Neurocomputing 129 (2014) 516–527

Contents lists available at ScienceDirect

Neurocomputing

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

Histogram of visual words based on locally adaptive regression kernels descriptors for image feature extraction

Jianjun Qian [a], Jian Yang [a,b,*], Nan Zhang [c,a], Zhangjing Yang [a]

[a] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, PR China
[b] Computation and Neural Systems, California Institute of Technology, Pasadena, CA 91125, USA
[c] Wuxi Institute of Technology, Wuxi 214000, PR China

ARTICLE INFO

Article history:
Received 29 August 2012
Received in revised form
24 June 2013
Accepted 3 September 2013
Communicated by Xiaofei He
Available online 19 October 2013

Keywords:
Locally adaptive regression kernels descriptors
Bag-of-visual-words
Feature extraction
Sparse representation

ABSTRACT

Image feature extraction is one of the most important problems for image recognition system. We tackle this by combing the locally adaptive regression kernel descriptors (LARK), bag-of-visual-words and sparse representation. Specifically, this paper makes two main contributions: (1) we introduce a novel method called histogram of visual words based on locally adaptive regression kernels descriptors (HWLD) for image feature extraction. LARK is used to describe the image local information and build the visual vocabulary. Each pixel of an image is assigned to the visual words and gets the corresponding weights. Image feature vector is obtained by subdividing the image and computing the accumulative weight histograms of visual words in these sub-blocks. (2) The K nearest neighbor based sparse representation (KNN-SR) is presented for assigning the visual words. Compared with nearest neighbors based method, KNN-SR has stronger discriminant power to identify different patches in the image. Experimental results on the AR face image set, the CMU-PIE face image set, the ETH80 object image set and the Nister image set demonstrate that our method is more effective than some state-of-the-art feature extraction methods.

1. Introduction

Image Feature extraction is one of the well-studied problems in pattern recognition and computer vision. In the past decades, numerous useful subspace based methods for image feature extraction have been developed [1]. Especially, principal component analysis (PCA) and liner discriminant analysis (LDA) are two traditional methods for feature extraction. Both of them have been widely used in pattern recognition and computer vision. They are known as the famous Eigenfaces method and Fisherfaces method that have become the most popular technologies for face recognition, respectively [2]. And many extensions of LDA have shown good performance in various applications [3–5]. In addition, many manifold learning algorithms for discovering intrinsic low-dimensional embedding of data have been proposed. The manifold learning methods include isometric feature mapping (ISOMAP) [6], Local linear embedding (LLE) [7], laplacian Eigenmap [8,9], Locality Preserving Projections (LPP) [9] and Unsupervised Discriminant Projection (UDP) [10]. Some experiments have shown that these methods can find perceptually meaningful embeddings for face images. They also yielded impressive results on the other artificial and real-world datasets. For image descriptors, Scale invariant feature transform (SIFT) is one of the most famous and popular image descriptors in the scenario of image matching and scene recognition [46]. SIFT is actually a 3-D histogram of gradient locations and orientations. The contribution of the location and orientation bins is determined by the gradient magnitude. In [47], Bosch introduced the dense SIFT (DSIFT) in conjunction with Hybrid Generative/Discriminative approach for scene classification. Local binary pattern (LBP) operator, which captures spatial structure of local image texture, is one of the best texture descriptors [11]. And it not only has been widely used in texture image classification, but also successfully applied in face recognition and verification [12,13]. Apart from these methods, there are some other approaches that have proven to be effective in many real-world applications [14–16].

Recently, Seo et al. presented a novel image descriptor locally adaptive regression kernels (LARK) [17], which capture the local structure information effectively. The method is inspired by their early work [18]. They also compared LARK with the sate-of-the-art image local descriptors as evaluated in the paper [19], and experimental results verified that their LARK descriptors give more discriminative power than others. On the basis of LARK descriptors, they not only designed a unified framework for static and space–time saliency detection [20], but also proposed a novel face representation approach for verification [21]. In addition, locally adaptive regression kernels were extended to space–time locally adaptive

* Corresponding author. Tel.: +86 25 8431 7297.
E-mail addresses: qjjtx@126.com (J. Qian), csjyang@njust.edu.cn (J. Yang), 554961959@qq.com (N. Zhang), yzzjjj@126.com (Z. Yang).

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.09.007

regression kernels which robustly preserve underlying space–time data structure for action recognition from one example [22]. It is concluded that there are three significant characteristics of LARK: (1) It is robust even in the presence of noise and blur; (2) normalized LARK yields invariance to illumination changes; and (3) it may be able to provide more stable local structure information of the data.

As an intermediate representation method, the bag-of-visual-words has been extensively used in visual recognition and scene classification [23–26]. First of all, descriptors computed from the training images are quantized into visual words with the *k*-means clustering algorithm. Each descriptor of an image is assigned to a visual word by using the nearest neighbor based method. We argue that it is not an optimal choice to directly assign the descriptor to its nearest neighbor for visual recognition. To address this problem, Jiang et al. proposed a straightforward soft-weighting method to weight the significance of the visual words [48]. Jiang and Chong-Wah Ngo also introduced a novel soft-weighting scheme and took into account two key points: the assignment of descriptor-to-word is one-to-many and the weight of assigned word is determined by their linguistic relationship [49]. M. Kogler et al. employed the fuzzy cluster technology for visual words assignment [50]. Jan C. Gemert et al. presented a soft assignment model, including two types of ambiguity between codewords: codeword uncertainty and codeword plausibility, for image classification [51]. Additionally, there are many soft-weighting schemes that have been proposed to weigh the significance of visual words. All these methods compute the weight term according to the distance metric between the descriptor and the visual word. However, sparse representation theory provides a new idea for assigning visual words from the linear representation point of view since it has been successfully applied to many vision tasks [27–30]. Especially, Wright et al. presented a sparse representation based classification (SRC) method and successfully applied it to real-world face recognition problems [28].

LARK (or SIFT) only take advantage of the within-image information to capture the local structural features but ignores the between-sample information. Actually, the between-sample information can be obtained by virtue of the bag-of-visual-words method. We use various LARK descriptors come from different sample images to build the visual vocabulary. Subsequently, each local descriptor of an image is represented by one visual word, and histogram is employed to compute the visual words frequency distribution. In this way, the obtained image feature not only contains the within-image information, but also includes the between-sample information.

Motivated by these ideas, we propose a novel image feature extraction method named histogram of visual words based on locally adaptive regression kernels descriptors (HWLD) that combines the strengths of the LARK, bag-of-visual-words and sparse representation. HWLD could be viewed as an extension of LARK. Dense LARK descriptors are computed and normalized first. In order to enhance the discriminant power and reduce computation complexity, PCA is thus used to reduce the dimension of LARK descriptors based on the training set. We performed *k*-means clustering of the random LARK descriptors from the training images to build a visual vocabulary. It is named LARK vocabulary in this paper. Subsequently, each pixel of an image is represented by indexes of its *K* nearest visual words and corresponding weights, which are obtained via *K* nearest neighbor based sparse representation (KNN-SR). One image is divided into $16(4\times 4)$ sub-blocks and then represented by accumulative weight histograms of visual words occurrences in these sub-blocks. The overview of the proposed method is illustrated in Fig. 1.

Specifically, this paper uses HWLD for image feature extraction, in conjunction with nearest neighbor classifier. Our method gives outstanding performance compared with some state-of-the-art approaches. Meanwhile, we also use LARK descriptors to combine with PCA, LDA and LPP for specific recognition tasks. They outperform the methods that use PCA (LDA and LPP) directly for recognition. However, HWLD is superior to all of them as well. Hence, HWLD not only preserves advantages of original LARK descriptors, but also shows its advantages of image feature representation for face recognition and object classification. Moreover, it obtains better result with small training samples under the same condition.

The remainder of this paper is organized as follows: Section 2 describes the related works; Section 3 introduces the detail of HWLD; Section 4, the proposed method is assessed using AR database, CMU-PIE database, ETH80 database as well as Nister database; finally, some conclusions are offered in Section 5.

2. Related works

2.1. Outline of locally adaptive regression kernels descriptors

LARK captures the local structure information of images by analyzing the radiometric differences based on estimated gradient, which also determine the shape and size of canonical kernel. The locally adaptive regression kernels is modeled as follows [17]:

$$K(\boldsymbol{x}_l-\boldsymbol{x}_i)=\frac{\sqrt{\det(\boldsymbol{C}_l)}}{2\pi h^2}\exp\left\{-\frac{(\boldsymbol{x}_l-\boldsymbol{x}_i)^T\boldsymbol{C}_l(\boldsymbol{x}_l-\boldsymbol{x}_i)}{2h^2}\right\} \quad (1)$$

where h is a global smooth parameter, $l\in(1\ldots p)$, and p is the

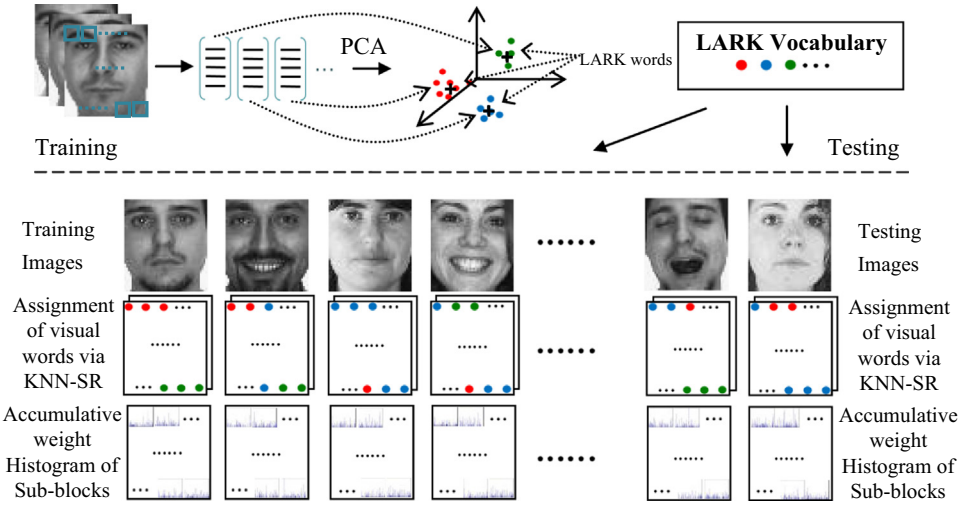

Fig. 1. The overview of HWLD.