Sparsely encoded local descriptor for face verification

Zhen Cui [a,b], Shiguang Shan [a,*], Ruiping Wang [a], Lei Zhang [c], Xilin Chen [a]

[a] *Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China*
[b] *School of Computer Science and Technology, Huaqiao University, Xiamen 361021, China*
[c] *Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China*

ARTICLE INFO

Article history:
Received 14 November 2013
Received in revised form
7 April 2014
Accepted 22 June 2014
Communicated by Qingshan Liu
Available online 10 July 2014

Keywords:
Local descriptor
Sparse coding
Non-negativity
Face verification
Labeled faces in the wild

ABSTRACT

A novel Sparsely Encoded Local Descriptor (SELD) is proposed for face verification. Different from traditional hard or soft quantization methods, we exploit linear regression (LR) model with sparsity and non-negativity constraints to extract more discriminative features (i.e. sparse codes) from local image patches sampled pixel-wisely. Sum-pooling is then imposed to integrate all the sparse codes within each block partitioned from the whole face image. Whitened Principal Component Analysis (WPCA) is finally used to suppress noises and reduce the dimensionality of the pooled features, which thus results in the so-called SELD. To validate the proposed method, comprehensive experiments are conducted on face verification task to compare SELD with the existing related methods in terms of three variable component modules: K-means or K-SVD for dictionary learning, hard/soft assignment or regression model for encoding, as well as sum-pooling or max-pooling for pooling. Experimental results show that our method achieves a competitive accuracy compared with the state-of-the-art methods on the challenging Labeled Faces in the Wild (LFW) database.

1. Introduction

Face recognition has attracted significant attention due to its wide potential applications in public security, law enforcement, etc. Numerous methods or techniques have been developed as surveyed in [1], and considerable progresses have been achieved in the past decades. Currently, state-of-the-art face recognition systems have been able to work well under well-controlled conditions with cooperative users. However, as discovered by LFW evaluation [2], face recognition under uncontrolled environment still remains a great challenge due to complex variations in pose, illumination, expression, aging, etc. To well address this problem, how to discriminatively represent face images plays a key role in the task of unconstrained face recognition.

In the past decade, local descriptors, modeling micro-patterns in images, have formed a blowout in face recognition area [3–9], due to their robustness to identity-irrelevant extrinsic variations. These methods usually fall into two categories: hand-crafted and auto-learned descriptors, which are briefly introduced in what follows.

Many manually designed local patterns have been developed for face recognition. For example, by combining the signs of the differences of central pixel intensity from those of its neighboring pixels, Local Binary Patterns (LBP) [6] implicitly encodes the micro-patterns of the input image such as flat areas, spots, edges, and corners. Because of its invariance to monotonic photometric changes, LBP is robust to lighting variation to some extent. After that, many variants of LBP were proposed. For instance, Zhao and Pietikainen extended LBP to the spatial-temporal domain [10]. In order to make LBP more robust to random and quantization noise in near-uniform face regions, Local Ternary Patterns (LTP) [11] were proposed. By combining Gabor filtering [12] with LBP, Local Gabor Binary Pattern (LGBP) [8] was proposed to endow LBP with capacity of encoding micro-patterns of multi-scale and multi-orientation. Later on, histogram of Gabor phase patterns [7] was further proposed to exploit the Gabor phase information. In addition, some local descriptors widely used in general object classification, such as Histogram of Oriented Gradients (HOG) [13] or Scale Invariant Feature Transform (SIFT) [9], were introduced into face recognition. In spite of its popularity, manually designing local patterns are non-trivial because it has to balance skillfully discriminative power and robustness against data variance.

In contrast to the above hand-crafted approaches, auto-learning based methods typically pursue some codewords (representative local visual primitives) from a large amount of low-level features (e.g. SIFT). Then, given an input image, its low-level features are encoded with these codewords by utilizing hard/soft quantization, followed by pooling operation to form mid-level

[*] Corresponding author. Tel.: +86 10 62600558.
E-mail address: sgshan@ict.ac.cn (S. Shan).

http://dx.doi.org/10.1016/j.neucom.2014.06.044
0925-2312/© 2014 Elsevier B.V. All rights reserved.

features. By learning the codewords directly on image patches with K-means clustering algorithm, Meng et al. [14] proposed Local Visual Primitives (LVP), which finally represented one face image by concatenating block-based histograms of the learned patterns for face recognition. Ahonen and Pietikainen [15] also tried K-means clustering to build local filter response codebook. Cao et al. [5] argued that quantized codes with K-means usually had uneven distribution and the encoded histogram would be less informative and less compact. To address the problem, they substituted random-projection tree for K-means clustering. In addition, hard quantization may lead to losing a lot of useful information especially subtle textural features in face images, since only one nearest atom is chosen as the agent for one input raw feature. In contrast, soft quantization based methods [16,17] encode the input features with multiple codewords so as to make the representation more accurate. For instance, van Gemert et al. [17] proposed to use Gaussian kernel to deal with visual word ambiguity for object classification.

Another recent progress in face recognition is sparse representation based methods [18–25]. In [18], Wright et al. sparsely encoded one face image by using the training set as the codebook and then sought for the subject whose samples result in the smallest reconstruction error by using their corresponding sparse coefficients. In the case of multiple well-aligned samples per person, they reported impressive results, especially for partially occluded faces. Further, some researchers tried to learn a robust codebook, such as the discriminative codebook [19] and the compact Gabor codebook [20]. Besides, Cui et al [25] apply sparse representation into video-based face recognition. However, these methods mostly focus on holistic representation, and thus are fragile to local appearance variations. Another limitation of these methods is that they only work for the scenario where each subject has multiple enrolled face images, i.e., they cannot be applied to face verification and face identification with single sample per person. To address these problems, more recently, face region descriptor (FRD) [4] is proposed to address still and video images with a similar framework.

Inspired by the above works, in this paper we propose a local descriptor via texton-learning with sparsity constraints. Specifically, our method first learns visual codewords locally on image patches with sparsity constraints. Then, non-negative sparse regression against the visual codewords is exploited to project each pixel-wise raw image patches into more discriminative sparse codes, which is quite different from the existing hard assignment methods [5,14,26] and soft assignment methods [16,17]. In the next step, sum-pooling is exploited to integrate the sparse codes within each image block, and at the same time endow the generated mid-level features more robustness to misalignment. Finally, Whitened Principal Component Analysis (WPCA) [27] is used to further reduce the dimensionality and suppress the noise of the pooled features, eventually resulting in our Sparsely Encoded Local Descriptor (SELD).

As an extension of our previous work [3], we further improve the conference work mainly on three aspects: (1) multiple block-partitioning modes on face images are used to retain more facial configuration information; (2) Distance Metric Learning (DML) is combined with SELD to utilize supervised information; and (3) extensive cross-validation experiments on the three component modules: dictionary learning, encoding and spatial pooling. As a whole, our contributions mainly lie in three folds: (1) propose an auto-learning face descriptor for face verification; (2) conduct extensive cross-validation experiments to validate the role of each module; and (3) achieve a competitive performance on the LFW dataset under its restrict protocol.

As our experiments are mainly conducted on the LFW dataset, here we briefly review the related state-of-the-art methods on it.[1] To achieve competitive, latest methods usually fuse multiple hand-crafted features, such as Gabor, LBP, TPLBP as in [28,29], or learn more efficient features by using Bag-of-Word (BoW) framework [5,4], or turn to deep learning [30]. To measure the similarity of features, distance metric learning methods are popular to enhance discriminability, as in [28,29,31]. Please note that, this paper only focuses on the restrict protocol of LFW, so we do not introduce methods depending on additional external dataset.

The remaining part of this paper is organized as follows. Section 2 presents the details of the proposed SELD, including the detailed description on the whole pipeline and three component modules. Section 3 discusses the fusion of multiple different partition modes, and the combination of Distance Metric Learning and SELD. Results and analysis of comprehensive experiments on LFW are presented in Section 4, followed by discussion and conclusion in the last section.

2. Sparsely encoded local descriptor

In this section, we first give an overview of the proposed SELD. Then we describe its three key components in detail: learning dictionary, encoding image patches and pooling codes. Finally, a discussion of WPCA is given.

2.1. Overview

SELD is essentially an enhanced texton-based method. It aims to learn robust local descriptors from face images. The overall schema of the proposed method is illustrated in Fig. 1. As shown in the figure, before extracting the SELD features, we first roughly align face images by fixing the eyes at the same position for all the face images, and then filter them with a Difference of Gaussian (DoG) so as to remove both high-frequency noises and low-frequency illumination variations. To preserve more texton information, we pixel-wisely sample raw image patches from the images by a pre-defined template. Each raw patch is vectorized into an intensity vector to form the original feature, which is then sparsely encoded into a higher level feature vector using an offline-learned over-complete dictionary (detailed in Section 2.2).

With the above sparse codes computed, the face image is spatially partitioned into a number of cells (or blocks), and the code vectors of all pixels within each cell are sum-pooled together to form a single descriptor for this cell. Finally, in order to suppress the noises of the pooled descriptors, we exploit whitened PCA to project them into a low-dimensional space, which finally results in our SELD.

In the above schema, if different cell-partitioning manners are applied, multiple SELDs can be generated for each face image. Given two face images, we may compute the similarity of their corresponding SELDs in the same cell-partition. The similarity scores from multiple partitioning manners can be either accumulated together followed by the simple Nearest Neighbor (NN) classifier for face identification, or fed into an SVM classifier for face verification.

2.2. Dictionary learning with K-SVD

In theory, sparse representation assumes a signal can be recovered from a very limited number of atoms contained in an over-complete dictionary. Thus, how to construct a good dictionary that can well support the sparse recovery is very crucial for subsequent representation and classification. To produce the

[1] http://vis-www.cs.umass.edu/lfw/results.html.