2014 22nd International Conference on Pattern Recognition

Person Re-Identification via Discriminative Accumulation of Local Features

Tetsu Matsukawa
The University of Tokyo, JAPAN
Email: te2@iis.u-tokyo.ac.jp

Takahiro Okabe
Kyushu Institute of Technology, JAPAN
Email: okabe@ai.kyutech.ac.jp

Yoichi Sato
The University of Tokyo, JAPAN
Email: ysato@iis.u-tokyo.ac.jp

Abstract—**Metric learning to learn a good distance metric for distinguishing different people while being insensitive to intra-person variations is widely applied to person re-identification. In previous works, local histograms are densely sampled to extract spatially localized information of each person image. The extracted local histograms are then concatenated into one vector that is used as an input of metric learning. However, the dimensionality of such a concatenated vector often becomes large while the number of training samples is limited. This leads to an over fitting problem. In this work, we argue that such a problem of over-fitting comes from that it is each local histogram dimension (*e.g.* color brightness bin) in the same position is treated separately to examine which part of the image is more discriminative. To solve this problem, we propose a method that analyzes discriminative image positions shared by different local histogram dimensions. A common weight map shared by different dimensions and a distance metric which emphasizes discriminative dimensions in the local histogram are jointly learned with a unified discriminative criterion. Our experiments using four different public datasets confirmed the effectiveness of the proposed method.**

I. Introduction

Re-identification of a person seen in dis-joint camera views is one of the important research topics in visual surveillance. By assuming targets do no change their clothes in short time, the problem of person re-identification can be posed as that of matching person images of a whole body with those captured in different camera views. However, it is highly challenging to perform accurate matching due to large intra-personal variations of appearance caused by viewpoint/illumination changes, partial occlusion, and background variations.

In general, there are two main steps to perform person re-identification. In the first step, a feature descriptor is extracted from each person image in both probe and gallery sets. In the second step, the feature descriptor of a probe image is compared to that of each gallery image with some distance metric.

A human pose often changes drastically among different observations, and it is known that, in such cases, histogram-based feature descriptors are more appropriate because the integral nature of histogram-based feature descriptors makes them less sensitive to human pose change. Some works [1], [2] construct robust histogram-based feature descriptors against pose change by accumulating features with a Gaussian-like weight map in which where image positions likely to be a foreground region have a high weight value. It is known spatially localized information of color/texture is effective for distinguishing different persons [4]. Therefore local histograms are densely extracted on horizontal or grid cells and these are often used for the input of metric learning. Supervised learning is widely applied to learn a distance metric for a feature descriptor [3], [4], [5], [6], [7], [8].

Some positions in input images such as corners of an image tend to be background. The effects of local histograms extracted from such positions should be weakened in distance calculation. On the other hands, some positions such as upper parts of a human body may have highly discriminative information. The local histograms extracted from such positions should be emphasized. In other words, the discriminativeness of each of densely sampled local histograms should depend on its extracted position. In the previous metric learning methods, a concatenated vector of local histograms extracted from all sampled positions is used as a feature vector. This means the discriminative positions are separately analyzed in each of local histogram dimensions (*e.g.* color brightness bins).

The number of available samples for training is generally limited since it is hard to correct all possible situations for each camera pairs. In such a case, the gap of sample distribution between training and test sets becomes large. For example, even though the training set does not contain the person who wears clothes that have a specific color, such person may appear in the test set. Therefore, this leads to an over-fitting problem. Namely, local histograms on discriminative positions would not be emphasized unless the persons in the training set have similar local histograms to that of the test samples on the same position. Although such over-fitting can be alleviated by using coarsely sampled histograms, the discriminative information that might commonly exist in both the training and the test set will also be reduced.

To solve this problem, we propose a method that analyzes discriminative image positions shared by different local histogram dimensions (*e.g.* color brightness bins). More specifically, we construct a feature vector as a weighted summed form of local histograms with a common weight map for each dimension. By sharing the weight map for each of local histogram dimensions, we can emphasize the discriminative positions even if similar color/texture is not observed in a training set. In this way, the problem of over-fitting in metric learning is alleviated without sacrificing the descriptiveness of the local histograms.

Some people may have highly distinctive positions on a part of whole body (*e.g.* upper body), and some people may have them on another parts (*e.g.* lower body). In these ways, the discriminative positions are different depending on each person. To effectively utilize such diverse discriminative positions, we construct multiple weight maps. Note that each of them is commonly applied to each of local histogram dimensions, and thus the robustness against over-fitting is maintained.

The flow of the proposed method is shown in Fig. 1. For each weight map, we construct weighted local histograms

1051-4651/14 $31.00 © 2014 IEEE
DOI 10.1109/ICPR.2014.681

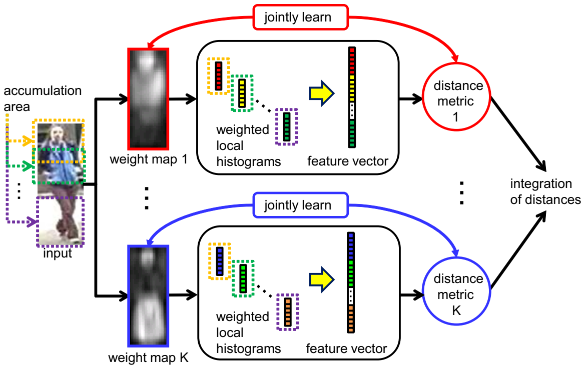

Fig. 1. Flow of the proposed method. Our method jointly learns each pair of a weight map for local histogram accumulations and a distance metric.

in pre-determined accumulation areas. The weighted local histograms are concatenated into a single feature vector. To emphasize discriminative dimension, a distance metric is applied to the feature vector. Using training data with correct person labels, we jointly learn an optimal weight map and a distance metric with a discriminative criterion that maximizes the local discriminative information on sample space. Because the important feature dimension calculated from different discriminative parts of a human body may be different, we learn different metric for the feature vector constructed from each weight map. Finally, distances calculated from multiple pairs of a weight map and a distance metric are integrated and used for re-identification.

As far as we know, the closest method to ours is *Fisher Weight Map (FWM)* [9] proposed for facial expression recognition. The major advantages of the proposed method to *FWM* are following two points. 1) *FWM* accumulates local features into a global histogram. If *FWM* is applied to person re-identification without any modification, a summed histogram within a person image which includes different parts (*e.g.*, head, upper body, and bottom body) becomes an input for metric learning. In contrast, we accumulate local features into several local accumulation areas. In general, a different part in each person image has different color and texture. Therefore, our method treats local features in different parts more appropriately. 2) In *FWM*, a distance metric is learned after weight maps were obtained. In contrast, we jointly learn them. Thus, the proposed method utilizes more discriminative information hidden in local features to obtain weight maps.

II. Weighted Accumulation of Local Features

The proposed method is associated with multiple pairs of a weight map and a distance metric. In each pair, we construct weighted local histograms using the weight map and then arrange them into a single feature vector (Fig. 1). In this section, we explain this process.

We firstly define grid cells in each input image, and a weight value for the histogram accumulation is set to each of the grid cells. Let $\boldsymbol{f}_r \in \mathbb{R}^{d'}$ be a vector of a local histogram (such as a color histogram) of r-th grid cell and w_r be a weight value of the r-th grid cell (Fig. 2(a)). We then define local accumulation areas to calculate a weighted histogram within them (areas shown by dotted lines in Fig. 2(a)). Each of the accumulation areas is composed of a set of nearby grid cells and it can overlap with different areas. Let $\mathcal{A}_i$ be the set of grid cells included in i-th accumulation area. Using them, the weighted local histogram of the i-th accumulation area $\boldsymbol{h}_i \in \mathbb{R}^{d'}$ is defined as $\boldsymbol{h}_i = \sum_{r\in\mathcal{A}_i} \boldsymbol{f}_r w_r$.

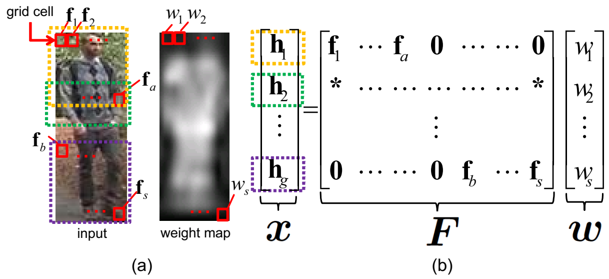
grid cell
$\mathbf{f}_1\ \mathbf{f}_2$
$\mathbf{f}_a$
$\mathbf{f}_b$
$\mathbf{f}_s$
input
$w_1\ w_2$
w_s
weight map
$\mathbf{h}_1$
$\mathbf{h}_2$
$\mathbf{h}_g$
$=$
$\mathbf{f}_1 \cdots \mathbf{f}_a\ \mathbf{0} \cdots\cdots \mathbf{0}$
$* \cdots\cdots\cdots\cdots *$
$\mathbf{0} \cdots\cdots \mathbf{0}\ \mathbf{f}_b \cdots \mathbf{f}_s$
w_1
w_2
w_s
$\boldsymbol{x}$
$\boldsymbol{F}$
$\boldsymbol{w}$
(a)
(b)

Fig. 2. Linear form of accumulated local histograms with a weight map.

We construct weighted local histograms calculated from different accumulation areas and then concatenate them as a single feature vector. Let g be the total number of accumulation areas. Then the concatenated feature vector is defined by $\boldsymbol{x} = [\boldsymbol{h}_1^T, ..., \boldsymbol{h}_g^T]^T \in \mathbb{R}^d$, where $d = d'g$. Note that when we set the number of accumulation areas as $g = 1$ and we include the all grid cells in $\mathcal{A}_1$, the feature vector $\boldsymbol{x}$ becomes the global weighted histogram used in *FWM* [9].

Now we define a weight map vector $\boldsymbol{w} = [w_1, ..., w_s]^T \in \mathbb{R}^s$, where s is the number of grid cells. Then we can rewrite the feature vector $\boldsymbol{x}$ as a following linear form:

$$\boldsymbol{x} = \boldsymbol{F}\boldsymbol{w}, \tag{1}$$

where $\boldsymbol{F} \in \mathbb{R}^{d\times s}$ is a feature matrix where the feature vectors of a local histogram of grid cells are arranged so that for each i-th set of d' rows of the matrix, the weight w_r acts only to the feature vector $\boldsymbol{f}_r$ extracted from the r-th grid cell that consist the i-th accumulation area (Fig. 2(b)).

III. Discriminative Learning

In this section, we explain how to jointly learn the multiple pairs of a weight map and a distance metric. To find these pairs in a discriminative way, we optimize them by the *Average Neighborhood Margin Maximization (ANMM)* criterion [10]. The *ANMM* criterion explores the discriminative information locally on sample space. It is known that the local discriminative information is more effective for recognition tasks than the global one in the *Fisher* criterion.

A. Optimization problem

Given N training feature matrices $\{\boldsymbol{F}_i\}_{i=1}^N$ extracted from training samples with correct person labels, the objective of the optimization is to seek optimal pairs of a weight map and a distance metric. The optimal pairs are sought such that for each sample, distances calculated between neighboring samples of the same person become as small as possible, while simultaneously distances calculated between neighboring samples of different persons become as large as possible.

At first, we see the k-th pair of a weight map and a distance metric. Using a weight map vector, all of feature matrices are transformed into the feature vectors as $\{\boldsymbol{x}_i = \boldsymbol{F}_i\boldsymbol{w}_k\}_{i=1}^N$. For any pair of the feature vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ of the i-th and the j-th images, a *squared Mahalanobis-like distance* is given by

$$D_{M_k}^2(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i - \boldsymbol{x}_j)^T \boldsymbol{M}_k (\boldsymbol{x}_i - \boldsymbol{x}_j), \tag{2}$$

where $\boldsymbol{M}_k \in \mathbb{R}^{d\times d}$ is a matrix which defines the distance metric of the k-th pair. In general, $\boldsymbol{M}_k$ is a valid metric if and only if it is a symmetric and *Positive Semi-Definite*

(PSD) matrix. In such a case, there exists a linear non-square projection matrix $\boldsymbol{L}_k \in \mathbb{R}^{d\times q}$ where $q \leq d$ that satisfies $\boldsymbol{M}_k = \boldsymbol{L}_k\boldsymbol{L}_k^T$. Using $\boldsymbol{L}_k$, Eq.(2) can be computed as

$$D^2_{w_k,L_k}(\boldsymbol{F}_i,\boldsymbol{F}_j) = \|\boldsymbol{L}_k^T\boldsymbol{F}_i\boldsymbol{w}_k - \boldsymbol{L}_k^T\boldsymbol{F}_j\boldsymbol{w}_k\|_2^2. \quad (3)$$

Next, we consider K pairs of a weight vector and a projection matrix corresponding to a distance metric; $\boldsymbol{\Omega} = \{\boldsymbol{w}_k,\boldsymbol{L}_k\}_{k=1}^K$. In this case, we integrate multiple *squared Mahalanobis-like distances* defined by each pair into one distance as a following equation:

$$D^2_\Omega(\boldsymbol{F}_i,\boldsymbol{F}_j) = \sum_{k=1}^K D^2_{w_k,L_k}(\boldsymbol{F}_i,\boldsymbol{F}_j). \quad (4)$$

To introduce the optimization problem, we firstly define two neighborhood sets for k-th pair of a weight map and a distance metric. Let $\mathcal{N}^S_{i,k}$ be the set of κ_s nearest neighborhood sample which are in the same person as $\boldsymbol{F}_i$ and $\mathcal{N}^D_{i,k}$ be the set of κ_d nearest neighborhood data which are in the different persons as $\boldsymbol{F}_i$[1].

Then the *average neighborhood margin* $\gamma_{i,k}$ of sample i for k-th pair is defined as

$$\gamma_{i,k} = \sum_{j\in\mathcal{N}^D_{i,k}} \frac{D^2_{w_k,L_k}(\boldsymbol{F}_i,\boldsymbol{F}_j)}{|\mathcal{N}^D_{i,k}|} - \sum_{j\in\mathcal{N}^S_{i,k}} \frac{D^2_{w_k,L_k}(\boldsymbol{F}_i,\boldsymbol{F}_j)}{|\mathcal{N}^S_{i,,k}|}. \quad (5)$$

For tractability, we optimize a sum of the *average neighborhood margin* per each pair of a weight map vector and a distance metric instead of the margin calculated by the final distance in Eq.(4). Summing up all training samples and all pairs, the total *average neighborhood margin* is given as

$$\gamma = \sum_{i=1}^N\sum_{k=1}^K \gamma_{i,k} = \sum_{k=1}^K J(\boldsymbol{w}_k,\boldsymbol{L}_k), \quad (6)$$

where $J(\boldsymbol{w}_k,\boldsymbol{L}_k) \triangleq \sum_{i=1}^N \gamma_{i,k}$.

Giving some constraints, the optimization problem results in

$$\begin{aligned} \max_{\boldsymbol{W},\boldsymbol{L}} \quad & \sum_{k=1}^K J(\boldsymbol{w}_k,\boldsymbol{L}_k) \\ s.t. \quad & \boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{I}, \\ & \boldsymbol{L}_k^T\boldsymbol{L}_k = \boldsymbol{I},\ k=1,...,K, \end{aligned} \quad (7)$$

where $\boldsymbol{W} = [\boldsymbol{w}_1,..,\boldsymbol{w}_K]$ and $\boldsymbol{I}$ is an identity matrix. The first constraint is introduced so that the weight map vectors are uncorrelated each other, and the second constraint ensures each of the matrix $\boldsymbol{M}_k = \boldsymbol{L}_k^T\boldsymbol{L}_k$ becomes a valid metric, *i.e.* symmetric and *PSD* matrix. Also both constraints prevent the objective value becomes unbounded.

B. Greedy solution

It is difficult to get the global solution of the optimization problem in Eq.(7). In order to get an approximation of the global optimum solution to the above problem, we propose to solve it by a greedy algorithm. The algorithm is shown in Algorithm.1. We separate the problem Eq.(7) into K steps and sequentially solve them. In each k-th step, we optimize the k-th pair of a weight map vector and a projection matrix $\{\boldsymbol{w}_k,\boldsymbol{L}_k\}$ such that

$$\begin{aligned} \max_{\boldsymbol{w}_k,\boldsymbol{L}_k} \quad & J(\boldsymbol{w}_k,\boldsymbol{L}_k) \\ s.t. \quad & \boldsymbol{w}_k^T\boldsymbol{w}_k = 1, \\ & \boldsymbol{w}_k^T\boldsymbol{w}_m = 0,\ m=1,...,k-1, \\ & \boldsymbol{L}_k^T\boldsymbol{L}_k = \boldsymbol{I}. \end{aligned} \quad (8)$$

The second constraint $\boldsymbol{w}_k^T\boldsymbol{w}_m = 0, m=1,...,k-1$ ensures to uncorrelate the k-th weight map vector $\boldsymbol{w}_k$ to already learned weight map vectors in previous steps $\{\boldsymbol{w}_m\}_{m=1}^{k-1}$. Assume that each row of each i-th feature matrix $\boldsymbol{F}_i$ is uncorrelated to $\{\boldsymbol{w}_m\}_{m=1}^{k-1}$, *i.e.* $\boldsymbol{F}_i\boldsymbol{w}_m = \boldsymbol{0}, m = 1,...,k-1$, where $\boldsymbol{0}$ is a vector whose all components are zero. In such a case, the second constraint can be omitted (see Appendix). This uncorrelation can be achieved by

$$\boldsymbol{F}'_i \leftarrow \boldsymbol{F}_i - \sum_{m=1}^{k-1}\big\{(\boldsymbol{1}_d\otimes\boldsymbol{w}_m^T)\odot(\boldsymbol{1}_s^T\otimes\boldsymbol{F}_i\boldsymbol{w}_m)\big\}. \quad (9)$$

where $\otimes$ is a kroneker product and $\odot$ is an element-wise product of matrices and $\boldsymbol{1}_d \in \mathbb{R}^d$ and $\boldsymbol{1}_s \in \mathbb{R}^s$ are column vectors whose all components are 1.

Now, the optimization problem in Eq.(8) is simplified to

$$\begin{aligned} \max_{\boldsymbol{w}_k,\boldsymbol{L}_k} \quad & J'(\boldsymbol{w}_k,\boldsymbol{L}_k) \\ s.t. \quad & \boldsymbol{w}_k^T\boldsymbol{w}_k = 1, \\ & \boldsymbol{L}_k^T\boldsymbol{L}_k = \boldsymbol{I}, \end{aligned} \quad (10)$$

where the J' is the objective value calculated using $\boldsymbol{F}$' instead of $\boldsymbol{F}$. The optimal solution of Eq.(10) can not be obtained at once. However, if we fixed one of $\{\boldsymbol{w}_k,\boldsymbol{L}_k\}$, we can maximize the object value in a closed form. Thus, we solve it by an alternative optimization. We start with constant weight map vectors $\boldsymbol{w}_k = \boldsymbol{1}_s/(\boldsymbol{1}_s^T\boldsymbol{1}_s)$ and repeat following processes $T_{\max}$ times.

On optimizing $\boldsymbol{L}_k$: By fixing $\boldsymbol{w}_k$ and transforming the feature matrices as $\{\boldsymbol{x}_i = \boldsymbol{F}'_i\boldsymbol{w}_k\}_{i=1}^N$, the optimization problem in Eq.(10) results in[2]

$$\boldsymbol{L}_k^* = \underset{\boldsymbol{L}_k}{\operatorname{argmax}}\, Tr\{\boldsymbol{L}_k^T(\boldsymbol{\Sigma}_D^{w_k} - \boldsymbol{\Sigma}_S^{w_k})\boldsymbol{L}_k\}\, s.t. \boldsymbol{L}_k^T\boldsymbol{L}_k = \boldsymbol{I}, \quad (11)$$

where

$$\begin{aligned} \boldsymbol{\Sigma}_D^{w_k} &= \sum_{i=1}^N\sum_{j\in\mathcal{N}^D_{i,k}} \frac{(\boldsymbol{x}_i-\boldsymbol{x}_j)(\boldsymbol{x}_i-\boldsymbol{x}_j)^T}{|\mathcal{N}^D_{i,k}|}, \\ \boldsymbol{\Sigma}_S^{w_k} &= \sum_{i=1}^N\sum_{j\in\mathcal{N}^S_{i,k}} \frac{(\boldsymbol{x}_i-\boldsymbol{x}_j)(\boldsymbol{x}_i-\boldsymbol{x}_j)^T}{|\mathcal{N}^S_{i,k}|}, \end{aligned} \quad (12)$$

and $Tr(\cdot)$ stands for a trace operator of matrices. Taking the Larange equation and setting its derivative of $\boldsymbol{L}_k$ to zero, we obtain the following eigen value problem:

$$(\boldsymbol{\Sigma}_D^{w_k} - \boldsymbol{\Sigma}_S^{w_k})\boldsymbol{L}_k = \lambda\boldsymbol{L}_k. \quad (13)$$

Thus, the optimal projection matrix is given as $\boldsymbol{L}_k^* = [\boldsymbol{l}_1^*,...,\boldsymbol{l}_q^*]$ where the column vectors are eigen vectors corresponding to the larges q eigen values in the above problem.

[1] Since $\boldsymbol{w}_k$ and $\boldsymbol{L}_k$ in the distance Eq.(3) are unknown, the neighborhood sets are initially searched by the *Frobenius norm* of matrix difference using $\boldsymbol{F}$. Then we update the neighborhood sets per k using the updated distance in each step of the optimization.

[2] It can be derived in a similar manner to *ANMM* of Wang *et al.* [10].

Algorithm 1 Algorithm for greedy solution

```
Require: A training feature matrices $\{\boldsymbol{F}_i\}_{i=1}^N$ ;
for  k = 1, 2, ..., $K$  do
    Initialize $\boldsymbol{w}_k = \mathbf{1}_s/(\mathbf{1}_s^T\mathbf{1}_s)$;
    for  t = 1, 2, ..., $T_{\max}$  do
        (a) Search neighborhood sets using $\{\boldsymbol{x}_i{=}\boldsymbol{F}_i\boldsymbol{w}_k\}_{i=1}^N$;
        (b) Compute $\Sigma_D^{w_k}$ and $\Sigma_S^{w_k}$ in Eq.(12);
        (c) Obtain $\boldsymbol{L}_k$ by the eigen value problem in Eq.(13);
        (d) Search neighborhood sets using $\{\boldsymbol{Y}_i{=}\boldsymbol{L}_k^T\boldsymbol{F}_i\}_{i=1}^N$;
        (e) Compute $\Sigma_D^{L_k}$ and $\Sigma_S^{L_k}$ in Eq.(15);
        (f) Obtain $\boldsymbol{w}_k$ by the eigen value problem in Eq.(16);
    end for
    Uncorrelate feature matrices as
    $\{\boldsymbol{F}_i\}_{i=1}^N \leftarrow \{\boldsymbol{F}_i - (\mathbf{1}_d \otimes \boldsymbol{w}_k^T) \odot (\mathbf{1}_s^T \otimes \boldsymbol{F}_i\boldsymbol{w}_k)\}_{i=1}^N$;
end for
Ensure: Pairs of a weight map and a projection $\{\boldsymbol{w}_k, \boldsymbol{L}_k\}_{k=1}^K$;
```

On optimizing $\boldsymbol{w}_k$: By fixing $\boldsymbol{L}_k$ and projecting the feature matrices as $\{\boldsymbol{Y}_i = \boldsymbol{L}_k^T\boldsymbol{F}'_i\}_{i=1}^N$, the optimization problem in Eq.(10) results in

$$\boldsymbol{w}_k^* = \operatorname*{argmax}_{\boldsymbol{w}_k} \boldsymbol{w}_k^T(\boldsymbol{\Sigma}_D^{L_k} - \boldsymbol{\Sigma}_S^{L_k})\boldsymbol{w}_k \quad s.t.\ \boldsymbol{w}_k^T\boldsymbol{w}_k = 1, \tag{14}$$

where

$$\boldsymbol{\Sigma}_D^{L_k} = \sum_{i=1}^N \sum_{j\in\mathcal{N}_{i,k}^D} \frac{(\boldsymbol{Y}_i - \boldsymbol{Y}_j)^T(\boldsymbol{Y}_i - \boldsymbol{Y}_j)}{|\mathcal{N}_{i,k}^D|},$$

$$\boldsymbol{\Sigma}_S^{L_k} = \sum_{i=1}^N \sum_{j\in\mathcal{N}_{i,k}^S} \frac{(\boldsymbol{Y}_i - \boldsymbol{Y}_j)^T(\boldsymbol{Y}_i - \boldsymbol{Y}_j)}{|\mathcal{N}_{i,k}^S|}. \tag{15}$$

Thus, the optimal $\boldsymbol{w}_k^*$ by fixing $\boldsymbol{L}_k$ is given by an eigen vector corresponding to the largest eigen value of the following eigen value problem:

$$(\boldsymbol{\Sigma}_D^{L_k} - \boldsymbol{\Sigma}_S^{L_k})\boldsymbol{w}_k = \gamma\boldsymbol{w}_k. \tag{16}$$

IV. Experiments

A. Setup

Datasets. We evaluated the proposed method on four commonly used public datasets; VIPeR [11], PRID2011 [12], GRID [13], and CAVIAR [14]. Following to other researches[3], we randomly divided the whole person images into a training and a test set. Every person in the training set has an image pair in different camera views. The test set consists of a probe and a gallery set. On PRID2011 and GRID, there exist persons that appear only in the gallery set, therefore the number of persons in this set is larger than that in the probe set. For every person in each probe/gallery set, five images are assigned for CAVIAR (*i.e.* multi-shot re-identification) and a single image is assigned for other datasets (*i.e.* single-shot re-identification). For the multi-shot re-identification, we calculate a distance for each possible pair of images between two persons, and the minimum distance of them [1] is used. The division setup of each dataset is shown in Table I. The whole evaluation was carried out 10 times by changing the random division of a training and a test set. We report the average results of them.

Features and accumulation areas. We used three types of visual features; HSV color histogram, color HOG, and a texture histogram. For the HSV color histogram, an 8-bin color histogram was calculated in each HSV component. Hence, the histogram results in 24 dimensional per each grid cell. For the color HOG, an 8-bin gradient orientation histogram was calculated in each YCbCr component. Hence, the orientation histogram results in 24 dimensional per each grid cell. For the texture histogram, 13 Schmid and 6 Gabor filters were applied to intensity components of an image and each filter response was discretized into 8-bin. Hence, the texture histogram results in 152 dimensional per each grid cell. We created the feature matrix defined in Sec.2 per each of the three visual features, and then concatenated them in a row and formed a single feature matrix. As an area configuration, we integrated 15×15 grid cells ($s = 225$) into 6×1 accumulation areas ($g = 6$). Both of them have overlap to different cells/areas.

TABLE I. SETUP OF EACH DATASET.

	traning set	test set		
Dataset	# of training persons	# of probe persons	# of gallery persons	shotnum/ person
VIPeR [11]	316	316	316	1
PRID2011 [12]	100	100	649	1
GRID [13]	125	125	900	1
CAVIAR [14]	36	36	36	5

Parameter settings. We set parameters of the proposed method as follows. The neighborhood size of different person was set to $\kappa_d = 10$. Because the number of samples in a single person is small in our experiments, we considered all samples in the same person as neighbors. The iteration number was set as $T_{\max} = 5$. The number of pairs of a weight map and a distance metric was set as $K = 10$. The dimensionality of the projection matrix was set as $q = 20$.

B. Performance comparison

Comparison of accumulation areas. To see the effects of the different configurations of the accumulation area, we varied the area configuration so that each configuration divides an input image into one of 1×1, 2×2, 4×4, 1×2, 1×4, 1×6, 2×1, 4×1, or 6×1 areas. For each configuration, each of the accumulation areas is equal size and overlaps with different areas. Fig. 3 shows the rank 1 and rank 10 recognition rates of the proposed method on different area configurations. The proposed method greatly improves the performance when more areas are used than the global histogram (1×1 area) that is used in *FWM* [9]. Among the area configurations, the horizontal strips (6×1, 4×1, and 2×1) work well on both VIPeR and PRID2011 datasets. From now on, we report the results of 6×1 areas for all datasets.

Comparison of the number of weight map and metric pairs. Fig. 6 shows the rank 1 and 10 recognition rates in different number of weight map and metric pairs. The performance is increasing as to increase the number of pairs and it is saturating around 10-20.

Comparison with different methods. First, we compared the proposed method with *FWM* [9], *ANMM* [10], *Local Fisher Discriminant Analysis (LF)* [8], and *KISS metric learning (KISSME)* [6] with common features to us. For *FWM*, we obtained weight maps by the *Fisher* criterion. Then the weighted histograms with learned weight maps were concatenated into one feature vector. The metric of the resulting feature vector was learned by *ANMM*. Thus, we refer it to *FWM + ANMM*. For other methods (*ANMM, LF,* and *KISSME*), we used a concatenation of local histograms on 15×15 overlapping grid cells as a feature vector. The same three types of

[3] For CAVIAR, several papers used this dataset in different settings. We followed the setting of Pedagadi *et al.* [8].

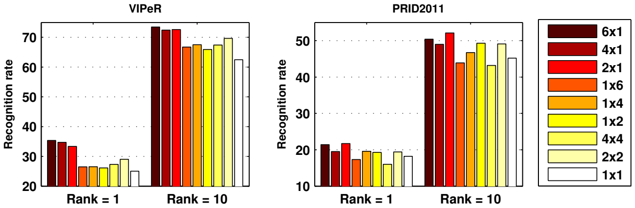

Fig. 3. Comparison of different configurations of the accumulation area.

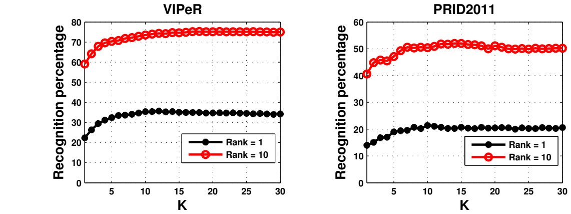

Fig. 4. Comparison of different number of weight map and metric pairs.

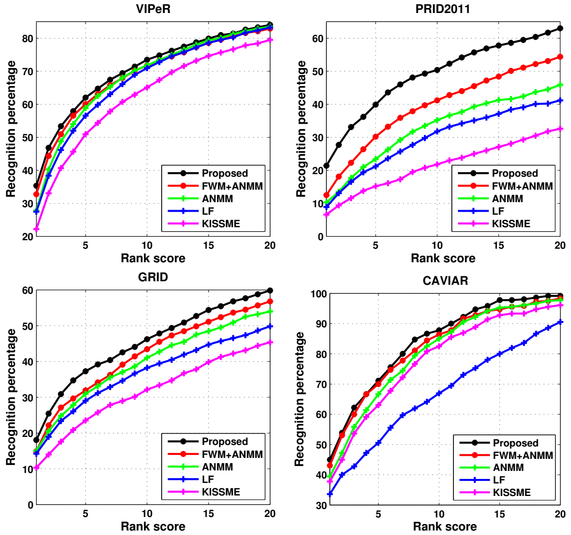

Fig. 5. Performance comparison on various datasets using CMC curves. Common features were used in all methods.

visual features to our method were used, and a feature vector was created for each feature type. PCA was applied to each of different types of feature vectors independently [8] and the compressed feature vectors were concatenated into one feature vector. The parameters of each method were tuned so that the best performance was obtained. Fig. 5 shows Cumulative Matching Characteristic (CMC) curves on each dataset. The results show that the proposed method outperforms all of the above base-line methods. Especially, the improvements on PRID2011 and GRID are high.

We then compared the results of the proposed method with previously reported them in other papers. The comparisons are shown in Table II. The rank-1 (rank-10) recognition rate of the proposed method is 35.35(73.48)% on VIPeR. This is a significantly better result compared with the results of the state-of-the art methods such as 27(69)% of *Relaxed Pairwise Learned Metric (RPLM)* [5]. Before now, *RPLM* reported the best re-identification result on PRID2011 and *Manifold Ranking (MR)* [15] reported it on GRID. The proposed method outperforms these methods with high margins. For CAVIAR, we followed the experimental settings of Pedagadi *et al.* [8][4]. The rank-1 recognition rate reported in their paper is 36.19% and the result of the proposed method is 44.16%. The proposed method achieves a better result than the previously reported result also on this dataset.

Comparison on different training sample sizes. On VIPeR, we compared our method on different sizes of training/test persons; 200 and 432 persons for a training and a test set, and 100 and 532 persons for a training and a test set. These divisions are commonly used in some previous works [1], [2], [5], [7]. Table III shows the comparison on these divisions. As to the number of training persons becomes low, the performances of all methods are decreased. In the table, *SDALF* [1] and *SCEALF* [2] are training free methods and the others are metric learning methods. Training free methods use several sophisticated features and matching methods such as graph matching, so these methods could produce high accuracies without training. It is known that metric learning methods are weak when the size of training set is small [2]. Hence, all of previous metric learning methods were worse than the training free methods when the number of training persons is 100. In all cases of person divisions, the performances of the proposed method are consistently better than other metric learning methods. In addition, the proposed method achieves better results than training free methods in latter ranks than 10 when the number of training persons is 100.

C. Analysis of learned position weights

We analyzed the learned position weights of the proposed method by visualizing weights corresponding to each grid cell. Because the standard metric learning methods also can be seen as weight learning methods of each dimension of a feature vector, as an example of them, the learned weights of *ANMM* was also visualized. The visualizations were done as follows. For *ANMM*, the k-th row of a projection matrix (*i.e.* projection vector) is considered as k-th position weights. Since PCA was applied to the input feature vector, the projection vectors were backed from PCA compression space to original feature space by the production of PCA bases. Within each of the projection vector, the absolute values of weights which are corresponding to the same grid cell were averaged and this value is considered as a position weight. For the proposed method, the combination of a weight map vector and a distance metric in the k-th pair is considered as k-th position weights. For each pair, absolute values of products of $\boldsymbol{L}_k$ and $\boldsymbol{w}_k$ that are corresponding to the same grid cell were averaged and this value is considered as a position weight. In both methods, resulting 225 (=15×15) dimensional position weights were resized into the original image size.

Fig. 6 shows the visualization results on VIPeR (number of training persons =100) and PRID2011 dataset. We can see that the position weights of *ANMM* contain more high weight values in the corners of the images than that of the proposed method. Such regions generally contain background information of each image and are non-relevant to the person. Although cropped regions of PRID2011 are wider than that of VIPeR and more background regions are contained in PRID2011, the effects of the background regions were adequately suppressed in the proposed method. This might be the reason why the proposed method achieved much better results than other baseline methods on this dataset. Although foreground points are one example of discriminative positions on person images, it seems to be evidence that shows the

[4] The reported result in Pedagadi *et al.* [8] is normalized into a scale of 1 to 100 ranks. So, we showed our experimental results in Table II.

TABLE II. COMPARISON WITH OTHER PUBLISHED RESULTS.

Rank score	1	5	10	20
VIPeR				
Proposed	**35.35**	**62.03**	**73.48**	**84.05**
RPLM [5]	27	60	69	83
RDC [7]	15.6	38.42	53.8	70.09
SCEAF [2]	26.49	49.80	60.29	73.54
SDALF [1]	19.11	38.97	51.07	65.29
PRID2011				
Proposed	**21.4**	**39.9**	**50.4**	**63.0**
RPLM [5]	15	33	42	54
GRID				
Proposed	**18.08**	**37.28**	**46.24**	**59.84**
MRank-L$_n$ (RankSVM) [15]	12.24	27.84	36.32	46.56
CAVIAR				
Proposed	**45.00**	**71.11**	**87.78**	**99.17**
LF [8]	33.61	50.55	66.94	90.55

TABLE III. COMPARISON ON DIFFRENT SETTINGS OF VIPeR.

Rank score	1	5	10	20
# of training persons = 200 (gallery size 432)				
Proposed	**25.93**	**50.32**	**63.19**	**76.3**
RDC [7]	12.29	31.55	44.49	59.91
SCEALF [2]	23.71	45.39	55.39	67.89
SDALF [1]	16.58	34.8	45.09	58.75
# of training persons = 100 (gallery size 532)				
Proposed	20.0	40.92	**53.46**	**66.67**
RPLM [5]	11	25	38	52
RDC [7]	9.12	24.19	34.40	48.55
SCEALF [2]	**22.13**	**42.72**	52.3	63.19
SDALF [1]	15.19	31.72	41,45	54.15

proposed method less tends to cause over-fitting compared with existing metric learning methods.

V. CONCLUSION

We have proposed a discriminative accumulation method of local histograms for person re-identification. The proposed method jointly learns pairs of a weight map for the accumulations and a distance metric which emphasizes discriminative histogram dimensions. Since each of the weight map is shared in each of the local histogram dimensions, the proposed method less tends to cause over-fitting. Through experiments, we showed that the proposed method can achieve better re-identification accuracies than other typical metric learning methods on various sizes of datasets. As other metric learning methods fix the distance metric for all input images, the learned values in each pair are also fixed in the proposed method. The extension of the proposed method to be adaptive for each input image is one of our possible future works.

REFERENCES

[1] L.Bazzani, M.Cristani, and V.Murino, "Symmetry-driven accumulation of local features for human characterization and re-identification," *Computer Vision and Image Understanding*, vol. 117, pp. 130–144, 2013.

[2] Y.Hu, S.Liao, Z.Lei, D.Yi, and S.Z.Li, "Exploring structual information and fusing multiple features for person re-identification," in *Proc. IEEE Workshop on Camera Networks and Wide Area Scene Analysis*, 2013.

[3] M.Dikmen, E.Akbas, T.Haung, and N.Ahuja, "Pedestrian recognition with learned metric," in *Proc. ACCV*, 2010.

[4] M.Hirzer, C.Beleznai, M.Köstinger, P.M.Roth, and H.Bischof, "Dense appearance modeling and efficient learning of camera transitions for person re-identification," in *Proc. ICIP*, 2012.

[5] M.Hirzer, P.Roth, M.Köstinger, and H.Bischof, "Relaxed pairwise learned metric for person re-identification," in *Proc. ECCV*, 2012.

[6] M.Köstinger, M.Hirzer, P.Wohlhart, P.M.Roth, and H.Bischof, "Large scale metric learning from equivalence constraints." in *Proc. CVPR*, 2012.

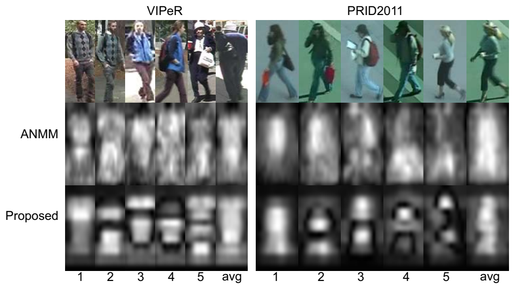

Fig. 6. Example of learned weights. Example images of each dataset are shown in the first row. Position weights corresponding to the first to fifth weights and the average of them are shown in the second and the third row.

[7] W.S.Zheng, S.Gong, and T.Xiang, "Re-identification by relative distance comparison," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 35, pp. 653–668, 2012.

[8] S.Pedagadi, J.Orwell, S.Velastin, and B.Boghossian, "Local fisher discriminant analysis for pedestrian re-identification," in *Proc. CVPR*, 2013.

[9] Y.Shinohara and N.Otsu, "Facial expression recognition using fisher weight maps," in *Proc. FG*, 2004.

[10] F.Wang, X.Wang, D.Zhang, C.Zhang, and T.Li, "marginface: A novel face recognition method by average neighborhood margin maximization," *Pattern Recognition*, vol. 42, pp. 2863–2875, 2009.

[11] D.Gray and H.Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Proc. ECCV*, 2008.

[12] M.Hirzer, C.Beleznai, P.M.Roth, and H.Bischof, "Person re-identification by descriptive and discriminative classification," in *Proc. SCIA*, 2011.

[13] C.C.Loy, T.Xiang, and S.Gong, "Time-delayed correlation analysis for multi-camera activity understanding," *International Journal of Computer Vision*, vol. 90, pp. 106–129, 2010.

[14] D.Cheng, M.Cristani, M.Stoppa, L.Bazzani, and V.Murino, "Custom pictorial structures for re-identification," in *Proc. BMVC*, 2011.

[15] C.C.Loy, C.Liu, and S.Gong, "Person re-identification by manifold ranking," in *Proc. ICIP*, 2013.

APPENDIX

Equivalence of the problems in Eqs.(8) and (10): We show the second constraint in Eq.(8) is satisfied in the solution of Eq.(10) and the objective values in Eqs.(8) and (10) are same.

The k-th weight map vector $\boldsymbol{w}_k$ can be decomposed into two vectors $\boldsymbol{w}_k = \boldsymbol{w}_{kc} + \boldsymbol{w}_{ku}$, where $\boldsymbol{w}_{kc} = \sum_{m=1}^{k-1} \boldsymbol{w}_k^T \boldsymbol{w}_m \boldsymbol{w}_m$ is a vector that shows the correlating components to previously learned weight map vectors and $\boldsymbol{w}_{ku}$ is a vector that shows the uncorrelating components to them, *i.e.* $\boldsymbol{w}_{kc}^T \boldsymbol{w}_{ku} = 0$. Since $\boldsymbol{w}_k^T \boldsymbol{w}_k = 1$, the following relation holds:

$$\boldsymbol{w}_{ku}^T \boldsymbol{w}_{ku} = 1 - \boldsymbol{w}_{kc}^T \boldsymbol{w}_{kc}. \tag{17}$$

After the uncorrelation in Eq.(9), the feature vector calculated using $\boldsymbol{w}_k$ becomes $\boldsymbol{x}_i = \boldsymbol{F}'_i \boldsymbol{w}_{kc} + \boldsymbol{F}'_i \boldsymbol{w}_{ku} = \boldsymbol{F}'_i \boldsymbol{w}_{ku}$. Let $\bar{\boldsymbol{w}}_{ku}$ be a normalized vector $\bar{\boldsymbol{w}}_{ku} = \boldsymbol{w}_{ku}/(\boldsymbol{w}_{ku}^T \boldsymbol{w}_{ku})$. Then the distance in Eq.(3) becomes $D_{\boldsymbol{w}_k, \boldsymbol{L}_k}(\boldsymbol{F}'_i, \boldsymbol{F}'_j) = \boldsymbol{w}_{ku}^T \boldsymbol{w}_{ku} \| \boldsymbol{L}_k^T \boldsymbol{F}'_i \bar{\boldsymbol{w}}_{ku} - \boldsymbol{L}_k^T \boldsymbol{F}'_j \bar{\boldsymbol{w}}_{ku} \|_2^2$. Therefore, the objective value in Eq.(10) becomes $J(\boldsymbol{w}_k, \boldsymbol{L}_k) = \boldsymbol{w}_{ku}^T \boldsymbol{w}_{ku} J(\bar{\boldsymbol{w}}_{ku}, \boldsymbol{L}_k)$. To maximize the objective value, the value $\boldsymbol{w}_{ku}^T \boldsymbol{w}_{ku}$ needs to be maximized among the weight vectors which have same $\bar{\boldsymbol{w}}_{ku}$. From Eq.(17), we can see that the value $\boldsymbol{w}_{ku}^T \boldsymbol{w}_{ku}$ is maximized when $\boldsymbol{w}_{kc}^T \boldsymbol{w}_{kc} = 0$. Since the $\boldsymbol{w}_{kc}$ is a linear combination of the orthogonal bases $\{\boldsymbol{w}_m\}_{m=1}^{k-1}$, all of its coefficients $\{\boldsymbol{w}_k^T \boldsymbol{w}_m\}_{m=1}^{k-1}$ should be zero to satisfy $\boldsymbol{w}_{kc}^T \boldsymbol{w}_{kc} = 0$. Thus, the solution of the problem in Eq.(10) needs to satisfy the second constraint in Eq.(8).

Let $\tilde{\boldsymbol{f}}_{i,v}$ and $\tilde{\boldsymbol{f}}'_{i,v}$ be the v-th row vector of the $\boldsymbol{F}_i$ and $\boldsymbol{F}'_i$, respectively. After the uncorrelation in Eq.(9), the following relation holds:

$$\tilde{\boldsymbol{f}}'_{i,v} = \tilde{\boldsymbol{f}}_{i,v} - \sum_{m=1}^{k-1} \tilde{\boldsymbol{f}}_{i,v} \boldsymbol{w}_m \boldsymbol{w}_m^T, \; v = 1, .., d. \tag{18}$$

By multiplying $\boldsymbol{w}_k$ into Eq.(18), the following relation holds for every v-th row vector of the $\boldsymbol{F}_i$ and $\boldsymbol{F'}_i$:

$$\tilde{\boldsymbol{f}}'_{i,v} \boldsymbol{w}_k = \tilde{\boldsymbol{f}}_{i,v} \boldsymbol{w}_k - \sum_{m=1}^{k-1} \tilde{\boldsymbol{f}}_{i,v} \boldsymbol{w}_m \underbrace{\boldsymbol{w}_m^T \boldsymbol{m}_k}_{=0} = \tilde{\boldsymbol{f}}_{i,v} \boldsymbol{w}_k. \tag{19}$$

Therefore it holds $\boldsymbol{F'}_i \boldsymbol{w}_k = \boldsymbol{F}_i \boldsymbol{w}_k$, and thus the objective values in Eqs.(8) and (10) are same.