Weighted Local Metric Learning for Person Re-identification

Xinqian Gu[1] and Yongxin Ge[1,2(✉)]

[1] School of Software Engineering, Chongqing University,
Chongqing 400044, China
xinqiangu@gmail.com, yongxinge@cqu.edu.cn
[2] Key Laboratory of Dependable Service Computing
in Cyber Physical Society Ministry of Education, Chongqing 400044, China

Abstract. Person re-identification aims to match individual across non-overlapping camera networks. In this paper, we propose a weighted local metric learning (WLML) method for person re-identification. Motivated by the fact that local metric learning has been exploited to handle the data which varies locally, we break down the pedestrian images into several local sub-regions, among which different metric functions are learned. Then we use structured method to learn the weight for each metric function and the final distance is calculated from a weighted sum of these metric functions. Our approach can also combine the local metric functions with global metric functions to exploit their complementary strengths. Moreover it is possible to integrate multiple visual features to further promote the recognition rate. Experiments on two challenging datasets validate the effectiveness of our proposed method.

Keywords: Person re-identification · Local metric learning · Structured learning

1 Introduction

Person re-identification aims to recognize people who have been observed from different disjoint cameras, which play a crucial role in video surveillance and visual information retrieval. Due to the large changes in appearances caused by variations in viewing angle, illumination, background clutter and occlusions, person re-identification is still a very challenging problem.

Recently proposed approaches which improve the person re-identification performance [1–9] can be mainly divided into two categories: (1) extracting robust descriptors to deal with the changes in person appearances; (2) designing discriminative metric functions to measure the similarity of person images. For the first category, several effective descriptors have been proposed, such as covariance descriptor [7], and local maximal occurrence (LOMO) [2]. For the second category, a discriminative metric is learned, under which the distance between the same persons and the distance between different persons are increased and decreased, respectively. Among them, Liong et al. [6] model and regulate the eigen-spectrums of covariance matrices in a parametric manner. Pedagadi et al. [9] learned the distance function by maximizing the between-class scatter

Z. You et al. (Eds.): CCBR 2016, LNCS 9967, pp. 686–694, 2016.
DOI: 10.1007/978-3-319-46654-5_75

matrix while minimizing the within-class scatter matrix using the Fisher discriminant objective. However, most metric learning methods only focus on the global measurement, neglecting the local discriminative power.

Chen et al. [1] proposed a similarity learning method with spatial constraints, which partitioned the person images into several sub-regions, and measured the similarity for each region, then, employed linear superposition to combine them together. He also collaborated the local measurements with global measurements and incorporated multiple visual cues to improve the performance. Experimental results show the effectiveness of this method. Considered different sub-regions and different features make different contribution to the final similarity, we proposed weighted local metric learning (WLML), which combines the similarity measurement of different sub-regions and different features by weighted summation. Then, we learn the weight by structured learning [3], instead of pre-defining. Experimental results on two widely used datasets demonstrate the efficacy of our proposed method.

2 Our Approach

In this section, we first propose weighted local metric learning and exploit structured method to learn the weight. Subsequently, we introduce metric method applied in our experiment.

2.1 Weighted Local Metric Learning

In this section, we introduce the overall similarity function first and then formulate the learning problem specifically.

Similarity Function. Given a pedestrian image, we divide it into R non-overlapping horizontal stripe regions and extract F types of color and texture features from each stripe region. After that, we can obtain the f-th descriptor $x^{r,f}$ for the r-th stripe region, where $r \in \{1,\ldots,R\}$ and $f \in \{1,\ldots,F\}$.

To measure the similarity between image descriptors $x_a, x_b \in \mathbb{R}^{d\times 1}$, where x_a and x_b are respectively from camera view A and camera view B, we employ the existing distance function $d(x_a, x_b)$ (will be introduced in Sect. 2.2) to calculate the distance between x_a and x_b. Correspondingly we define the similarity function for the f-th descriptor of the r-th stripe region as:

$$d^{r,f}(x_a, x_b) = d\left(x_a^{r,f}, x_b^{r,f}\right) \tag{1}$$

Considering one specific type of visual feature may not be powerful enough to discriminate individuals with similar visual appearance, we employ a weighted summation method to combine these features. The similarity function for r-th stripe region can be written as:

$$d^{r}(x_a, x_b) = \sum_{f=1}^{F} w_{r,f} d^{r,f}(x_a, x_b) \tag{2}$$

where $w_{r,f} \geq 0$ is the weight of $d^{r,f}(x_a, x_b)$. Since different regions make different contributions to the local similarity score. For all R regions, the local similarity function is represented as:

$$d^{local}(x_a, x_b) = \sum_{r=1}^{R} w_r d^{r}(x_a, x_b) \tag{3}$$

where $w_r \geq 0$ is the weight of $d^{r}(x_a, x_b)$. Since the horizontal stripe regions are non-overlapping, and the local descriptors can't describe the matching of large patterns across the stripes. We combine local similarity with global similarity, and the overall similarity function can be written as:

$$d(x_a, x_b) = d^{local}(x_a, x_b) + w_G d^{global}(x_a, x_b) \tag{4}$$

where $w_G \geq 0$ is the weight of $d^{global}(x_a, x_b)$, and the global similarity function $d^{global}(x_a, x_b)$ is defined as:

$$d^{global}(x_a, x_b) = \sum_{f=1}^{F} w_{G,f} d^{G,f}(x_a, x_b) \tag{5}$$

where $d^{G,f}(x_a, x_b) = d(x_a^{G,f}, x_b^{G.f})$ and $x_a^{G,f}$, $x_b^{G,f}$ are the f-th type global visual feature for image a and image b. Then, expansion formula of Eq. (4) can be written as:

$$d(x_a, x_b) = \sum_{r=1}^{R} \sum_{f=1}^{F} w_r w_{r,f} d^{r,f}(x_a, x_b) + \sum_{f=1}^{F} w_G w_{G.f} d^{G,f}(x_a, x_b) \tag{6}$$

Equation (6) can be simplified as follows by replacing $w_r w_{r,f}$ and $w_G w_{G,f}$ with $w'_{r,f}$ and $w'_{G,f}$:

$$\begin{aligned} d(x_a, x_b) =& \sum_{r=1}^{R} \sum_{f=1}^{F} w'_{r,f} d^{r,f}(x_a, x_b) + \sum_{f=1}^{F} w'_{G,f} d^{G,f}(x_a, x_b) \\ =& w^{T} \cdot d \end{aligned} \tag{7}$$

where

$$w = [w'_{1,1}, \ldots w'_{R,F}, w'_{G,1}, \ldots w'_{G,F}]$$

and

$$d = [d^{1,1}(x_a, x_b), \ldots, d^{R,F}(x_a, x_b), d^{G,1}(x_a, x_b), \ldots, d^{G,F}(x_a, x_b)].$$

In next section, we will introduce the learning of the weighted similarity function Eq. (7) which makes $d(x_a, x_b)$ smaller when x_a and x_b are from the same individual.

Structured Learning of Similarity Model. In the training process, we randomly selected one image per individual form the gallery set and the remaining images are used to form the probe set. We denote the training set as $\chi = \{x_p^q\}_{p=a,b}^{q=1,\ldots,N_p}$, where x_p^q denotes the q-th pedestrian form camera view p. We refer to probe set as camera view a, and gallery set as camera view b. We also denote the ground-truth ranking structure as $y^* = \left\{y_{ij}^*\right\}$, and $y_{ij}^* = 1$ if x_a^i and x_b^j are the same person; otherwise, $y_{ij}^* = 0$. Then training process can be formulated as the following structured learning problem:

$$\begin{aligned} &\min_{w,\xi} \frac{1}{2}\|w\|_2^2 + C\xi \\ &s.t.\ w^T(\psi(\chi, y^*) - \psi(\chi, y)) \geq \Delta(y^*, y) - \xi, \forall y \in \mathcal{Y}, \xi \geq 0 \end{aligned} \tag{8}$$

where w is the weight vector in Eq. (7), $y \in \mathcal{Y}$ denote any arbitrary predicted ranking structure, $\|\cdot\|_2$ denotes the l_2-norm of a vector, and $C > 0$ is the regularization parameter. We define the feature map $\psi(\chi, y)$ as:

$$\psi(\chi, y) = \frac{1}{N_a} \sum_{i=1}^{N_a} \sum_{k \in \chi_i^+} \sum_{j \in \chi_i^-} (1 - y_{ij}) \frac{d(x_a^i, x_b^j) - d(x_a^i, x_b^k)}{|\chi_i^+| \cdot |\chi_i^-|} \tag{9}$$

where χ_i^- denotes irrelevant individuals set of x_a^i, and χ_i^+ denotes the relevant individual set of x_a^i correspondingly. Since we use single-shot training, $|\chi_i^+| = 1$ and $|\chi_i^-| = (N_b - 1)$, Eq. (9) can be simplified as:

$$\psi(\chi, y) = \frac{1}{N_a(N_b - 1)} \sum_{i=1}^{N_a} \sum_{j \in \chi_i^-} (1 - y_{ij})(d(x_a^i, x_b^j) - d(x_a^i, x_b^k)), k \in \chi_i^+ \tag{10}$$

The goal of constraints is to enforce the distance between irrelevant individuals and relevant individuals of the ground-truth ranking structure to be the largest among any arbitrary ranking structures. Following the large margin framework, we define the loss function as:

$$\Delta(y^*, y) = \frac{1}{N_a(N_b - 1)} \sum_{i=1}^{N_a} \sum_{j \in \chi_i^-} (y_{ij} - y_{ij}^*) \tag{11}$$

which denotes the mean loss incurred by predicting ranking structures instead of the ground-truth ranking structure. Note that other convex loss functions can also be applied.

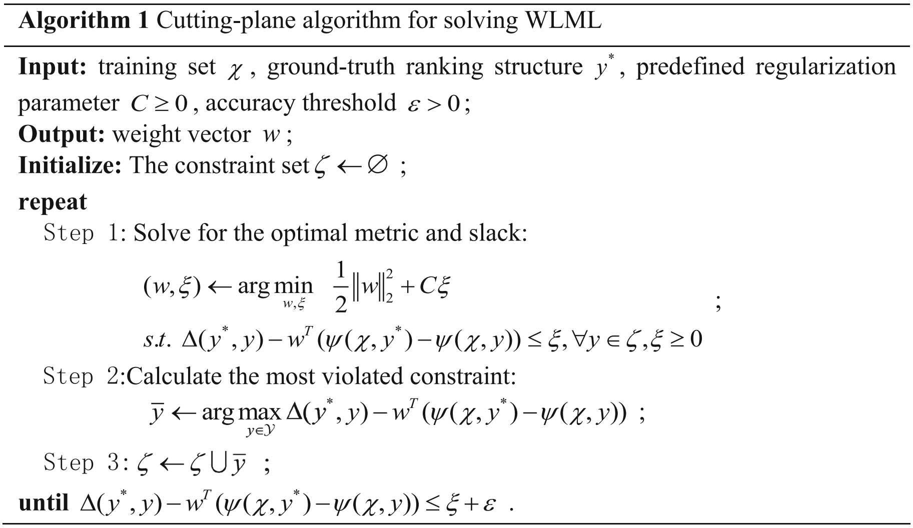
Optimization. In principle we can solve the structured learning using cutting-plane algorithm [12]. The basic idea of cutting-plane algorithm is that it is sufficient to obtain a ε-approximate solution of optimization problem by using a small subset of all constraints. We list the algorithm steps in Algorithm 1. It begins with a null constraint set. At each iteration, we solve the optimization problem to find a suitable w over current constraint set. Based on the w, we can find the most violated ranking structure $\bar{y}$ and add it to constraint set. The cutting-plane algorithm repeats the above steps until it converges.

The calculation of the most violated constraint (Algorithm 1, step 2) can be written as:

$$
\begin{aligned}
\bar{y} &= \arg\max_{y \in \mathcal{Y}} \Delta(y^*, y) - w^T(\psi(\chi, y^*) - \psi(\chi, y)) \\
&= \arg\max_{y \in \mathcal{Y}} \frac{1}{N_a(N_b - 1)} \sum_{i=1}^{N_a} \sum_{j \in \chi_i^-} y_{ij} - \frac{1}{N_a(N_b - 1)} \sum_{i=1}^{N_a} \sum_{j \in \chi_i^-} y_{ij} w^T (d(x_a^i, x_b^j) - d(x_a^i, x_b^k)) \\
&= \arg\max_{y \in \mathcal{Y}} \frac{1}{N_a(N_b - 1)} \sum_{i=1}^{N_a} \sum_{j \in \chi_i^-} (y_{ij}(1 - w^T d_{ij}))
\end{aligned} \tag{12}
$$

where $d_{ij} = d\left(x_a^i, x_b^j\right) - d\left(x_a^i, x_b^k\right)$. Obviously, $\bar{y}$ can be written as:

$$
\bar{y}_{ij} = \begin{cases} 1, if\ w^T d_{ij} \leq 1 \\ 0, \text{otherwise.} \end{cases} \tag{13}
$$



Algorithm 1 Cutting-plane algorithm for solving WLML

```
Input: training set $\chi$, ground-truth ranking structure $y^*$, predefined regularization
parameter $C \geq 0$, accuracy threshold $\varepsilon > 0$;
Output: weight vector $w$;
Initialize: The constraint set $\zeta \leftarrow \varnothing$ ;
repeat
  Step 1: Solve for the optimal metric and slack:
        $(w, \xi) \leftarrow \arg\min_{w,\xi} \frac{1}{2}\|w\|_2^2 + C\xi$ ;
        $s.t.\ \Delta(y^*, y) - w^T(\psi(\chi, y^*) - \psi(\chi, y)) \leq \xi, \forall y \in \zeta, \xi \geq 0$
  Step 2:Calculate the most violated constraint:
        $\bar{y} \leftarrow \arg\max_{y \in \mathcal{Y}} \Delta(y^*, y) - w^T(\psi(\chi, y^*) - \psi(\chi, y))$ ;
  Step 3: $\zeta \leftarrow \zeta \cup \bar{y}$ ;
until $\Delta(y^*, y) - w^T(\psi(\chi, y^*) - \psi(\chi, y)) \leq \xi + \varepsilon$ .
```

2.2 Metric Method

Metric learning can be divided into linear [6, 9] and non-linear methods [2, 4]. As to linear method, a projection matrix M is sought, so that the distance between x_a and x_b can be denoted as $d(x_a, x_b) = (x_a - x_b)^T M(x_a - x_b)$, which will be small if x_a and x_b are from the same person and large otherwise. By kernelization, linear method can be extended to non-linear method easily. The distance of non-linear method can be written as $d(x_a, x_b) = (\phi(x_a) - \phi(x_b))^T M(\phi(x_a) - \phi(x_b))$, where $\phi(x)$ is mapping from feature to kernel space.

In our experiments, we used kernel Local Fisher Discriminant Analysis (kLFDA) [4], which is a non-linear extension to previously proposed LFDA [9]. Unlike LFDA, kLFDA learns projection matrix M in the kernel space $\phi(x)$. Note that there are more metric learning methods can be used in our framework and we just choose the kLFDA for our experiments.

3 Experiments

We evaluate the proposed WLML method on two widely used person re-identification datasets, namely the VIPeR [10] and i-LIDS [11] databases. The following describe the details of our experiments and results.

3.1 Feature Extraction

We divide a pedestrian image into 4 non-overlapping horizontal stripe regions. For each stripe region, we extract 4 types of basic features multi-HS, multi-RGB, SILTP [13] and dense SIFT [5], which describe different aspects of person images. Among them, multi-HS and multi-RGB are 8×8 and $8 \times 8 \times 8$ joint histograms respectively. SILTP and dense SIFT are texture descriptors extracted at RGB and LAB channel, respectively. Then each histogram feature is normalized with the l_2-norm. Finally, two visual cues F_1, F_2 are organized as multi-HS/SILTP and multi-RGB/SIFT. For global descriptor, we also extract HS and RGB concatenated histograms with each channel having 32 bins and concatenate them with HOG [14] histogram.

3.2 Settings

In our experiments, we used the single-shot training and testing where one image per person was randomly selected to form the gallery set and the remaining images were used to form the probe set. For our WLML method, we set the regularization parameter C as $10^{2.8}$ and accuracy threshold ε as 10^{-6} for all experiments. For the non-linear metric method kLFDA [4], we set the regularization parameter for class scatter matrix as 0.01 and apply the RBF-χ^2 kernel for all features. In this experiment, we set the value of σ^2 to be the same as the first quantile of all distances [4]. To evaluate our proposed method, the average cumulative matching curve (CMC) where a match is found at the top-*n* ranks by repeating the experiments 10 times.

3.3 Evaluation on the VIPeR Dataset

The VIPeR dataset [10] is one of the most popular datasets for person re-identification. It consists of 632 persons captured from two cameras with a viewpoint change of 90° and varying illumination conditions. We randomly select 316 persons to form the training set and the remaining 316 persons are used to form test set.

Table 1 show the matching results compared to 6 representative methods, which includes kLFDA [4], LOMO+XQDA [2], ME [3] and SCSP [1]. While kLFDA, LOMO+XQDA, ME and SCSP are state-of-the-art techniques that presented promising performance in person re-identification. We see that our proposed WLML method outperforms most existing methods. It achieved 50.9 % rank-1 accuracy in VIPeR dataset.

Table 1. Matching rates (%) of different metric learning methods on the VIPeR dataset

Rank	1	5	10	20
WLML	50.9	77.5	88.6	96.2
SCSP [1]	**53.5**	**82.6**	**91.5**	**96.6**
ME [3]	44.9	76.3	88.2	94.9
LOMO+XQDA [2]	40.0	68.0	80.5	91.1
kLFDA [4]	32.3	65.8	79.7	90.9

To validate the effectiveness of our proposed method, we compare our proposed method with other methods using multi-feature in Table 2. Since we haven't the source code of SCSP, we compare our proposed method with ME and the linear superposition of the local metric in our method (SLML). The difference between SCSP and SLML is that they used different local metric and kernel. To fairly compare these methods, all these approaches use the same features and testing/training set. We see that our proposed WLML method outperforms SLML and ME methods with as high as approximately 4 % and 3 % rank-1 accuracy, which validates the effectiveness of our proposed weighted method and local metric, respectively.

Table 3. Matching rates (%) of different metric learning methods on the i-LIDS dataset

Rank	1	5	10	20
WLML	**61.4**	**76.0**	**82.3**	**93.8**
ME [3]	50.3	–	–	–
kLFDA [4]	38.0	65.1	77.4	89.2
LFDA [9]	33.8	57.4	69.7	82.8

Table 2. Matching rates (%) of 3 methods using the same features and testing/training set

Rank	1	5	10	20
WLML	**50.9**	**77.5**	**88.6**	**96.2**
SLML	46.5	75.6	88.6	96.2
ME	47.5	76.9	87.0	94.9

3.4 Evaluation on the i-LIDS Dataset

The i-LIDS dataset consists of 476 images from 119 persons captured from eight disjoint cameras [11]. The number of images for each individual varies from 2 to 8. The dataset presents severe occlusions caused by busy crowd and luggage. We randomly select 59 persons to form the training set and the remaining 60 persons are used to form test set. Table 3 shows the matching rates compared to state-of-the-art method. We see that our proposed WLML method outperforms all other methods.

4 Conclusion

In this paper, we have proposed a weighted local metric learning (WLML) for person re-identification. The proposed method learns the weight of local metric and combines the similarity measurement of different sub-regions and different visual cues by weighted summation. Experimental results on two widely used re-identification datasets have shown the effectiveness of the proposed method.

References

1. Chen, D., Yuan, Z., Chen, B., Zheng, N.: Similarity learning with spatial constraints for person re-identification. In: CVPR, pp. 1268–1277 (2016)
2. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: CVPR, pp. 2197–2206 (2015)
3. Paisitkriangkrai, S., Shen, C., Hengel, A.V.D.: Learning to rank in person re-identification with metric ensembles. In: CVPR, pp. 1846–1855 (2015)
4. Xiong, F., Gou, M., Camps, O., Sznaier, M.: Person re-identification using kernel-based metric learning methods. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part VII. LNCS, vol. 8695, pp. 1–16. Springer, Heidelberg (2014)
5. Zhao, R., Ouyang, W., Wang, X.: Unsupervised salience learning for person re-identification. In: CVPR, pp. 3586–3593 (2013)
6. Liong, V.E., Lu, J., Ge, Y.: Regularized Bayesian metric learning for person re-identification. In: Agapito, L., Bronstein, M.M., Rother, C. (eds.) ECCV 2014 Workshops. LNCS, vol. 8927, pp. 209–224. Springer, Heidelberg (2015)
7. Ma, B., Su, Y., Jurie, F.: Covariance descriptor based on bio-inspired features for person re-identification and face verification. Image Vis. Comput. **32**(6–7), 379–390 (2014)
8. Ma, B., Su, Y., Jurie, F.: Local descriptors encoded by fisher vectors for person re-identification. In: Fusiello, A., Murino, V., Cucchiara, R. (eds.) ECCV 2012 Ws/Demos, Part I. LNCS, vol. 7583, pp. 413–422. Springer, Heidelberg (2012)

9. Pedagadi, S., Orwell, J., Velastin, S., Boghossian, B.: Local fisher discriminant analysis for pedestrian re-identification. In: CVPR, pp. 3318–3325 (2013)
10. Gray, D., Brennan, S., Tao, H.: Evaluating appearance models for recognition, reacquisition, and tracking. In: PETS (2007)
11. Zheng, W.S., Gong, S., Xiang, T.: Associating groups of people. In: BMVC, pp. 1–11 (2009)
12. Joachims, T., Finley, T., Yu, C.-N.J.: Cutting-plane training of structural svms. Mach. Learn. **77**, 27–59 (2009)
13. Liao, S., Zhao, G., Kellokumpu, V., Pietikäinen, M., Li, S.Z.: Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes. In: CVPR (2010)
14. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR (2005)