Image and Vision Computing 44 (2015) 44–58

Contents lists available at ScienceDirect

Image and Vision Computing

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

Boosting Fisher vector based scoring functions for person re-identification☆

Stefano Messelodi, Carla Maria Modena *

Fondazione Bruno Kessler, Via Sommarive 18, I-38123 Trento, Italy

ARTICLE INFO

Article history:
Received 30 August 2014
Received in revised form 17 August 2015
Accepted 16 September 2015
Available online 19 October 2015

Editor: Nicu Sebe

Keywords:
Person re-identification
Fisher vector
Adaptive boosting
Likelihood ratio
Similarity ranking

ABSTRACT

In recent years, much effort has been put into the development of novel algorithms to solve the person re-identification problem. The goal is to match a given person's image against a gallery of people. In this paper, we propose a single-shot supervised method to compute a scoring function that, when applied to a pair of images, provides a score expressing the likelihood that they depict the same individual. The method is characterized by: (i) the usage of a set of local image descriptors based on Fisher vectors, (ii) the training of a pool of scoring functions based on the local descriptors, and (iii) the construction of a strong scoring function by means of an adaptive boosting procedure. The method has been tested on four data-sets and results have been compared with state-of-the-art methods clearly showing superior performance.

1. Introduction

The problem to automatically retrieve a selected person from video streams is of fundamental importance to video analysis. Applications vary from searching for suspicious individuals in a network of surveillance cameras, to maintaining person identity from one camera to the other for behavior analysis. Several factors contribute making the problem very hard, in fact a person's appearance can vary greatly through scenes due to changes in viewpoints, illumination conditions, pose and orientation, or to the possible usage of different acquisition devices. Other disturbing factors are the presence of shadows, occlusions, or individuals in the scene with similar appearance.

Person re-identification consists of matching observations of individuals across disjoint camera views. In very recent years, this problem has received a considerable attention, and various surveys and reviews are available, pointing out different aspects of this challenging topic [1–6]. For this reason, we direct the reader to these papers for a detailed discussion on the challenges posed by the problem, and for an overview of state-of-the-art methods along with their performance on publicly available data-sets.

Broadly speaking, in order to address the problem, people have to be detected in videos and be represented by descriptors which aim to capture their visual appearance. The descriptors are then used to compare different individuals and to determine the correspondence among them. Re-identification methods proposed in literature usually avoid to consider the detection phase and assume to work with images whose content is restricted to a bounding box around the person. They differ on the descriptor construction that can refer to a single view of the person (single-shot methods) or to multiple views obtained by briefly tracking (tracklet) the person's movements (multi-shot methods), and on the comparison of descriptors, which can be direct (unsupervised) or based on similarity measures learned using a set of labeled samples (supervised).

Although re-identification can be regarded as a binary classification problem over pairs of people descriptors, it is clear that a binary answer (same person or not) becomes harder as the gallery size increases. Thus the evaluation of a re-identification system is accomplished by regarding re-identification as a ranking problem rather than a classification one: the algorithms return a sorted list of candidates and the best performance is obtained if the correct correspondence is in most cases at, or close to, the first position of the returned list.

Using a standard taxonomy, the method proposed in this paper is a supervised single-shot recognition method. The major novelty of the proposed method, named *BFiVe*, consists of combining the power of Fisher vector descriptors with the ability of boosting procedures to select the most appropriate local descriptors to build a strong scoring function.

Starting from low-level features computed at pixel level in regions obtained from a coarse to fine image subdivision, an image is initially represented by a family of local descriptors based on Fisher vectors

☆ This paper has been recommended for acceptance by Jakob Verbeek.

* Corresponding author. Tel.: +39 0461314508; fax: +39 0461314501.
E-mail addresses: messelod@fbk.eu (S. Messelodi), modena@fbk.eu (C.M. Modena).

http://dx.doi.org/10.1016/j.imavis.2015.09.008
0262-8856/

that are then dimensionally reduced to an optimal size using Principal Component Analysis. In the training phase, a pool of weak scoring functions is generated using the local descriptors. Finally, the construction of a strong scoring function by means of an adaptive boosting procedure is performed using a minimum error procedure on the weak learners. The error is computed by analyzing the position of the right match in the ranked output of the weak scoring function. In this way, the regions that better contribute to collocate the right match in the very first positions weigh more in the global scoring function.

Previously published methods, to the best of our knowledge, aggregate local descriptors in order to build a single image descriptor, and learn a single metric to provide the final ranking. The novelty of *BFiVe* is that it learns a proper metric for each subimage, i.e. there are as many learnt rankers as the regions the image is divided into. A second learning step is performed using a ranking-based boosting approach, which combines local rankers to establish the final ranking function.

The proposed method has been experimentally validated on four challenging data-sets: VIPeR, 3DPeS, PRID 2011 and i-LIDS-119. The obtained figures clearly outperform the best previously published results on all of them.

The rest of the paper is organized as follows. Section 2 briefly describes the state-of-the-art methods included in the supervised single-shot category. Section 3 presents synthetically the *BFiVe* method. Sections 4 and 5 explain the techniques we propose for the description of images and for the learning of the scoring functions, respectively, while Section 6 illustrates the on-line usage of the method. Section 7 presents the experimental validation of *BFiVe* including a comparison with the state-of-the-art, the methodology followed for parameters selection, and an analysis of the computational complexity. Section 8 analyzes several aspects of the proposed method, discussing its main features. Section 9 concludes the paper.

2. Related works

In this section, we review several works in recent literature that fall into the supervised, single-shot re-identification category. Methods in this class are characterized by specific features used to describe the images and by specific procedures that make use of a labeled data-set to learn a metric by enforcing small distances among data of the same class (images depicting the same person). The usage of common data-sets and evaluation protocols is mandatory for a direct and meaningful comparison of the method's performance.

In Ma et al. [7], the color image is firstly divided into large, fixed, non-overlapping rectangular regions and each pixel is described by simple feature vectors. The feature vectors of the pixels that fall in each region are encoded and aggregated into Fisher vectors, which are then concatenated and dimensionally reduced with Principle Component Analysis (PCA) to obtain the final signature of the image. Using Pairwise Constrained Component Analysis (PCCA) [8] a similarity metric, sLDFV, is learnt, i.e. a projection into a low-dimensional space where distances between pairs of signatures respect the desired matching constraints.

In Pedagadi et al. [9] images are described by very high dimensional features based on local color histograms and their statistics in HUV and HSV color spaces, separately. The feature vectors can be exploited in an efficient way using a dimensionality reduction approach that combines unsupervised and supervised techniques, namely PCA and local Fisher discriminative analysis (LF). The Euclidean metric is then used for the comparison. In the same paper, a novel statistic is introduced to characterize re-identification performance, called Proportion of Uncertainty Removed (PUR) index. It is invariant to test set size, and we use it to evaluate our method's performance.

In [10–12], the main focus is on metric learning rather than on feature selection specific to the re-identification task. In [10], a support vector machine framework is proposed to obtain an optimized metric for nearest neighbor classification called large margin nearest neighbor with rejection (LMNN-R), i.e. the classifier returns no matches if all neighbors are beyond a certain distance. The signature is built by applying a PCA reduction to the concatenation of histograms of color channels (RGB and HSV) extracted from a grid of rectangular overlapping windows.

In [11], relaxed pairwise distance metric learning, RP-MeL, is used to address the problem of maximizing the probability that a pair of images depicting the same person has a smaller distance than a pair of different individuals. Once the metric has been learnt, only linear projections are necessary at search time, where a nearest neighbor classification is performed. The image descriptor is obtained by merging local color and texture features computed on overlapping rectangular regions, then reduced with PCA. In [12], a "keep it simple and straightforward metric" (KISSME) was introduced to learn a distance metric from equivalence constraints. The method is applied on a variety of challenging benchmarks including person re-identification across spatially disjoint cameras, using the same descriptors as [11].

The KISSME metric learning algorithm is also used by Ma et al. in [13] to improve the discriminative ability of their proposed descriptors: To gain robustness to illumination variations, scale and shifts, the image representation relies on the combination of biologically inspired features [14] based on covariance descriptors. This approach, named kBiCov, that focuses on feature selection and on metric learning, produces one of the best results currently present in literature.

In the re-identification task, one of the main problems is the different responses of the camera due to sensor variability, illumination changes, and aiming angle. Hirzer et al. [15], address the 'different camera properties problem' by learning a transition function from one camera to another. This is realized by learning a Mahalanobis metric using pairs of images coming from different cameras. The mean color values from small image regions are combined with a histogram of Local Binary Patterns to represent an image, and then pairwise sample differences are learnt for re-identification, considering correspondent people and also impostors that invade the perimeter of a given pair (Efficient Impostor-based Metric Learning, EIMeL).

In [28] the authors formulate a relative distance comparison (RDC) model, to maximize the likelihood of a pair of true matches that have a relatively smaller distance compared to an incorrect matching pair in a soft discriminant manner. The descriptors are obtained by dividing the images into six horizontal stripes. For each stripe, color features and texture features are extracted, giving rise to an image descriptor vector in a 2784 dimensional feature space. The model is based on logistic functions which are learnt with an iterative optimization algorithm on subsets of the data and then combined in an ensemble way to obtain the final RDC.

Li and Wang [16] propose locally aligned feature transforms, LAFT, for matching people across camera views that can have complex cross-view variations. Images to be matched are softly assigned to different local experts of a gating network according to the similarity of cross-view transforms, then they are projected to a common feature space and matched with a locally learnt discriminative metric.

An original framework is proposed in [17], where a reference set of images is used to generate reference-based descriptors for probe and gallery people. The starting signatures are built from color and texture features following the approach in [11]. In the training phase, a reference set of image pairs is used to learn a subspace in which the data of the same subjects from different cameras are maximally correlated using Regularized Canonical Correlation Analysis (RCCA). The so-called reference descriptors (RDs) of probe and, respectively, gallery images are then obtained by projecting the original feature vectors into the RCCA subspace using the two learnt matrices. Re-identification is performed by comparing the RDs of the probes and the RDs of the gallery images. In this way, a direct comparison of probes and gallery images is avoided.

3. Outline of the *BFiVe* method

In this section, we provide an overview of the proposed re-identification method, which is outlined in Fig. 1. As labeled data-sets are crucial for developing supervised methods, we briefly explain