Image and Vision Computing 32 (2014) 379–390

Contents lists available at ScienceDirect

Image and Vision Computing

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

Covariance descriptor based on bio-inspired features for person re-identification and face verification☆

Bingpeng Ma [a,1], Yu Su [b], Frédéric Jurie [b,*]

[a] School of Computer and Control Engineering, University of China Academy Science, Beijing, China
[b] GREYC — CNRS UMR 6072, University of Caen Basse-Normandie, Caen, France

ARTICLE INFO

Article history:
Received 9 August 2013
Received in revised form 10 January 2014
Accepted 2 April 2014
Available online 12 April 2014

Keywords:
Image representation
Person re-identification
Face verification
Biologically inspired features
Covariance descriptor

ABSTRACT

Avoiding the use of complicated pre-processing steps such as accurate face and body part segmentation or image normalization, this paper proposes a novel face/person image representation which can properly handle background and illumination variations. Denoted as gBiCov, this representation relies on the combination of Biologically Inspired Features (BIF) and Covariance descriptors [1]. More precisely, gBiCov is obtained by computing and encoding the difference between BIF features at different scales. The distance between two persons can then be efficiently measured by computing the Euclidean distance of their signatures, avoiding some time consuming operations in Riemannian manifold required by the use of Covariance descriptors. In addition, the recently proposed KISSME framework [2] is adopted to learn a metric adapted to the representation. To show the effectiveness of gBiCov, experiments are conducted on three person re-identification tasks (VIPeR, i-LIDS and ETHZ) and one face verification task (LFW), on which competitive results are obtained. As an example, the matching rate at rank 1 on the VIPeR dataset is of 31.11%, improving the best previously published result by more than 10.

1. Introduction

The task of person re-identification consists in recognizing an individual through different cameras in a distributed network or through the same camera capturing images at different time. This is a challenging problem that has attracted a lot of attention in recent years. The key issue of such systems lies in their ability to measure the similarity between two person-centered bounding boxes, i.e. to predict if they represent to the same person, despite changes in illumination, pose, viewpoint, background, partial occlusions and low resolution. In order to tackle this problem, the dominant strategy is to combine feature sets into templates, used as person descriptors, and to measure the similarity between templates to predict persons' identities. Descriptors adapted to the re-identification of faces are usually different than those for person re-identification. Indeed, face verification required to be able to capture smaller details of the input image, as intra-class and inter-class variations is smaller than that for person re-identification. It is challenging for a descriptor to handle both tasks at the same time. Finally, even if such person or face descriptors have received a lot of attention during the last decade, they still need some improvement before they can be used for real applications. This is the motivation for the presented work.

Extending the work presented in [3], this paper presents a novel image representation for person re-identification and face verification. Specifically, the proposed image representation allows to measure efficiently the similarity between two persons/faces without any preprocessing step (e.g., precise background subtraction or body part segmentation). This paper mainly focuses on person re-identification which has received less attention than face verification, however we experimentally demonstrate that the proposed representation also works well for face verification. In both scenarios, we assume that pedestrians/faces have been previously detected and cropped.

The proposed method, denoted as gBiCov, includes three steps. In the first step, Biologically Inspired Features (BIF) [4] are extracted. BIFs are based on the study of human visual system and have shown excellent performances on several computer vision tasks [5–7]. In particular, we use the S1 layer (Gabor filters) and C1 layer (MAX operator) of BIF. While the Gabor filters can improve robustness to illumination variations, the MAX operator increases the tolerance to scale changes and image shifts. In the second step, a covariance descriptor is used to compute the similarity of BIF features taken at neighboring scales. Covariance descriptors can capture shape, location and color information, and their performance have been shown to be better than other methods in many situations, as rotations and illumination changes are absorbed by the covariance matrix [1]. Furthermore, we argue that measuring the similarity of BIF at neighboring scales decreases the influence

☆ This paper has been recommended for acceptance by Shishir Shah.

* Corresponding author at: GREYC — CNRS UMR 6072, University of Caen Basse-Normandie, Caen, France. Tel.: +33(0)231 567 498.
E-mail addresses: bpma@ucas.ac.cn (B. Ma), nanosuyu@gmail.com (Y. Su), frederic.jurie@unicaen.fr (F. Jurie).
[1] Tel./Fax: +86 10 6967 1794.

http://dx.doi.org/10.1016/j.imavis.2014.04.002
0262-8856/© 2014 Elsevier B.V. All rights reserved.

of the background (see Section 3.5 for details). In the third step, BIF and covariance descriptors are combined into a single representation. Finally, we show that the performance of the proposed representation can be further enhanced by the use of metric leaning (we use the KISSME framework of [2]). Since the resulting representation is robust to illumination, scale and background changes, the performance for person re-identification and face verification can be significantly improved.

In addition to presenting an approach performing well on real datasets, one interesting contribution of the paper lies in the use of covariance descriptors in a novel way. In traditional covariance-based method, the similarity of two images can be obtained by comparing their covariance descriptors [8–10], which is a time-consuming process. In contrast, in the proposed approach, the similarities of covariance descriptors between consecutive bands of BIF features in the same image are measured. These similarities are then concatenated to produce image signatures, and the similarity between probe and gallery images is obtained by simply computing the *L2* distance between their signatures. It avoids the expensive computation of the similarity between covariance descriptors of probe image and each gallery image, which can be extremely time-consuming when the gallery is large.

The proposed method is experimentally validated on three public datasets for person re-identification: VIPeR, i-LIDS and ETHZ. They are among the most challenging ones, since all the above-described issues such as pose changes, viewpoint and lighting variations, and occlusions, are present. As an illustration of the performance, the matching rate at rank 1 (i.e., considering only the most similar image of the gallery) is of 31.11% on the VIPeR dataset (10% better than best previously published result). Knowing that the matching rate at low ranks is the most important criterion for real-life applications, this improvement is very significant. The proposed method is also validated on a face verification dataset, the Labeled Faces in the Wild (LFW) dataset, and compared to recently published state-of-the-art approaches.

The remaining of this paper is organized as follows: Section 2 reviews the related works on person's re-identification and face verification. Section 3 describes the proposed method in details and discusses its advantages. Experimental validations are given in Section 4. Finally, Section 5 concludes the paper.

2. Related work

Person/face re-identification – which is the task of associating the same person through different cameras or at different time – is a challenging problem as the association has to be done despite view point, illumination and pose changes. It has received a lot of attention in the recent literature, reflecting the interest for the important applications that can be addressed with these technologies.

More formally, the task of person re-identification can be defined as finding the correspondences between the images of a *probe set* representing a single person and the images of a *gallery set*. Depending on the number of available images per individual (size of the probe set), the scenarios can be defined as: (a) single-shot [11,12], if only one frame per individual is available both in the probe and gallery sets; and (b) multiple-shot [11,12], if multiple frames per individual are available both in the probe and gallery sets.

One of the key ingredient of face/person re-identification approaches lies in the encoding of images into visual signatures that can be compared more efficiently than raw pixel intensities. The recent literature abounds with such image descriptors for person re-identification. They can be based on (i) color, widely used since the color of clothing constitutes simple but efficient visual signatures, usually encoded within histograms of RGB or HSV values [11], (ii) shape, e.g. using HOG based signature [13,14], (iii) texture, often represented by Gabor filters [15,10,16], differential filters [15,16], Haar-like representations [17] and co-occurrence matrices [14], (iv) interest points, e.g. SURF [18] and SIFT [19,20] and (v) image regions [13,11,12]. Besides these generic representations, there are some more specialized representations, e.g. Epitomic Analysis [21], Spin Images [22,23], Bag-of-Word-based description [20], Implicit Shape Model (ISM) [19] and Panoramic Map [24]. Since different elementary features capture different and complementary aspects of the image, better performance is obtained by combining several signatures. We point this out in the following section.

Among these methods, those based on representing humans by a collection of parts have attracted more and more attention. Part-based methods split the human body into different parts and encode each part separately. In [11,12], the authors use Maximally Stable Color Regions (MSCR) to build a representation of human body. MSCR consists in grouping pixels having similar colors into maximally stable regions during a clustering process. The regions are subsequently described by their area, centroid, second moment matrices and average colors. Interestingly, covariance descriptors have also been widely used for representing regions [8–10]; the pixels within a region are represented by a feature vector consisting of intensity, texture and shape statistics, while the regions are represented by the covariance matrix of these feature vectors.

As mentioned above, since different elementary features (color, shape, texture, etc.) capture different and complementary characteristics of the image, they are often combined to give a richer signature. For example, [15] combines 8 color features with 21 texture filters (Gabor and differential filters). [11,12] combine MSCR descriptors with weighted color histograms, achieving state-of-the-art results on several wildly-used person re-identification datasets. The Covariance descriptor can be generalized to any type of images (three channel color images, infrared images, etc.), and can be used to combined different descriptors. For example, in [10], Gabor features and Local Binary Patterns (LBP) are combined with a covariance descriptor which handles, to some extent, illumination and viewpoint changes as well as non-rigid deformations.

Different representations usually require different similarity measures. For example, representations based on histograms can be compared with the Bhattacharyya distance [11,21,12] or the Earth Movers Distance (EMD). When the representation includes two or more different features/channels, the similarity is usually computed by combining their respective similarities (late fusion) e.g. using a linear combination [21,12,11]. Regarding the methods based on the covariance descriptor, even if the similarity of two regions is computed by estimating the distance between two covariance matrices in a pairwise manner, the similarity of human body described themselves by a set of covariance matrices has to combine several region similarities. This combination can be based, for example, on the mean covariance distance between corresponding regions [25] or by the minimum difference between corresponding body regions [10]. To capture the correlation between body parts, [17] uses spatial pyramid matching and designs a new similarity measure between human signatures. In [9], the authors argue that the covariance matrices lie in a Riemannian manifold, and combine the efficiency of the mean Riemannian covariance descriptor with the spatial information carried out by a dense grid structure. In [8], the authors propose a multi-scale covariance descriptor which describes an image quadrant through its corresponding sub-tree.

In order to improve the performance of these representations in the context of person re-identification, several papers have proposed to use discriminative classifiers on top of them: these classifiers can be based on Adaboost [16,17], Rank SVM [15], Partial Least Squares (PLS), multi-feature learning [26] or multiple instance learning [27,28].

Different from these classifiers, metric learning can provide a way to adapt a similarity function to the given task. Simple but efficient are the metric learning methods based on Mahalanobis-like distance functions. Approaches such as Large Margin Nearest Neighbors (LMNN) [29], Information Theoretic Metric Learning (ITML) [30], Logistic Discriminant Metric Learning (LDML) [31], Pairwise Constrained Component Analysis (PCCA) [32] and Keep It Simple and Straightforward Metric Learning (KISSME) [2] have been used successfully in the context of face verification and person re-identification.