Pattern Recognition Letters 112 (2018) 234–240

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Accelerated low-rank sparse metric learning for person re-identification

Niki Martinel

Department of Mathematics and Computer Science, University of Udine, Via Delle Scienze, 206, Udine 33100, Italy

ARTICLE INFO

Article history:
Received 12 January 2018
Available online 31 July 2018

Keywords:
Person re-identification
Metric learning
Low-rank manifold
Proximal gradient optimization

ABSTRACT

Person re-identification is an open and challenging problem in computer vision. A surge of effort has been spent design the best feature representation, and to learn either the transformation of such features across cameras or an optimal matching metric. Metric learning solutions which are currently in vogue in the field generally require a dimensionality reduction pre-processing stage to handle the high-dimensionality of the adopted feature representation. Such an approach is suboptimal and a better solution can be achieved by combining such a step in the metric learning process. Towards this objective, a low-rank matrix which projects the high-dimensional vectors to a low-dimensional manifold with a discriminative Euclidean distance is introduced. The goal is achieved with a stochastic accelerated proximal gradient method. Experiments on two public benchmark datasets show that better performances than state-of-the-art methods are achieved.

1. Introduction

The problem of tracking pedestrian *within* a single camera field-of-view (FoV), i.e.intra-camera tracking, has been on the edge of computer vision for decades (e.g.[13,24,37,46]). Only recently, motivated by the increasing request of more safety, there has been raising interest in the problem of tracking a person moving *across* disjoint camera FoVs, i.e., inter-camera tracking with no temporal constraints. This is known as the person re-identification problem. Such a problem is very attractive since, for video surveillance applications, knowing whether a person is present in the monitored area at a precise time instant is of paramount importance.

The problem has received increasing attention (see [41] for a recent survey) due to its intrinsic open challenges. Person images generally have low spatial resolution which makes the acquisition of discriminating biometric features unreliable. This motivated the community to mainly exploit visual appearance features, thus opening to the complex variations of individuals clothing appearance (due to viewing angles, lighting, background clutter, and occlusions).

A swell of efforts has been devoted to address the stated problems by following three main approaches. These are (i) Discriminative signature based methods –which seek to design the best visual representation that is robust to the aforementioned challenges. (ii) Feature transformation based methods –that aim to model the transformation of visual features that occur between pairs of cameras. (iii) Metric Learning based approaches –which focus on learning discriminant metrics yielding an optimal matching score/distance between gallery and probe images.

Despite such efforts, the first family of works suffer from significant limitations including the fact that a hand-crafted visual representation is generally not sufficiently robust to the illumination, pose, occlusion challenges that are widely common in re-identification. Indeed, as shown in Fig. 1, the projection of a same visual feature in the feature space of a disjoint camera is likely to return a wrong list of matches. On the other hand, if a distance function can be learned in such a way that features of a same person acquired from two cameras projected onto a shared feature space will be "closer" than features of different persons, then the re-identification goal can be better tackled (Fig. 2).

Motivation and contribution: While widely explored, metric learning-based solutions generally require a pre-processing stage to handle the high-dimensionality of the adopted feature representation. This usually translates into the application of Principle Component Analysis (PCA) before the learning process starts. It is reasonable to believe that such an approach is suboptimal and may induce severe misclassification errors due to the rejection –by PCA– of low-variance components which may carry relevant information for the re-identification task. A better solution could be achieved by combining both the metric learning and the dimensionality reduction steps in such a way that relevant components are identified by means of the re-identification performance. The core contribution of this work is a metric learn-

E-mail address: niki.martinel@uniud.it

https://doi.org/10.1016/j.patrec.2018.07.033
0167-8655/

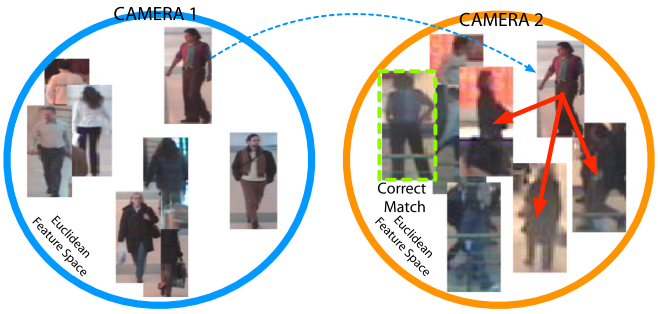

Fig. 1. Example of tackling the re-identification problem by projection of visual features between camera dependent Euclidean feature spaces.

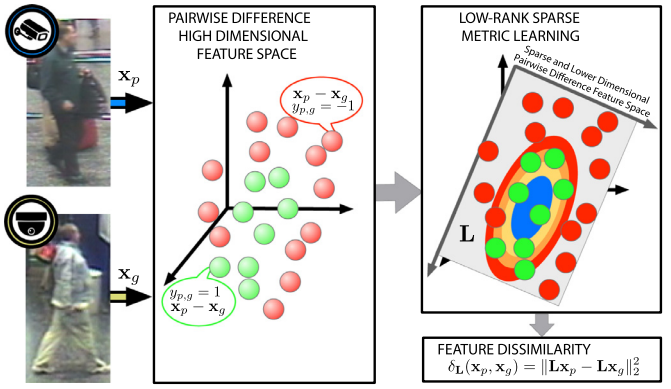

Fig. 2. Proposed approach pipeline. High-dimensional visual features are extracted from image pairs acquired by non-overlapping cameras. Common metric learning solutions apply a pre-processing stage to handle such a high dimensional space. In the proposed approach, the metric learning and the dimensionality reduction steps are combined in such a way that relevant components are identified by means of the re-identification performance.

ing approach, named Accelerated Low-Rank sparse Metric learning (ALRM), which produces a low-rank solution that self-determines the discriminative dimensions of the underlying manifold.

In the experimental section, results show that ALRM has similar or better performances than the ones of such state-of-the-art solutions on two benchmark datasets.

2. Related work

Person re-identification has recently become an expansive field of research [41]. A surge of effort has been devoted to attack the problem from different perspectives, ranging from partially seen persons [49] to low resolution images [19], which can eventually be synthesized in the open-world re-identification idea [53]. In the following, a brief overview of the literature is given by discussing relevant works following the three main research directions.

Discriminative signature based methods have been the most widely used ones. Salient color names [45] and the color distribution structure [16] were proposed as robust feature descriptors. To tackle pose changes, correlation between random patches extracted from pair of images were also exploited [28]. To handle background and illumination variations, feature representations based on the combination of Biologically Inspired Features (BIF) [26] and Covariance descriptors [21] were introduced. Works going beyond appearance features and integrating a semantic aspect into image representations were proposed by leveraging on the interaction between attributes and appearance [14] and by describing the relations among the low-level part features, middle-level clothing attributes, and high-level re-identification labels [17].

Feature transformation methods have addressed the re-identification problem by modeling the transformation of visual features that occur between pairs of cameras. An early work was proposed in [11], where the Brightness Transfer Function (BTF) computed between the appearance features was used to match persons across camera pairs. More recently, local eigen-dissimilarities between multiple features extracted from image pairs [29] as well as warp function space of feature transformations [27] were explored. Cameras viewpoint information was also explored to train binary [8,18] and dictionary-based classifiers [12] according to the similarity of cross-view transforms.

Metric Learning approaches focus on learning discriminant metrics which aim to yield an optimal matching score/distance between a gallery and a probe image. Since the early work of [44], many different solutions have been introduced [4]. More specifically, in the re-identification field metric learning approaches have been proposed by relaxing [10] or enforcing [22] the PSD conditions as well as by considering equivalence constraints [40]. While most of the existing methods capture the global structure of the dissimilarity space, local solutions [20,34] have been proposed too. Motivated by the success of both solutions, approaches combining them in metric ensembles [30,33] have been introduced.

Different solutions yielding similarity measures have also been investigated by proposing to learn listwise [6] and pairwise [52] similarities as well as mixture of polynomial kernel-based models [5]. Related to these similarity learning models are the deep architectures which have been exploited to tackle the task [39,54].

With respect to all such methods, two recent works [21,22] share the idea of finding discriminative low-rank projections, however, there are significant differences with the proposed method. Specifically, this work introduces: (i) An accelerated convex optimization solver which reduces the per-epoch computational complexity while improving the convergence rate. This is obtained by formulating the problem in a stochastic fashion and by avoiding to visit all the training samples in each epoch before updating the metric parameters. (ii) An additional sparsity regularizer on the low-rank projection that allows to self-discover the relevant components of the underlying manifold, thus avoiding the specification of its dimensionality beforehand.

3. Methodology

3.1. Preliminaries and definitions

Let $\mathcal{P} = \{\mathbf{I}_p\}_{p=1}^{|\mathcal{P}|}$ and $\mathcal{G} = \{\mathbf{I}_g\}_{g=1}^{|\mathcal{G}|}$ be the set of probe and gallery images acquired by two disjoint cameras. Let $\mathbf{x}_p \in \mathbb{R}^d$ and $\mathbf{x}_g \in \mathbb{R}^d$ be the feature representations of $\mathbf{I}_p$ and $\mathbf{I}_g$ of two persons p and g. Let $\mathcal{X} = \{(\mathbf{x}_p, \mathbf{x}_g; y_{p,g})^{(i)}\}_{i=1}^{n}$ denote the training set of $n = |\mathcal{P}| \times |\mathcal{G}|$ probe-gallery pairs where $y_{p,g} \in \{-1, +1\}$ indicates if p and g are the same person $(+1)$ or not (-1). Finally, let an *iteration* be a parameter update computed by visiting a single sample and let an *epoch* denote a complete cycle on the training set.

3.2. Low-rank sparse metric learning

In this section, the core contribution of this work is discussed: a low-rank sparse metric learning approach along with an accelerated stochastic solver.

Objective The image feature representations $\mathbf{x}$ might be very high-dimensional and contain undiscriminative components. To address such a problem, metric learning approaches generally apply PCA before the learning process starts.

To embed similar PCA capabilities in discovering the discriminative components within the learned metric, a matrix $\mathbf{L} \in \mathbb{R}^{r \times d}$ that