Pattern Recognition 47 (2014) 3767–3778

Contents lists available at ScienceDirect

Pattern Recognition

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

Leveraging local neighborhood topology for large scale person re-identification

Svebor Karaman [a,*], Giuseppe Lisanti [a], Andrew D. Bagdanov [b], Alberto Del Bimbo [a]

[a] Media Integration and Communication Center (MICC), University of Florence, Viale Morgagni 65, Firenze 50134, Italy
[b] Computer Vision Center, Barcelona, Universitad Autonoma de Barcelona, Bellaterra, Spain

ARTICLE INFO

Article history:
Received 8 December 2013
Received in revised form
3 April 2014
Accepted 7 June 2014
Available online 17 June 2014

Keywords:
Re-identification
Conditional random field
Semi-supervised
ETHZ
CAVIAR
3DPeS
CMV100

ABSTRACT

In this paper we describe a semi-supervised approach to person re-identification that combines discriminative models of person identity with a Conditional Random Field (CRF) to exploit the local manifold approximation induced by the nearest neighbor graph in feature space. The linear discriminative models learned on few gallery images provides coarse separation of probe images into identities, while a graph topology defined by distances between all person images in feature space leverages local support for label propagation in the CRF. We evaluate our approach using multiple scenarios on several publicly available datasets, where the number of identities varies from 28 to 191 and the number of images ranges between 1003 and 36 171. We demonstrate that the discriminative model and the CRF are complementary and that the combination of both leads to significant improvement over state-of-the-art approaches. We further demonstrate how the performance of our approach improves with increasing test data and also with increasing amounts of additional unlabeled data.

1. Introduction

Person re-identification is the problem of identifying previously seen individuals on the basis of one or more images captured from one or more cameras. It is an important aspect of modern surveillance systems as it is a way of maintaining identity information about targets in multiple views over potentially long periods of time. Re-identification is normally formulated in terms of a set of *gallery images* for each individual of interest, and a set of *probe images* which should be re-identified by determining the corresponding gallery individual. The problem is difficult due to changes in illumination and pose, occlusions, similarity of appearance, and changes in camera view. All of these render difficult the search for discriminative feature representations that capture identity and are robust to changing imaging conditions.

Re-identification performance is traditionally evaluated as a retrieval problem. The given data at training time is a gallery set consisting of a single or several images of known individuals. At test time, for each probe image or group of probe images of an unknown person, the goal of re-identification is usually to return a ranked list of individuals from the gallery. In the re-identification literature there are three main scenarios:

- *Single-vs-Single (SvsS) re-identification:* in which *exactly one* example image of each person is provided in the gallery and *at least one* instance of each person is present in the probe set.
- *Multi-vs-Multi (MvsM) re-identification:* in which a *group* of *M* examples of each individual is given in the gallery and a *group* of *M* examples of each individual is given to be re-identified in the probe set.
- *Multi-vs-Single (MvsS) re-identification:* in which *multiple images* of each person are given as groups in the gallery image set, and *exactly one* example image of each person is given in the probe set. The MvsS re-identification modality is little used and at times misinterpreted in the literature. It is not a realistic reflection of real-world application scenarios and we do not consider it further in this work.

This breakdown of re-identification scenarios as used in the literature focuses primarily on the number of examples of each person given in the gallery and probe image sets. We feel that is more useful, instead, to think of re-identification problems as either "structured" or "unstructured". Structured re-identification problems are ones in which knowledge that multiple images depict the same person is given *a priori* in the problem definition. A typical example of *structured re-identification* is the multi-versus-multi

* Corresponding author. Tel.: +39 055 275 1390; fax: +39 055 275 1396.
E-mail addresses: svebor.karaman@unifi.it (S. Karaman), giuseppe.lisanti@unifi.it (G. Lisanti), bagdanov@cvc.uab.es (A.D. Bagdanov), alberto.delbimbo@unifi.it (A. Del Bimbo).

http://dx.doi.org/10.1016/j.patcog.2014.06.003
0031-3203/

(MvsM) scenario in which perfect, hard group structure is known about both the gallery and the test image sets.[1] We refer to the grouping structure in MvsM re-identification as *hard* because the groups of images corresponding to the same person are perfectly known. Note that perfect group knowledge is almost never known in practice, and knowledge of group structure significantly simplifies re-identification problems. Group knowledge in probe image sets can be used to apply voting schemes or similar methods to obtain more robust re-identification than assignment of individual probe image to nearest neighbors, for example.

In contrast, "unstructured" scenarios may contain multiple observations of a person without explicitly giving the corresponding group structure. An example of *unstructured re-identification* is the single-versus-single (SvsS) shot scenario in which there may be multiple images of the same person in the probe image set, but this grouping structure is not known. We sometimes refer to this type of re-identification problem as single-versus-all (SvsAll) to emphasize the fact that the task is to identify *all* individual probe images [1]. Note that unstructured problems are much more difficult than structured ones. Fig. 1 illustrates the difference between structured and unstructured re-identification. In the structured case, the unit of re-identification is a *group* of images of the same person, while in the unstructured case the unit of re-identification is each *individual image*.

One of the greatest benefits in structured re-identification is that having multiple aspects of the same person in both the probe and gallery greatly increases the likelihood of finding at least one good match. Our objective with this work is to bring some of the benefits of structured re-identification problems to unstructured ones. To this end, we propose to infer a sort of soft group structure that can then be used to improve re-identification accuracy. We use the nearest-neighbor topology of all available gallery and probe imagery in feature space to accomplish this. This topology, which is essentially an estimation of the local data manifold, is used to adapt simple discriminative models of person identity to better reflect the structure of the data in feature space. A unique advantage of our approach is that we are able to exploit not only gallery imagery, but all available imagery when performing re-identification.

Re-identification has largely been limited to relatively small benchmarks that contain a few hundred or a few thousand images. In this work we also demonstrate how our semi-supervised approach scales well to very large datasets containing tens of thousands of images to label, and we are the first to perform re-identification experiments on such large datasets. Furthermore, in contrast to most discriminative techniques, the performance of our approach *improves* with increasing amounts of test data. Even the addition of unlabeled, anonymous images for which we do not desire labels helps improve the performance of our approach due to better manifold sampling in the nearest-neighbor topology.

In the next section we review relevant work from the literature on person re-identification. We describe our approach to combining structure discovery with discriminative models in Section 3. In Section 4 we report on a series of experiments we performed to explore the performance of our algorithm and compare our approach to the state-of-the-art. Finally, we conclude in Section 5 with a discussion of our contribution.

2. Related work

In this section we review the major trends in re-identification research, which can be broadly categorized into approaches that focus mostly on sophisticated appearance models for re-identification and approaches that focus more on learning discriminative classifiers or metrics for re-identification. We give an overview of our proposed approach with respect to previous work in Section 2.3.

2.1. Appearance modeling for re-identification

The majority of existing research on person re-identification has concentrated on the development of sophisticated features for describing the visual appearance of targets. In [2] were introduced discriminative appearance-based models using Partial Least Squares (PLS) over texture, gradients and color features. The authors of [3] use an ensemble of local features learned using a boosting procedure, while in [4] the authors use a covariance matrix of features computed in a grid of overlapping cells. The SDALF descriptor introduced in [5] exploits axis symmetry and asymmetry and represents each part of a person by a weighted color histogram, maximally stable color regions and texture information from recurrent highly structured patches. In [6] the authors fit a Custom Pictorial Structure (CPS) model consisting of head, chest, thighs and legs part descriptors using color histograms and Maximally Stable Color Region (MSCR). The Global Color Context (GCC) of [7] uses a quantization of color measurements into color words and then builds a color context modeling the self-similarity for each word using a polar grid. The Asymmetry-based Histogram Plus Epitome (AHPE) approach in [8] represents a person by a global mean color histogram and recurrent local patterns through epitomic analysis. More recently, the authors of [9] suggested that re-identification can be performed by exploiting small salient regions in each person image. They applied adjacency-constrained patch matching to build dense correspondence between image pairs through an unsupervised saliency learning method that does not require identity labels during training.

Two limitations of appearance-based approaches are that they often attempt to fit complex appearance models to target images of limited quality, and that they typically use aggregate or mean appearance models over multiple observations of the same individual (for multi-shot modalities). Both of these can be serious limitations in practice, since much surveillance imagery is resolution-limited and mean appearance models may not exploit well all available imagery.

2.2. Learning for re-identification

Differently than the approaches mentioned above, learning-based ones concentrate specifically on the classification or ranking technique applied to re-identify probe images. Techniques based on learning can be roughly grouped into metric learning approaches and those that learn strong discriminative models for classification or ranking. The approach in [10] is a supervised technique that uses pairs of similar and dissimilar images and a relaxed RankSVM algorithm to rank probe images. A set-based discriminative ranking approach was also recently proposed which alternates between optimizing a set-to-set geometric distance and a feature space projection, resulting in a discriminative set-distance-based model [11]. The Probabilistic Relative Distance Comparison approach learns a metric under which the probability of an incorrect match having a small distance is less than that of a correct one [12]. Camera transfer approaches have also been proposed that use images of the same person captured from different cameras to learn metrics [13,14].

The authors of [15] propose a method for person appearance matching across disjoint camera views by learning a model that selects the most descriptive features for a specific class of objects.

[1] Note that throughout the paper we use *gallery images*, *training examples*, *training data* and *training samples* interchangeably to refer to images from the gallery. Similarly, we use *probe images*, *test images*, *test samples* and *test data* interchangeably to refer to images from the probe set.