Pattern Recognition 73 (2018) 275–288

Contents lists available at ScienceDirect

Pattern Recognition

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

Deep adaptive feature embedding with local sample distributions for person re-identification

Lin Wu[a], Yang Wang[b,*], Junbin Gao[c], Xue Li[a]

[a] The University of Queensland, Queensland 4072, Australia
[b] The University of New South Wales, NSW 2052, Australia
[c] Discipline of Business Analytics, The University of Sydney Business School, The University of Sydney, NSW 2006, Australia

ARTICLE INFO

Article history:
Received 22 April 2017
Revised 25 August 2017
Accepted 27 August 2017
Available online 31 August 2017

Keywords:
Deep feature embedding
Person re-identification
Local positive mining

ABSTRACT

Person re-identification (re-id) aims to match pedestrians observed by disjoint camera views. It attracts increasing attention in computer vision due to its importance to surveillance systems. To combat the major challenge of cross-view visual variations, deep embedding approaches are proposed by learning a compact feature space from images such that the Euclidean distances correspond to their cross-view similarity metric. However, the global Euclidean distance cannot faithfully characterize the ideal similarity in a complex visual feature space because features of pedestrian images exhibit unknown distributions due to large variations in poses, illumination and occlusion. Moreover, intra-personal training samples within a local range which are robust to guide deep embedding against uncontrolled variations cannot be captured by a global Euclidean distance. In this paper, we study the problem of person re-id by proposing a novel sampling to mine suitable *positives* (*i.e.,* intra-class) within a local range to improve the deep embedding in the context of large intra-class variations. Our method is capable of learning a deep similarity metric adaptive to local sample structure by minimizing each sample's local distances while propagating through the relationship between samples to attain the whole intra-class minimization. To this end, a novel objective function is proposed to jointly optimize similarity metric learning, local positive mining and robust deep feature embedding. This attains local discriminations by selecting local-ranged positive samples, and the learned features are robust to dramatic intra-class variations. Experiments on benchmarks show state-of-the-art results achieved by our method.

1. Introduction

The re-identification (re-id) of individuals across spatially disjoint camera views has attracted tremendous attention in computer vision community due to its practice into security and surveillance systems. Despite years of great efforts, person re-id still remains a challenging task due to its large variations in terms of view points, illuminations and different poses (See examples in Fig. 1 (a)). Existing approaches to person re-id can be summarized into two categories. The first category focuses on developing robust descriptors to describe a person's appearance against challenging factors (lighting, pose, etc) while preserving identity information [1–4]. Low-level features such as color [4], texture (Local Binary Patterns [1], and Gabor [3]) are commonly used for this purpose. However, direct matching pedestrians based on hand-crafted features is not distinctive and reliable enough to severe changes and misalignment across camera views. The second category [5–10] comes up with the metric learning problem which is to discriminate distance metrics from training data consisting of cross-camera matched pairs, under which inter-class and intra-class variations of pedestrian samples are maximized and minimized, respectively. They, however, consider feature extraction and metric learning as two independent components, leading to a suboptimal performance. Moreover, such methods focus on optimizing a linear transformation on the input, which has a limited number of parameters and fail to model the higher-order correlations over the original data dimensions.

More recent studies on deep embedding methods [5,11–17] aim at learning a compact feature embedding $f(x) \in \mathbb{R}^d$ from image x via a deep convolutional neural network (CNN). The embedding objective is usually modeled over Euclidean space: the Euclidean distance $D(x_i, x_j) = ||f(x_i) - f(x_j)||_2$ between feature vectors should preserve the semantic relationship encoded in

* Corresponding author.
E-mail addresses: lin.wu@uq.edu.au, Jolin.lwu@gmail.com (L. Wu), wangy@cse.unsw.edu.au (Y. Wang), junbin.gao@sydney.edu.au (J. Gao), xueli@itee.uq.edu.au (X. Li).

http://dx.doi.org/10.1016/j.patcog.2017.08.029
0031-3203/

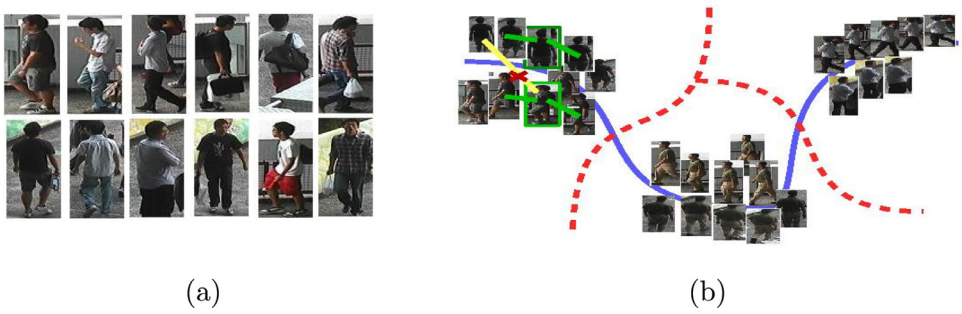

Fig. 1. (a) Samples of pedestrian images from the CUHK03 dataset [11]. Each column shows two images of the same identity observed by two disjoint camera views. (b) Highly-curved manifolds of 3 identities. Positive samples in a local range (green lines) should be selected to guide deep feature embedding while those in large distance (yellow line with cross) should not be sampled to respect the manifold structure. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

pairwise (by contrastive loss [11–14]), in triplets [16,17], or even high-order relationships [18]. Among these methods, hard sample mining is crucial to ensure the quality and the learning efficiency, due to the fact that there are many more easy examples than those meaningful hard examples. Thus, they usually choose hard samples to compute the convenient Euclidean distance in the embedding space. However, these deep embedding methods suffer from inherent limitations: First, they adopt a global Euclidean distance metric to evaluate the hard samples whereas recent manifold learning in person re-id [19] suggests that pedestrian samples are distributed as highly-curved manifolds. Euclidean distance can only be adopted in local range to approximate the geodesic distance via graphical relationship between samples (as illustrated in Fig. 1 (b)). Second, these methods are conditioned on individual samples in term of pairs/triplets to categorize the inputs as depicting either the same or different subjects. Such mapping to a scalar prediction of similarity score based on person identities would make the optimization on CNN parameters over-fitting because the supervision binary similarity labels (0 for dissimilar and 1 for similar) tend to push the scores independently. In practice, the similarity scores of positive and negative pairs live on a 1-D manifold following the distribution on heterogeneous data [20]. Finally, when training the CNN with contrastive or triplet loss for embedding, existing methods use the Euclidean distance indiscriminately with all the positive samples. Nonetheless, we observe that selecting positive samples within local ranges (pairs in green lines in Fig. 1 (b)) is critical for training whilst enforce training with the positive samples of long distance may distort the manifold structure (the yellow line with red cross in Fig. 1 (b)). Moreover, objective functions defined on triplet loss involve sampling on divergent triplets, which is not necessarily consistent, and thus impedes the convergence rate and training efficiency.

Our Approach. Mitigating the aforementioned issues, in this paper we propose a principled approach to learn a local-adaptive similarity metric, which will be exploited to search for suitable positive samples in a local neighborhood to facilitate a more effective yet efficient deep embedding learning. The key challenge lies in the design of robust feature extraction and the loss function that can jointly consider 1) similarity metric learning; 2) suitable positive sample selection; and 3) deep embedding learning. Existing deep embedding studies [11,14,15,19,21,22] only consider the two later objectives but not jointly with the first important aspect. To this end, we propose a principled approach to train a deep network that transforms the input data into a deep feature space where the local data distribution structure within classes can be captured. We formulate the feature extractor as stacked convolutional Restricted Boltzmann Machines (CRBMs) [23] to initialize the parameters that define the mapping from input images to their representation space. We remark that CNN has generic parameterization while in person re-id case, body parts exhibit different visual modalities due to the combinations of view points, poses, and photometric settings. Thus, a single/generic CNN filter cannot capture the inter-camera variations while some fine-grained information such as “texture in clothes” and “bags” are very helpful in reducing intra-personal variations. As such, CRBMs serve as hierarchical feature model to faithfully describe pedestrian samples containing dramatic variations. We formulate the training of CRBMs adaptively to search the suitable positive samples within local range so that it learns locally adaptive metric (instead of global Euclidean distance). Furthermore, to improve training efficiency, we employ variance reduced Stochastic Gradient Descent (SGD) [24] to share and reuse past stochastic gradients across data samples by exploiting their neighborhood structure. As shown in Fig. 2, the proposed metric yields similarity scores in mini-batch, from which positive samples constituting a hard quadruplet are mined and used to optimize the feature embedding space. The similarity metric learning and embedding learning in the associated CRBMs are jointly optimized via a novel large-margin criterion.

Contributions. The main contributions of our work are four-fold: (1) An improved deep embedding approach is presented to construct a representation amenable to similarity metric computation in person re-identification by jointly optimizing robust feature embedding, local adaptive similarity learning, and suitable positive mining. (2) The proposed method enhances the quality of learned representations and the training efficiency by accessing Euclidean distance of samples in local range w.r.t highly-curved structure. This allows adaptive similarity access in local range and achieves minimization of intra-class variations by local-ranged positive sample mining. (3) We provide alternative to CNN embedding by formulating a stacked CRBMs into local sample structure in deep feature space, and thus enables local adaptive similarity metric learning as well as plausible positive mining. (4) Our method achieves state-of-the-art results on four benchmark datasets: VIPeR [25], CUHK03 [11], CUHK01 [26], and Market-1501 [27].

2. Related work

2.1. Metric learning in person re-identification

Metric learning algorithms have been extensively applied into person re-identification to learning discriminative distance metrics or subspaces for matching persons across views [2,5–7,9,10,19,26,28–31]. They essentially perform a two-stage pipeline where hand-crafted features are extracted for each image, and then a Mahalanobis form metric is learned. This corresponds to a