Computer Vision and Image Understanding 000 (2016) 1–15

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Structured learning of metric ensembles with application to person re-identification

Sakrapee Paisitkriangkrai[a], Lin Wu[a,b], Chunhua Shen[a,b,*], Anton van den Hengel[a,b]

[a] School of Computer Science, The University of Adelaide, Adelaide, SA 5005, Australia
[b] Australian Centre for Robotic Vision, Adelaide, Australia

ARTICLE INFO

Article history:
Received 5 January 2016
Revised 4 October 2016
Accepted 19 October 2016
Available online xxx

Keywords:
Person re-identification
Learning to rank
Metric ensembles
Structured learning

ABSTRACT

Matching individuals across non-overlapping camera networks, known as person re-identification, is a fundamentally challenging problem due to the large visual appearance changes caused by variations of viewpoints, lighting, and occlusion. Approaches in literature can be categorized into two streams: The first stream is to develop reliable features against realistic conditions by combining several visual features in a pre-defined way; the second stream is to learn a metric from training data to ensure strong inter-class differences and intra-class similarities. However, seeking an optimal combination of visual features which is generic yet adaptive to different benchmarks is an unsolved problem, and metric learning models easily get over-fitted due to the scarcity of training data in person re-identification. In this paper, we propose two effective structured learning based approaches which explore the adaptive effects of visual features in recognizing persons in different benchmark data sets. Our framework is built on the basis of multiple low-level visual features with an optimal ensemble of their metrics. We formulate two optimization algorithms, $\text{CMC}^{\text{triplet}}$ and CMC^{top}, which directly optimize evaluation measures commonly used in person re-identification, also known as the Cumulative Matching Characteristic (CMC) curve. The more standard $\text{CMC}^{\text{triplet}}$ formulation works on the triplet information by maximizing the relative distance between a matched pair and a mismatched pair in each triplet unit. The CMC^{top} formulation, modeled on a structured learning of maximizing the correct identification among top candidates, is demonstrated to be more beneficial to person re-identification by directly optimizing an objective closer to the actual testing criteria. The combination of these factors leads to a person re-identification system which outperforms most existing algorithms. More importantly, we advance state-of-the-art results by improving the rank-1 recognition rates from 40% to 61% on the iLIDS benchmark, 16% to 22% on the PRID2011 benchmark, 43% to 50% on the VIPeR benchmark, 34% to 55% on the CUHK01 benchmark and 21% to 68% on the CUHK03 benchmark.

1. Introduction

The task of person re-identification (re-id) is to match pedestrian images observed from different and disjoint camera views. Despite extensive research efforts in re-id (Gheissari et al., 2006; Li et al., 2013; Pedagadi et al., 2013; Roth et al., 2014; Xiong et al., 2014; Zhao et al., 2013a, 2013b, 2014; Zheng et al., 2011), the problem itself is still a very challenging task due to (a) large variation in visual appearance (person's appearance often undergoes large variations across different camera views); (b) significant changes in human poses at the time the image was captured; (c) large amount of illumination changes, background clutter and occlusions; (d) relatively low resolution and the different placement of the cameras. Moreover, the problem becomes increasingly difficult when there are high variations in pose, camera viewpoints, and illumination, etc.

To address these challenges, existing research has concentrated on the development of sophisticated and robust features to describe visual appearance under significant changes. Most of them use appearance-based features that are viewpoint invariant such as color and texture descriptors (Farenzena et al., 2010; Gheissari et al., 2006; Gray and Tao, 2008; Liu et al., 2012; Wang et al., 2007). However, the system that relies heavily on one specific type of visual cues, e.g., color, texture or shape, would not be practical and powerful enough to discriminate individuals with similar visual appearance. Some studies have tried to address the above problem by seeking a combination of robust and distinctive feature representation of person's appearance, ranging from color

* Corresponding author.
E-mail address: chhshen@gmail.com (C. Shen).

http://dx.doi.org/10.1016/j.cviu.2016.10.015
1077-3142/

histogram (Gray and Tao, 2008), spatial co-occurrence representation (Wang et al., 2007), LBP (Xiong et al., 2014), to color SIFT (Zhao et al., 2013b). The basic idea of exploiting multiple visual features is to build an ensemble of metrics (distance functions), in which each distance function is learned using a single feature and the final distance is calculated from a weighted sum of these distance functions (Farenzena et al., 2010; Xiong et al., 2014; Zhao et al., 2013b). These works often pre-define distance weights, which need to be re-estimated beforehand for different data sets. However, such a pre-defined principle has some drawbacks.

- Different real-world re-id scenarios can have very different characteristics, e.g., variation in view angle, lighting and occlusion. Simply combining multiple distance functions using predetermined weights may be undesirable as highly discriminative features in one environment might become irrelevant in another environment.
- The effectiveness of distance learning heavily relies on the quality of the feature selected, and such selection requires some domain knowledge and expertise.
- Given that certain features are determined to be more reliable than others under a certain condition, applying a standard distance measure for each individual match is undesirable as it treats all features equally without differentiation on features.

In these ends, it necessarily demands a principled approach that is able to automatically select and learn weights for diverse metrics, meanwhile generic yet adaptive to different scenarios.

Person re-identification problem can also be cast as a learning problem in which either metrics or discriminative models are learned (Chopra et al., 2005; Davis et al., 2007; Kedem et al., 2012; Kostinger et al., 2012; Li et al., 2013; Mignon and Jurie, 2012; Weinberger et al., 2006; Weinberger and Saul, 2008; Wu et al., 2011; Xiong et al., 2014; Zheng et al., 2013), which typically learn a distance measure by minimizing intra-class distance and maximizing inter-class distance simultaneously. Thereby, they require sufficient labeled training data from each class [1] and most of them also require new training data when camera settings change. Nonetheless, in a person re-id benchmark, available training data are relatively scarce, and thus inherently undersampled for building a representative class distribution. This intrinsic characteristic of person re-id problem makes metric learning pipelines easily overfitted and unable to be applicable in small image sets.

To combat above difficulties simultaneously, in this paper, we introduce two structured learning based approaches to person re-id by learning weights of distance functions for low-level features. The first approach, $\text{CMC}^{\text{triplet}}$, optimizes the relative distance using the triplet units, each of which contains three person images, i.e., one person with a matched reference and a mismatched reference. Treating these triplet units as input, we formulate a large margin framework with triplet loss where the relative distance between the matched pair and the mismatched pair tends to be maximized. An illustration of $\text{CMC}^{\text{triplet}}$ is shown in Fig. 1. This triplet based model is more natural for person re-id mainly because the intra-class and inter-class variation may vary significantly for different classes, making it inappropriate to require the distance between a matched/mismatched pair to fall within an absolute range (Zheng et al., 2013). Also, training images in person re-id are relatively scarce, whereas the triplet-based training model is to make comparison between any two data points rather than comparison between any data distribution boundaries or among clusters of data. This thus alleviates the over-fitting problem in person re-id given undersampled data. The second approach, CMC^{top}, is developed to maximize the average rank-k recognition rate, in which k is chosen to be small, e.g., $k < 10$. Setting the value of k to be small is crucial for many real-world applications since most surveillance operators typically inspect only the first ten or twenty items retrieved. Thus, we directly optimize the testing performance measure commonly used in CMC curve, i.e., the recognition rate at rank-k by using structured learning.

The main contributions of this paper are three-fold:

- We propose two principled approaches, $\text{CMC}^{\text{triplet}}$ and CMC^{top}, to build an ensemble of person re-id metrics. The standard approach $\text{CMC}^{\text{triplet}}$ is developed based on triplet information, which is more tolerant to large intra and inter-class variations, and alleviates the over-fitting problem. The second approach of CMC^{top} directly optimizes an objective closer to the testing criteria by maximizing the correctness among top k matches using structured learning, which is empirically demonstrated to be more beneficial to improving recognition rates.
- We perform feature quantification by exploring the effects of diverse feature descriptors in recognizing persons in different benchmarks. An ensemble of metrics is formulated into a late fusion paradigm where a set of weights corresponding to visual features are automatically learned. This late fusion scheme is empirically studied to be superior to various early fusions on visual features.
- Extensive experiments are carried out to demonstrate that by building an ensemble of person re-id metrics learned from different visual features, notable improvement on rank-1 recognition rate can be obtained. In addition, our ensemble approaches are highly flexible and can be combined with linear and non-linear metrics. For non-linear base metrics, we extend our approaches to be tractable and suitable to large-scale benchmark data sets by approximating the kernel learning.

2. Related work

Many person re-id approaches are proposed to seek robust and discriminative features such that they can be used to describe the appearance of the same individual across different camera views under various changes and conditions (Bazzani et al., 2012; Cheng et al., 2011; Farenzena et al., 2010; Gheissari et al., 2006; Gray and Tao, 2008; Li et al., 2014; Wang et al., 2007; Zhao et al., 2013b, 2014). For instance, Bazzani et al. represent a person by a global mean color histogram and recurrent local patterns through epitomic analysis (Bazzani et al., 2012). Farenzena et al. propose the symmetry-driven accumulation of local features (SDALF) which exploits both symmetry and asymmetry, and represents each part of a person by a weighted color histogram, maximally stable color regions and texture information (Farenzena et al., 2010). Gray and Tao introduce an ensemble of local features which combines three color channels with 19 texture channels (Gray and Tao, 2008). Schwartz and Davis propose a discriminative appearance based model using partial least squares where multiple visual features: texture, gradient and color features are combined (Schwartz and Davis, 2009). Zhao et al. combine dcolorSIFT with unsupervised salience learning to improve its discriminative power in person re-id (Zhao et al., 2013b) (eSDC), and further integrate both salience matching and patch matching into a unified RankSVM framework (SalMatch (Zhao et al., 2013a)). They also propose mid-level filters (MidLevel) for person re-identification by exploring the partial area under the ROC curve (pAUC) score (Zhao et al., 2014). Lisanti et al. (2015) leverage low-level feature descriptors to approximate the appearance variants in order to discriminate individuals by using sparse linear reconstruction model (ISR).

Another line to approach the problem of matching people across cameras is to essentially formalize person re-id as a

[1] Images of each person in a training set form a class.

Please cite this article as: S. Paisitkriangkrai et al., Structured learning of metric ensembles with application to person re-identification, Computer Vision and Image Understanding (2016), http://dx.doi.org/10.1016/j.cviu.2016.10.015