Neurocomputing 149 (2015) 1048–1057

Contents lists available at ScienceDirect

Neurocomputing

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

Regularized discriminant embedding for visual descriptor learning

Kye-Hyeon Kim [a], Rui Cai [b], Lei Zhang [b], Seungjin Choi [a,*]

[a] Department of Computer Science and Engineering, Pohang University of Science and Technology (POSTECH), 77 Cheongam-ro, Nam-gu, Pohang 790-784, Republic of Korea

[b] Microsoft Research Asia, No. 5 Dan Ling Street, Haidian District, Beijing 100080, China

ARTICLE INFO

Article history:
Received 1 August 2013
Received in revised form
26 May 2014
Accepted 21 July 2014
Communicated by Deng Cai
Available online 4 August 2014

Keywords:
Discriminant analysis
Kernel method
Local descriptor
Object retrieval

ABSTRACT

Visual descriptor learning seeks a projection to embed local descriptors (e.g., SIFT descriptors) into a new Euclidean space where pairs of matching descriptors (positive pairs) are better separated from pairs of non-matching descriptors (negative pairs). The original descriptors often confuse the positive pairs with the negative pairs, since local points labeled "non-matching" yield descriptors close together (irrelevant-near) or local points labeled "matching" yield descriptors far apart (relevant-far). This is because images differ in terms of viewpoint, resolution, noise, and illumination. In this paper, we formulate an embedding as a *regularized discriminant analysis*, which emphasizes relevant-far pairs and irrelevant-near pairs to better separate negative pairs from positive pairs. We then extend our method to nonlinear mapping by employing recent work on explicit kernel mapping. Experiments on object retrieval for landmark buildings in Oxford and Paris demonstrate the high performance of our method, compared to existing methods.

1. Introduction

Comparing images by matching their local interest points[1] [1–3] is a fundamental preliminary task in many multimedia applications and computer vision problems, including object recognition [4,5], near-duplicate media detection [6,7], image retrieval [8–10], and scene alignment [11]. Local descriptors are extracted from a small image patch around each local interest point, and then any two local points belonging to different images are matched if their local descriptors are close enough in the feature space. Scale-invariant feature transform (SIFT) [3] is a well-known method for extracting interest points and their local descriptors from a given image. The SIFT features are robust to minor appearance changes engendered in a local image patch by varying environmental conditions (e.g., viewpoint, illumination, noise, and resolution).

However, the robustness of SIFT features is limited to small changes only. The appearance near a local point can vary widely due to significant changes in environmental conditions, leading to a large variance of the SIFT features extracted from the point. Such instability of local features is a great difficulty in image matching and its applications. Another issue is distinctiveness. Since a local descriptor represents very limited information about a local point, two local descriptors having different contexts can be rather close in the feature space when their local image patches look similar to each other. Such ambiguity is also a major limitation of local features.

Extensive research has been conducted to overcome these limitations. One line of research is to develop more robust and distinctive local features, which include PCA-SIFT [12], multi-step feature extraction [13], the Walsh–Hadamard transform [14], and kernel descriptors [15]. However, compared to SIFT, those customized features are rather complicated to compute and not widely proven in their general performance or applicability.

Another line of research is *descriptor learning* [16–19], which consists in learning a projection that maps given local features (e.g., SIFT features) to a new feature space where matching descriptors are closer to each other and non-matching descriptors are farther from each other. To this end, two categories of training data are required:

1. *Relevant* descriptors (or matching descriptors) that belong to the same class and thus should be *closer* to each other for better robustness to intra-class changes.
2. *Irrelevant* descriptors (or non-matching descriptors) that belong to different classes and thus should be *farther* from each other for more inter-class discriminative power in local description.

* Corresponding author: Tel.: +82 54 279 2259; fax: +82 54 279 2299.
E-mail addresses: fenrir@postech.ac.kr (K.-H. Kim), ruicai@microsoft.com (R. Cai), leizhang@microsoft.com (L. Zhang), seungjin@postech.ac.kr (S. Choi).

[1] There are many synonyms for these, including local points, interest points, and physical points. In this paper, we use each synonym according to context, while all have the same meaning.

http://dx.doi.org/10.1016/j.neucom.2014.07.029
0925-2312/

Descriptors that belong to the same class are extracted from the same local point in various images taken under different environmental conditions. Compared to customized features, descriptor learning can easily be incorporated into any existing local features to improve their intra-class robustness and inter-class distinctiveness. We address descriptor learning also because of its wide applicability.

In this paper, we present a novel learning strategy to further improve the performance gain of descriptor learning. First, we show that the *pairwise distance* between local descriptors in the original feature space can be a strong clue for determining which kind of training data are essential for descriptor learning. We define four categories of local descriptor pairs according to the pairwise distance and relevance of each pair:

1. *Relevant-Near* (*Rel-Near*): A relevant pair lying quite close to each other in the original feature space. We define a pair as Rel-Near if both descriptors belong to the same class and one descriptor is among the *k nearest neighbors* (*k* NNs, $k=5$) of the other descriptor, considering all descriptors in that class. Because it is already well matched, such a pair is *not* very worthwhile as training data for improving the matching performance of local features.
2. *Relevant-Far* (*Rel-Far*): A relevant pair, but not close enough. We define Rel-Far pairs as all pairs of relevant descriptors except for Rel-Near pairs. For a Rel-Far pair, two descriptors are extracted from the same local point, but have significant differences in their feature values due to varying environmental conditions. Thus, Rel-Far pairs are important for training in order to improve the robustness of local features against intra-class variations.
3. *Irrelevant-Near* (*Irr-Near*): An irrelevant pair, but close enough to be easily mistaken as matching descriptors. We define an Irr-Near pair as a local descriptor and its *k* NN descriptors, considering all irrelevant descriptors. The small pairwise distance implies that Irr-Near pairs mostly lie near a boundary between different classes. Thus, Irr-Near pairs are important for training in order to improve the inter-class distinctiveness of local features.
4. *Irrelevant-Far* (*Irr-Far*): An irrelevant pair far apart. We define Irr-Far pairs as all pairs of irrelevant descriptors except for Irr-Near pairs. Irr-Far pairs contain the overall scattering information between classes, but most of the pairs are already well separated in the original feature space, so these are *not* very important as training data.

Fig. 1(a) shows the distribution of the pairwise distances in the SIFT feature space, where 2×10^4 pairs are randomly chosen in each category from among 5×10^5 SIFT descriptors. According to their distances, Rel-Near and Irr-Far pairs are already well separated in the SIFT space. By contrast, a significant overlap exists between Rel-Far and Irr-Near distributions ($\sim$30% of their area), i.e., many Rel-Far pairs lie farther than Irr-Near pairs in the SIFT space. Thus, *the success of descriptor learning highly depends on the success in differentiating between Rel-Far and Irr-Near pairs*.

In order to further emphasize Rel-Far and Irr-Near pairs for learning, we propose a regularized learning framework in which each category of training pairs is weighted differently in costs. We seek a linear projection $\boldsymbol{T}$ that maximizes the ratio of variances between matching and non-matching differences:

$$J(\boldsymbol{T})=\frac{\beta_{IN}\sum_{(i,j)\in\mathcal{P}_{IN}}d_{ij}(\boldsymbol{T})+\beta_{IF}\sum_{(i,j)\in\mathcal{P}_{IF}}d_{ij}(\boldsymbol{T})}{\beta_{RN}\sum_{(i,j)\in\mathcal{P}_{RN}}d_{ij}(\boldsymbol{T})+\beta_{RF}\sum_{(i,j)\in\mathcal{P}_{RF}}d_{ij}(\boldsymbol{T})}, \tag{1}$$

where $d_{ij}(\boldsymbol{T})$ denotes the squared distance $||\boldsymbol{T}(\boldsymbol{x}_i-\boldsymbol{x}_j)||^2$, and $\mathcal{P}_{RN}$, $\mathcal{P}_{RF}$, $\mathcal{P}_{IN}$, $\mathcal{P}_{IF}$ denote the training sets belonging to Rel-Near, Rel-Far, Irr-Near, and Irr-Far, respectively. We introduce four regularization constants $\beta_{RN}, \beta_{RF}, \beta_{IN}, \beta_{IF}$ to control the importance of each category appropriately. In [16], $\beta_{RN}=\beta_{RF}=\beta_{IN}=\beta_{IF}=1$, i.e., all pairs are equally important regardless of their pairwise distances. We propose setting the regularization constants as follows:

1. $\beta_{RN}\ll\beta_{RF}$ to enhance the contribution of Rel-Far pairs;
2. $\beta_{IN}\gg\beta_{IF}$ to enhance the contribution of Irr-Near pairs.

In this way, our method can separate Rel-Far and Irr-Near pairs more clearly in the projected feature space (Fig. 1(b)).

We also provide an extension of our method to nonlinear learning. By adapting recent work on kernelization with *explicit feature maps* [20], our nonlinear learning method has more discriminative power but imposes the same computational cost as its linear counterpart.

In image retrieval experiments on landmark buildings in Oxford and Paris, our selective learning scheme outperformed existing descriptor learning methods and achieved a considerable

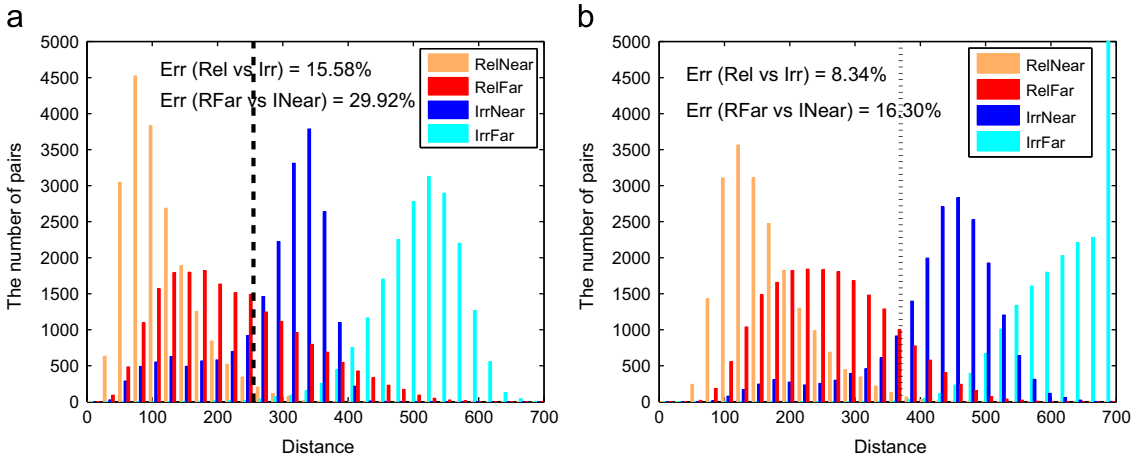

Fig. 1. (a) Distribution of Euclidean distance in SIFT feature space for four categories of local descriptor pairs: Rel-Near, Rel-Far, Irr-Near, and Irr-Far. The Bayes optimal error rates are also shown. *Err (Rel vs. Irr)* measures the proportion of overlapping region between {Rel-Near, Rel-Far} and {Irr-Near, Irr-Far}, while *Err (RFar vs. INear)* measures the overlap between Rel-Far and Irr-Near. (b) Distance distribution and Bayes optimal error are obtained by our learning method. Compared to the original SIFT space, the regions between relevant and irrelevant pairs or between Rel-Far and Irr-Near pairs are more separable.