Learning to rank in person re-identification with metric ensembles

Sakrapee Paisitkriangkrai, Chunhua Shen, Anton ven den Hengel
The University of Adelaide, Australia; and Australian Centre for Robotic Vision

The task of person re-identification (re-id) is to match pedestrian images observed from multiple cameras. It has recently gained popularity in research community due to its several important applications in video surveillance. An automated re-id system could save a lot of human labour in exhaustively searching for a person of interest from a large amount of video sequences.

We propose an effective structured learning based approach to the problem of person re-identification which outperforms the current state-of-the-art on most benchmark data sets evaluated. Our framework is built on the basis of multiple low-level hand-crafted and high-level visual features. We then formulate two optimization algorithms, which directly optimize evaluation measures commonly used in person re-identification, also known as the Cumulative Matching Characteristic (CMC) curve. Our new approach is practical to many real-world surveillance applications as the re-identification performance can be concentrated in the range of most practical importance. 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 on person re-identification by improving the rank-1 recognition rates from 40% to 50% on the iLIDS benchmark, 16% to 18% on the PRID2011 benchmark, 43% to 46% on the VIPeR benchmark, 34% to 53% on the CUHK01 benchmark and 21% to 62% on the CUHK03 benchmark.

The main contributions of this paper are twofold: 1) We propose principled approaches to build an ensemble of person re-id metrics. The first approach aims at maximizing the relative distance between images of different individuals and images of the same individual such that the CMC curve approaches one with a minimal number of returned candidates. The second approach directly optimizes the probability that any of these top k matches are correct using structured learning. Our ensemble-based approaches are highly flexible and can be combined with linear and non-linear metrics. 2) 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. Experimental results show that our approach achieves the state-of-the-art performance on most person re-id benchmark data sets evaluated. In addition, our ensemble approach is complementary to any existing distance learning methods.

Notation Bold lower-case letters, *e.g.*, w, denote column vectors and bold upper-case letters, *e.g.*, P, denote matrices. We assume that the provided training data is for the task of single-shot person re-identification, *i.e.*, there exist only two images of the same person – one image taken from camera view A and another image taken from camera view B. We represent a set of training samples by $\left\{(x_i, x_i^+)\right\}_{i=1}^m$ where $x_i \in \mathbb{R}^D$ represents a training example from one camera (*i.e.*, camera view A), and x_i^+ is the corresponding image of the same person from a different camera (*i.e.*, camera view B). Here m is the number of persons in the training data. From the given training data, we can generate a set of triplets for each sample x_i as $\left\{(x_i, x_i^+, x_{i,j}^-)\right\}$ for $i = 1, \cdots, m$ and $i \neq j$. Here we introduce $x_{i,j}^- \in \mathcal{X}_i^-$ where $\mathcal{X}_i^-$ denotes a subset of images of persons with a different identity to x_i from camera view B. We also assume that there exist a set of distance functions $d_t(\cdot,\cdot)$ which calculate the distance between two given inputs. Our goal is to learn a weighted distance function: $d(\cdot,\cdot) = \sum_{t=1}^T w_t d_t(\cdot,\cdot)$, such that the distance between x_i (taken from camera view A) and x_i^+ (taken from camera view B) is smaller than the distance between x_i and any $x_{i,j}^-$ (taken from camera view B). The better the distance function, the faster the cumulative matching characteristic (CMC) curve approaches one.

Relative distance based approach (CMC$^{\text{triplet}}$) In order to minimize k such that the rank-k recognition rate is equal to 100%, we consider learning an ensemble of distance functions based on relative comparison of triplets. Given a set of triplets $\left\{(x_i, x_i^+, x_{i,j}^-)\right\}_{i,j}$, in which x_i is taken from camera view A and $\{x_i^+, x_{i,j}^-\}$ are taken from camera view B, the basic idea is to learn a distance function such that images of the same individual are closer than any images of different individuals, *i.e.*, x_i is closer to x_i^+ than any $x_{i,j}^-$. For a triplet $\left\{(x_i, x_i^+, x_{i,j}^-)\right\}_{i,j}$, the following condition must hold $d(x_i, x_{i,j}^-) > d(x_i, x_i^+), \forall j, i \neq j$. Following the large margin framework with the hinge loss, the condition $d(x_i, x_{i,j}^-) \geq 1 + d(x_i, x_i^+)$ should be satisfied. This condition means that the distance between two images of different individuals should be larger by at least a unit than the distance between two images of the same individual. Since the above condition cannot be satisfied by all triplets, we introduce a slack variable to enable soft margin. By generalizing the above idea to the entire training set, the primal problem that we want to optimize can be written as,

$$
\begin{aligned}
\min_{w,\xi} \; & \frac{1}{2}\|w\|_2^2 + \nu \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j=1}^{m-1} \xi_{ij} \\
\text{s.t. } & w^\top (d_j^- - d_i^+) \geq 1 - \xi_{ij}, \forall \{i,j\}, i \neq j; \\
& w \geq 0; \; \xi \geq 0.
\end{aligned} \tag{1}
$$

Here $\nu > 0$ is the regularization parameter and $d_j^- = [d_1(x_i, x_{i,j}^-), \cdots, d_t(x_i, x_{i,j}^-)]$, $d_i^+ = [d_1(x_i, x_i^+), \cdots, d_t(x_i, x_i^+)]$ and $\{d_1(\cdot,\cdot), \cdots, d_t(\cdot,\cdot)\}$ represent a set of base metrics.

Results The algorithm proposed in [3] achieves state-of-the-art results on iLIDS and 3DPeS data sets (40.3% and 54.2% recognition rate at rank-1, respectively). Our approach outperforms [3] on the iLIDS (50.3%) and achieve a comparable result on 3DPeS (53.3%). Zhao *et al*. propose mid-level filters for person re-identification [4], which achieve state-of-the-art results on the VIPeR and CUHK01 data sets (43.39% and 34.30% recognition rate at rank-1, respectively). Our approach outperforms [4] by achieving a recognition rate of 45.89% and 53.40% on the VIPeR and CUHK01 data sets, respectively. Table 1 compares our results with other state-of-the-art methods on other person re-identification data sets.

Data set	# Individuals		Prev. best	Ours
	train	test		
iLIDS	59	60	40.3% [3]	**50.3**%
3DPeS	96	96	**54.2**% [3]	53.3%
PRID2011	100	100	16.0% [2]	**17.9**%
VIPeR	316	316	43.4% [4]	**45.9**%
CUHK01	486	485	34.3% [4]	**53.4**%
CUHK03	1260	100	20.7% [1]	**62.1**%

Table 1: Rank-1 recognition rate of existing best reported results and our results. The best result is shown in boldface.

[1] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2014.

[2] P. M. Roth, M. Hirzer, M. Köstinger, C. Beleznai, and H. Bischof. Mahalanobis distance learning for person re-identification. In *Person Re-Identification*, pages 247–267. Springer, 2014.

[3] F. Xiong, M. Gou, O. Camps, and M. Sznaier. Person re-identification using kernel-based metric learning methods. In *Proc. Eur. Conf. Comp. Vis.*, 2014.

[4] R. Zhao, W. Ouyang, and X. Wang. Learning mid-level filters for person re-identfiation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2014.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.