J. Vis. Commun. Image R. 38 (2016) 115–129

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

Modeling feature distances by orientation driven classifiers for person re-identification ☆

Jorge García [a], Niki Martinel [b], Alfredo Gardel [a], Ignacio Bravo [a], Gian Luca Foresti [b], Christian Micheloni [b],*

[a] Department of Electronics, University of Alcala, Alcalá de Henares 28801, Spain
[b] Department of Mathematics and Computer Science, University of Udine, Udine 33100, Italy

ARTICLE INFO

Article history:
Received 14 May 2015
Accepted 16 February 2016
Available online 27 February 2016

Keywords:
Person re-identification
Pose and orientation recovery
Appearance models
Surveillance systems
Scene understanding
Tracking
Feature transformation
Dissimilarity learning

ABSTRACT

To tackle the re-identification challenges existing methods propose to directly match image features or to learn the transformation of features that undergoes between two cameras. Other methods learn optimal similarity measures. However, the performance of all these methods are strongly dependent from the person pose and orientation. We focus on this aspect and introduce three main contributions to the field: (i) to propose a method to extract multiple frames of the same person with different orientations in order to capture the complete person appearance; (ii) to learn the pairwise feature dissimilarities space (PFDS) formed by the subspaces of similar and different image pair orientations; and (iii) within each subspace, a classifier is trained to capture the multi-modal inter-camera transformation of pairwise image dissimilarities and to discriminate between positive and negative pairs. The experiments show the superior performance of the proposed approach with respect to state-of-the-art methods using two publicly available benchmark datasets.

1. Introduction

The person re-identification problem, formally defined as the problem of associating a given person acquired by a camera to the same person previously acquired by any other camera in the network at any location and at any time instant, is increasingly gaining attention by the community. This challenging task is very important for surveillance applications such as inter-camera tracking [1], multi-camera behavior analysis [2], and access granting [3]. Despite the problem can be alleviated by deploying a large number of sensors such as all the areas of the monitored environment are covered by camera field-of-views (FoVs), the costs of system installation, and maintenance lead to a non-feasible solution. Thus, in a real scenario, we have to deal with partial area coverage that yields to the re-identification problem.

1.1. Motivation

Despite much effort has been spent by the community to find the best signature (e.g. [4–7]), to learn the feature transformation that undergoes between camera pairs (e.g. [8–11]) and to find the optimal similarity measure (e.g. [12–15]), re-identify a person that moves across disjoint cameras still remains an open issue. Almost all existing works assume that a uni-modal inter-camera transformation of features occurs between two camera views. However, we believe that the deployment and the configuration of the cameras (it is a combination of view points, illumination changes, and photometric settings, etc.) together with the appearance of a person give rise to multi-modal inter-camera transformations (see Fig. 1 for an example). In particular, current methods highly suffer the strong pose and orientation changes that may occur when a person moves between cameras FoVs.

1.2. Contribution

Motivated by these, and inspired by the fact that as the transformation between appearance features is multi-modal, so is the transformation of the distances between them [16], we introduce the following contributions:

1. we build upon the idea that the transformation learned for the same person seen from different viewpoints may be less reliable than the one learned for the same person seen from the same point of view. Hence, we introduce a method to recover images of the same person with different orientations so as we can capture the multi-modal appearance of a person with higher reliability;

☆ This paper has been recommended for acceptance by Zicheng Liu.

* Corresponding author.
E-mail addresses: jorge.garcia@depeca.uah.es (J. García), niki.martinel@uniud.it (N. Martinel), alfredo@depeca.uah.es (A. Gardel), ibravo@depeca.uah.es (I. Bravo), gianluca.foresti@uniud.it (G.L. Foresti), christian.micheloni@uniud.it (C. Micheloni).

http://dx.doi.org/10.1016/j.jvcir.2016.02.009
1047-3203/

Fig. 1. Images of a persons acquired by the same camera. Person appearances look very different among all the images due to the changes in pose and illumination conditions.

2. the person orientations are used to learn different models for the different transformations that exist between pairwise feature dissimilarities. We form the *pairwise feature dissimilarities space (PFDS)* and divide it into two main regions, the region for which the dissimilarities are computed between pair of images with similar orientation and the region containing all other different pairwise orientations;
3. within each region, a classifier is trained to capture all the possible multi-modal transformation of feature dissimilarities. Those are used to discriminate between pairwise images of the same person (positive pair) and pairwise images of two different persons (negative pair). This also allows to pose the re-identification as a binary classification problem.

The rest of the paper is organized as follows. A brief description of the related work is given in Section 2. In Section 3 a system overview and details of the modules that compose our re-identification approach are described. The superior performance of our approach over existing state-of-the-art methods is shown in Section 4. Finally, conclusions are drawn in Section 5.

2. Related work

In the recent past, the community has proposed to tackle the problem of person re-identification across non-overlapping cameras using several approaches that differs from the way the person body is modeled, to which features are used (i.e. biometrics or appearance), to how matches between individuals are computed, etc. [17,18]. The person appearance is defined as a set of features which are extracted from the person image region. This image representation through a set of features is not only used in person re-identification, face recognition [19], video summarization [20] or age classification [21] problems also include such stage in their framework. While recent works in the field of person re-identification can be grouped on the basis of any of such categories, we group them as follows: (i) methods that use discriminative appearance-based signatures, (ii) methods that capture the transformation of features across camera pairs, and (iii) methods that learn the optimal distance metric between appearance features.

Discriminative signature based methods are the most commonly explored approaches for person re-identification. In [5] particular interest has been focused on finding the best set of features that can be exploited to match persons across cameras. In [22] the objective was to model the spatial distribution of the appearance relative to each of the object part. In [23] multiple local and global features were used to create an invariant signature. A discriminative signature computed using the Mean Riemannian Covariance patches was used in [24]. In [4], frames were used to built a collaborative representation that best approximates the query frames. In [6], the distribution of color features projected in the log-chromaticity space was described using the shape context descriptor. In [25] an unsupervised framework was proposed to extract distinctive features, then a patch matching method was used together with adjacency constraints. In [7] a combination of biologically inspired features and covariance descriptors was proposed. In [26] local feature descriptors were encoded by fisher vectors and pooled to provide a global image representation. In [27] an articulated multiple-instance-based compositional template was proposed to model person appearance. Appearance features and similarities with a reference set of persons were used in [28]. Human saliency was also explored in [25,29–31] to reject body parts that are non-discriminative for the re-identification task. Recently, horizontal occurrence of local features [32], clothing attributes [33] and 3D body models [34] were explored to propose a stable representation against viewpoint changes. A sparse-based local matching technique aiming to find the optimal correspondence between image patches was proposed in [35]. Pose priors and subject-discriminative features were also investigated in [36].

Those methods rely only on the discriminative power of appearance features to perform a pure feature matching. While no training is required and good results can be achieved when images are similar, this is still an unreliable solution as such methods generally assume that features are not transformed between cameras.

Transformation learning based methods were explored in [8] to capture the transformation across non-overlapping cameras in a tracking scenario. Similarly, the problem of capturing the nonlinear transformation between features was addressed in [37]. In [38] pairwise dissimilarity profiles between categories were learned and exploited in a nearest neighbor classification framework. In [9] the implicit transformation function of features was learned by concatenating appearance feature vectors of persons viewed by different cameras. In [10] a Weighted Brightness Transfer Function that assigns unequal weights to observations based on how close they are to test observations was proposed. In [16], the error on the transformation function was modeled by a binary