Optical Engineering

SPIEDigitalLibrary.org/oe

Coupled metric learning for single-shot versus single-shot person reidentification

Wei Li
Yang Wu
Masayuki Mukunoki
Michihiko Minoh

SPIE

Coupled metric learning for single-shot versus single-shot person reidentification

Wei Li
Kyoto University
Department of Intelligence Science and Technology
Graduate School of Informatics
Kyoto 606-8501, Japan
E-mail: liwei@mm.media.kyoto-u.ac.jp

Yang Wu
Masayuki Mukunoki
Michihiko Minoh
Kyoto University
Academic Center for Computing and Media Studies
Kyoto 606-8501, Japan

Abstract. Person reidentification tackles the problem of building a correspondence between different images of the same person captured by distributed cameras. To date, attempts to solve this problem have focused on either feature representation or learning methods. Usually, the greater the number of the samples for each person, the better the reidentification performance is. However, in the real world, we may not be able to acquire enough samples to give acceptable performance. Here, we focus on the so-called "single-shot versus single-shot" problem: matching one image of a person to another. Because of the extremely small sample class size, there is limited scope to statistically weaken the empirical risk for handcrafted feature representation. Therefore, we resort to metric learning methods, such as the ranking-specialized metric learning to rank (MLR) and the classification-based maximally collapsing metric learning (MCML). Taking advantage of the complementarity between them, we propose a novel "coupled metric learning" approach. This searches for the optimal linear projection for the original feature space using MCML before minimizing the ranking loss via MLR. Experiments on widely used benchmark datasets show encouraging results. [DOI: 10.1117/1.OE.52.2.027203]

Subject terms: person reidentification; coupled metric learning; single-shot versus single-shot.

Paper 121319 received Sep. 13, 2012; revised manuscript received Jan. 10, 2013; accepted for publication Jan. 14, 2013; published online Feb. 4, 2013.

1 Introduction

Inter-camera person image correspondence, known as person reidentification, is one of the most challenging issues in computer vision applications. Its difficulty is chiefly due to variations in illumination, pose and viewpoint, and the resemblance of human clothes, gait, and body shape in nonoverlapping cameras.[1]

Generally, according to the sample class size, person reidentification can be categorized as: "multishot versus multishot" (M versus M), "single-shot versus single-shot" (S versus S), or "single-shot versus multishot" (S versus. M).[2] In M versus M: both the query set and the corpus set contain multiple images of each person. This can be regarded as the topological problem of set-to-set correspondence. In S versus S: each query set and each corpus set contain only one image per person. Topologically, this is a point-to-point correspondence problem. S versus M can refer to either one image in the query set and multiple images in the corpus set or one image in the corpus set and multiple images in the query set for very person. This problem can be treated as a point-to-set correspondence.

Current M versus M approaches follow two main directions. The first pays attention to either feature/signature design or feature/signature selection across sequential images obtained from surveillance tracking. Such feature/signature design considers not only appearance information but also spatial-temporal information. Haar-like features and dominant color descriptors,[3] relying on the AdaBoost scheme can build a satisfactorily invariant and discriminative signature. Histogram Plus Epitome,[4] which focuses on the overall chromatic contents via histogram representation and recurrent local patches via epitomic analysis, can extract the complementary global and local features of the human appearance. Bak et al.[5] emphasized that different regions of an object's appearance ought to be matched using various strategies. They attempted to obtain a distinctive representation by selecting the most descriptive features for a specific class of objects. Relying on an entropy-driven criterion in a covariance metric space, the authors formulated the appearance matching problem as the task of feature selection. The second M versus M direction addresses the problem from the perspective of distance/dissimilarity crafting or optimization. One typical method uses a mean riemannian covariance grid (MRCG).[6] Relying on the dense grid structure, MRCG not only records information about spatial-temporal changes in a person's appearance using a set of covariance descriptors in the Riemannian space but also constructs a suitable dissimilarity measurement for mean Riemannian discriminants. Another exemplary method is set based discriminative ranking,[7] which iteratively constructs convex hulls for set-to-set distance measurements and optimizes the metric for ranking based on these measurements.

For S versus S, there is only one sample image as a query and one in the corpus for each person, and so it is impossible to design a signature by exploiting spatio-temporal information. Thus, many researchers are currently concentrating on feature design. Related methods try to incorporate appearance information and human-body-structure information simultaneously, or learn a suitable metric space to distinguish the sample class. One representative method of feature design is symmetry-driven accumulation of local features

(SDALF),[2] which adopts three powerful features to account for symmetry and asymmetry of body structures. Another feature design method is custom pictorial structure.[8] It focuses on improving the localization of body parts to obtain more reliable visual characteristics. Using metric learning, this could be applied to construct a more discriminative space than the original feature space for better ranking/classification. The large margin nearest neighbor with rejection (LMNN-R)[9] framework converts the issue of person reidentification into a classification problem. It resorts to LMNN[10] to optimize a classification in a traditional support vector machine (SVM) framework. Zheng Wei-Shi et al.[11] recast person reidentification as a learning problem for relative distance comparison in order to learn an optimal similarity measurement between a pair of images of the same person. Technically, they maximized the likelihood of a pair of true matches having a relatively smaller distance than that of a wrong match pair in a soft discriminant manner. RankSVM[12] reformulates the reidentification problem as a ranking problem and learns a metric space in which the potential true match is given the highest rank instead of using a direct distance measurement. Though RankSVM utilizes structural SVM[13] to learn a good pair-wise ranking space, it neglects the role of the loss function in its optimization framework. Standing on the shoulders of metric learning to rank (MLR),[14] optimizing mean reciprocal rank (OMRR)[15] designs the loss function in a structural SVM framework to learn a metric for a list-wise, rather than pair-wise, ranking. The OMRR achieves more satisfactory improvement than previous metric learning methods for person reidentification.

The S versus M lies somewhere between M versus M and S versus S, and therefore, a combination of M versus M methods and S versus S methods are generally applicable.

By way of contrast, the S versus S problem seems more difficult and its accuracy is far lower than the other two. The S versus S not only shares the same challenges as M versus M and S versus M but also has a much smaller sample size for each person, and this is the thing that leads to a bottleneck.

It is obvious that multishot images can be considered as multiple single-shot images for each person. If the S versus S problem could be conquered, the problems of M versus M and S versus M would be readily solved as well. Therefore, from a research perspective, S versus S is important for solving the general person reidentification problem. Moreover, from an application perspective, research on the S versus S problem is very valuable, as it cannot be guaranteed that we will always have multiple samples for each person. For example, we commonly want to retrieve a person with only one query image for whom there is only one registered image in the corpus. Even in visual surveillance systems, we may only have one image of sufficient quality to serve as the query and in the corpus due to occlusion or other unexpected disturbances.

The S versus S person reidentification problem considered in this paper must handle an extremely small sample class size—not only a single-shot sample image as the query but also a single-shot sample image in the corpus for each person. The difficulty is quite significant in this case because the nature of the problem weakens the capability of statistical techniques. There are no enough samples per class to design a robust signature or learn a reliable distance measurement in a round. To solve this problem, we take advantage of the complementarity of two existing metric learning methods to present a new "coupled metric learning" (CML) method. The CML maps the original feature space using a learned linear projection, and this provides a good platform for further optimization by MLR, thus resulting in a better ranking space.

The rest of the paper is structured as follows. In Sec 2, we introduce the CML framework by discussing and analyzing linear projections using maximally collapsing metric learning (MCML) for an original feature space, which is beneficial for the construction of a more reliable ranking space with MLR. In Sec. 3, we justify the reasonability of our modeling approach by comparing with other modeling choices. We then experimentally demonstrate the superiority of CML over MCML, MLR, and other applicable state-of-the-art approaches in Sec. 4, using the recommended feature representation of widely used benchmark datasets. Our conclusions and future research directions are discussed in Sec. 5.

2 Coupled Metric Learning

The notion of distance is fundamental for many data mining/machine learning algorithms. Traditionally, the distance metric has been specified by an a priori assignment. However, metric learning emphasizes that the distance measurement should be learnt from training data.

Dimensional reduction techniques, which exploit the embedding of data, can be categorized as unsupervised metric learning methods such as principal component analysis (PCA), regularized linear discriminant analysis and so on. Supervised metric learning methods, such as Information-Theoretic Metric Learning[16] and Cosine similarity metric learning,[17] utilize objective functions and constraints to optimize the distance measurement. Although these methods have not been directly applied to the issue of person reidentification, they appear to have some potential in this field. When we treat the features of person images as high-dimensional points, the learned metric is able to map them into a new space to improve their intra-class compactness and inter-class separation.

LMNN has already been introduced to address the issue of person reidentification, and it was shown that optimizing a metric space for ranking is more effective than optimizing a metric space for classification when the sample class size is small.[15] A satisfactory ranking space expects the intra-class distances of all samples to be smaller than the inter-class distances. The MLR is based on such an intuitive concept, and OMRR is an application of MLR to the problem of reidentification. It uses a structural SVM framework to optimize the metric for ranking, paying heed to the design of the loss function.

The performance of MLR is more or less determined by the feature representation and sample class size. Therefore, the main idea of our proposed CML method is that, before metric learning, some projective space is optimally searched for the original feature representation by another metric learning method. We will show that MCML could provide a good platform for ranking optimization by MLR. It will be also shown that these two metric learning methods are infact different with a degree of complementary abilities, and their combination could ensure a better performance.

More concretely, CML consists of the application of MCML followed by MLR in this paper.

2.1 *Metric Learning to Rank*

Given a query set $\mathcal{Q} = \{q|q \in \mathbb{R}^d\}$ and a corpus set $\mathcal{X} = \{x_{ql}|x_{ql} \in \mathbb{R}^d\}$, let $\phi_{ql}(x_{ql}, q)$ denote the relative feature representation of a corpus sample x_{ql} w.r.t. q and let w denote the metric we intend to optimize. The desired ranking model could be $g_w(x_{ql}) = w^T\phi_{ql}(x_{ql}, q)$, which scores each x_{ql}. Let $y \in \mathcal{Y}$ be a ranking of $\mathcal{X}$ w.r.t the query q, and $\psi(q, y, \mathcal{X}) \in \mathbb{R}^d$ be a vector-valued joint feature map as defined in Ref. 15. Then, optimizing w for the ranking model $g_w(x_{ql})$ is equivalent to optimizing the following model based on $\psi(q, y, \mathcal{X})$.

$$\arg\min_w \frac{1}{2}\|w\|^2 + \frac{C}{|\mathcal{Q}|}\sum_q \xi_q \tag{1}$$

subject to

$$w^T\psi(q, y_q^*, \mathcal{X}) \geq w^T\psi(q, y, \mathcal{X}) + \Delta(y_q^*, y) - \xi_q,$$
$$\forall\ q, y \neq y_q^*; \quad \xi_q \geq 0, \quad \forall\ q,$$

where y_q^* is the ground truth ranking of $\mathcal{X}$ for a given $q \in \mathcal{Q}$, ξ_q is the slack variable, C is the trade-off parameter, and $\Delta(y_q^*, y)$ is the loss function to penalize the prediction of y instead of y_q^*.

For person reidentification, the most widely used evaluation criterion is the cumulative matching characteristic (CMC) curve. The CMC illustrates how the performance (recognition/re-acquisition rate) improves as the number of requested images increases. Nevertheless, the CMC curve is not a single-trial measurement, and therefore cannot be directly optimized. Despite this, some other criteria may potentially coincide with the CMC curve to some extent, with mean reciprocal rank (MRR) being the closest among existing candidates. Conceptually, the reciprocal rank of a query response is the multiplicative inverse of the rank of the first correct match, and MRR is the average of such reciprocal ranks of results over the whole query set.[15] To a degree, it coincides with such a practical performance expectation, and is thus a good alternative to the CMC curve for performance evaluation. Since only the rank of the first correct match is counted, the ranks of both other correct matches and any incorrect matches are arbitrary. Thus, there are large amounts of ranking instantiations of a given ground truth. In practice, such a choice is quite reasonable, and there seems to be no significantly better options.

According to Ref. 14, for the ranking model MLR, the loss function $\Delta_{\text{mrr}}(y_q^*, y)$ based on MRR is an effective choice, which can be as simple as

$$\Delta_{\text{mrr}}(y_q^*, y) = 1 - S_{\text{mrr}}(q, y), \tag{2}$$

in which

$$S_{\text{mrr}}(q, y) = \frac{1}{|\mathcal{Q}|}\sum_{q\in\mathcal{Q}}\begin{cases} 1/r_q, & r_q < K; \\ 0, & r_q \geq K, \end{cases} \tag{3}$$

where r_q is the position of the first relevant item w.r.t. q in the ranking y, and K is the number of top ranked items to be considered; for the ground truth y_q^*, $S_{\text{mrr}}(q, y) = 1$.

The optimization of MLR can be solved using a cutting-plane algorithm.[18] After learning, a new metric space with more ranking capability than the original feature space can be constructed.

Based on a structural SVM framework, MLR learns a metric such that data rankings induced by their distances from a query can be optimized against various ranking measures. In the MLR framework, it has been shown that optimizing the metric for partial-order features $\psi(q, y_q, \mathcal{X})$ is equivalent to optimizing the metric for $\phi_{ql}(x_{ql}, q)$.[14] $\phi_{ql}(x_{ql}, q)$ is a kind of feature mapping that characterizes the relationship between the query sample q and the corpus sample l. As $\|q - l\|_W^2 = \langle W, (q - l)(q - l)^T\rangle_F$, it follows that $\phi_{ql} \triangleq -(q - l)(q - l)^T$. It is obvious that ϕ_{ql} describes differential information for each sample pair. It contains information about one sample relative to the other, so ϕ_{ql} can also be treated as a kind of "relative feature", though in matrix form rather than vector form, different from traditional features. Naturally, if such a matrix holds pair-wise information about samples in the same class, we name it the "intra-class relative feature", and if it describes pair-wise information on samples from different classes, we name it the "inter-class relative feature". Using these relative features, the ranking model optimization can be viewed from a classification perspective. It is reasonable to demand that the relative features are of sufficient quality if we hope to provide a good partial-order feature space for MLR. According to the definitions of relative features, it is natural to require that samples of the same class should stay as close to each other as possible while samples from different classes should remain far away from one another. Certainly and notably, such a requirement is sufficient but not necessary for providing a good partial-order feature space. There are other ways to make relative features meet the requirement, like feature design. Because the sample class size is extremely small, there is only a limited space for each class to statistically weaken the empirical risk in the process of hand designing features. Hence, we recommend searching a suitable projective space w.r.t. the original feature representation. The MLR itself performs a linear projection on the original data using a structural SVM framework, which is capable of dealing with the high-dimension small-sample problem, although the small sample class size limits its power. Accordingly, we resort to some other auxiliary projection approaches for MLR that are beneficial for mapping a new feature space, though not for directly optimizing the ranking. The newly mapped space is expected to have the property that samples in the same class stay close to each other and samples from different classes are kept far apart. Undoubtedly, MCML[19] coincides with such requirements. However, as a metric learning method, it is difficult to apply MCML for feature mapping. To overcome this barrier, we simplify the projection into a linear form and perform some mathematical derivations, shown in Eq. (4).

Suppose x is the feature point, and that the linear mapping is expressed by $g(x) = xL$. The projected relative feature $\phi_{\text{projected}}$ can then be presented as

$$\begin{aligned}\phi_{\text{projected}} &= -[g(q) - g(l)][g(q) - g(l)]^T \\ &= -(q - l)LL^T(q - l)^T \\ &= -(q - l)A(q - l)^T,\end{aligned} \tag{4}$$

where A is a positive semi-definite (PSD) matrix. This matrix is considered the Mahalanobis metric to be optimized in MCML, and can be decomposed as $A = L \times L^T$. Optimizing A is equivalent to optimizing the feature mapping in Eq. (4), which could make the projected relative feature more representative and conducive to partial-order feature space construction, potentially benefiting MLR.

2.2 *Maximally Collapsing Metric Learning*

The MCML relies on the geometric intuition that all points in the same class should be mapped to a single location in the feature space and all points in other classes should be mapped to other locations. Basically, MCML obtains a compact low-dimensional feature representation of the original input space.[19]

The MCML uses Kullback-Leibler divergence in the objective function to make intra-class distances as small as zero and inter-class distances as large as infinite.

Given n labeled samples (x_i, y_i), where $x_i \in \mathcal{R}^r$ and $y_i \in \{1 \ldots H\}$, the distance between any different points indexed by i and j can be defined as

$$d(x_i, x_j|A) = d_{ij}^A = (x_i - x_j)^T A (x_i - x_j), \tag{5}$$

where A is a PSD matrix.

To learn a metric that approximates the ideal geometric intuition, for each training point, a conditional distribution over other points has been introduced. Specifically, for each x_i, a conditional distribution over any other x_j, where $j \neq i$, is defined as

$$P^A(j|i) = \frac{e^{-d_{ij}^A}}{Z} = \frac{e^{-d_{ij}^A}}{\sum_{k \neq i} e^{-d_{ik}^A}}, \quad j \neq i, \tag{6}$$

where j means any sample other than i. The framework of MCML is as below:

$$\arg \min_A \sum_i KL[P_0(j|i)]|[P^A(j|i)] \tag{7}$$

subject to

$$A \in \text{PSD}, \tag{8}$$

where Z is the normalizing factor, and $P^A(j|i)$ takes a pseudoprobabilistic form to describe the conditional distribution over points. $P_0(j|i)$ is the ideal bi-level distribution as in Eq. (9). If all points in the same class were mapped to a single point and infinitely far from points in different classes, we would have the ideal bi-level distribution:

$$P_0(j|i) = \begin{cases} 1, & y_i = y_j; \\ 0, & y_i \neq y_j. \end{cases} \tag{9}$$

In the optimizing process, at each iteration, MCML takes a small step in the direction of the negative gradient of the objective function, then the MCML metric is projected back onto the PSD cone by taking the eigen-decomposition of A and substituting zero for the components with negative eigenvalues.[19]

MCML can perform a dimensional reduction by spectral decomposition, and the reduced dimension t can be determined a priori. The eigen-decomposition of a metric A can be written as

$$A = \sum_{h=1}^{r} \lambda_h v_h v_h^T, \tag{10}$$

where r is the number of nonnegative eigenvalues, which is equivalent to the original feature dimension, λ_h are the eigenvalues of A, and v_h are the corresponding eigenvectors.

The matrix A that has less than full rank corresponds to Mahalanobis distance based on the low-dimensional projection. Hence, we can then select the largest t eigenvalues and their eigenvectors to obtain the low-rank metric:

$$A_t = \text{diag}(\sqrt{\lambda}_1, \ldots \sqrt{\lambda}_t)[v_1^T; \ldots; v_t^T]. \tag{11}$$

Generally speaking, the low-dimensional space projection is not guaranteed to be the same as the projection corresponding to minimizing the objective function of MCML, subject to a rank constraint on the optimal metric, unless the rank of the optimal metric is less than or equal to t. However, as demonstrated in Ref. 19, for practical problems, it is often the case that the optimal metric has an eigenspectrum which is rapidly decaying, so that many of its eigenvalues are indeed very small. This suggests the low rank solution will be close to optimal. So A_t can be applied to map the original space to a new low-dimensional space, in which intra-class compactness and inter-class separation can be simultaneously achieved to as great a degree as possible.

Usually, the reduce dimension t can be determined a-priori. Although t is determined by heuristic, we need to avoid the case that t is too large or too small. If t is too large, the new feature space of the reduced dimension will be too noisy. If t is too small, the new feature space of the reduced dimension will be too sensitive. Both cases will damage the effect of intra-class compactness and inter-class separation. Here, we suggest t to be approximate 80% of the original feature dimension.

3 Modeling Justification

In the previous sections, we determined that MCML could construct a linear projective space that is beneficial to MLR.

Specially, this implies that the modeling performance of CML could outstrip that of MLR. We denote this by "MCML + MLR $>_{\text{perf}}$ MLR". (Here, $>_{\text{perf}}$ means the former performs better than the latter. Similarly, in the following, $<_{\text{perf}}$ means the latter performs better than the former, $=_{\text{perf}}$ means the former performs almost the same as the latter, $\geq_{\text{perf}}$ means the former performs no worse than the latter, and $\leq_{\text{perf}}$ means the latter performs no worse than the former.)

To further highlight the advantages of MCML + MLR, in this section, we compare it with other modeling choices by discussing the relationship between MCML and MLR.

3.1 *Analysis of CML*

It is impossible to give a direct, mathematical proof for the superiority of MCML + MLR due to the difficulty of unifying MCML and MLR into a single optimized framework. Alternatively, we will provide evidence of the benefits of coupling MCML and MLR based on their complementarity,

which means that MCML contributes toward MLR learning a more satisfactory space.

The role of MCML in CML can be understood from the point of view of noise reduction. Low-dimensional space projection by MCML has the effectiveness of data de-noising that coincides with MLR's target.

Traditionally, PCA is widely used to find low-dimensional embedding for de-noising before supervised learning. Although it has a similar functionality as MCML, PCA does not have the discriminative ability as that of MCML, and the space mapping given by it is not guaranteed to improve the performance of MLR.

The objective of MLR is that the distance between the samples in the same class should be smaller than that between the samples from different classes.

In the ideal case, when MCML meets its target perfectly, MLR's objective will also be indirectly and perfectly achieved, because intra-class distances will all be zero and inter-class distances will all be infinite. Hence, in general, MCML works in the same direction as MLR.

Although, in the real world, it is impossible to obtain ideal data for MCML or MLR, it is intuitive that if samples from the same class become closer together and samples from different classes become farther away, it will be easier for MLR to make intra-class distances smaller than inter-class distances.

Therefore, MCML is able to contribute toward MLR learning a more satisfactory space, namely "MCML + MLR $>_{\text{perf}}$ MLR". If all the contributions from MCML is redundant to MLR, which is very unlikely to happen, at least we can get "MCML + MLR $=_{\text{perf}}$ MLR". However, it has never appeared in our experimental results to be presented later.

3.2 *Comparison with Other Modeling Choices*

We will compare MCML + MLR with other modeling choices for CML, including MLR + MLR, MCML + MCML, and MLR + MCML, where the "+" sign denotes a strict "left-to-right" order (i.e., MLR maps the original space to a new linear projective space for further optimization by MCML).

One simple and direct reason to reject MLR + MLR and MCML + MCML is that the metric learned by MLR has a unique solution given certain training samples, and so does MCML. Therefore, from a performance perspective, "MLR + MLR $=_{\text{perf}}$ MLR" and "MCML + MCML $=_{\text{perf}}$MCML".

A more in-depth explanation is required for rejecting MLR + MCML, or indeed for not simply implementing MCML.

It is easy to discern the similarity between MCML and MLR. Obviously, both use a convex optimization framework to learn a Mahalanobis metric. The key difference is that MCML forces intra-class distances to be zero and inter-class distances to be infinite. So the behavior of pulling samples in the same class close together is conditionally independent to the behavior of pushing samples from different classes far away, which could be denoted as "independent measurement". Independent measurement indicates there is not direct relationship between intra-class distance and inter-class distance. The MCML aims to minimize the sum of the loss function for both behaviors, described by Kullback-Leibler divergence. Therefore, MCML may seek to balance the compactness of samples in the same class and the separation between samples from different classes. Undoubtedly, an optimal combination of intra-class compactness and inter-class separation will be the most beneficial for classification.

On the contrary, MLR concerns the correlative relationship between samples in the same class and samples from different classes. It forces intra-class distances to be smaller than inter-class distances for all samples, which could be denoted as "correlative measurement". It is different from the optimal combination of intra-class compactness and inter-class separation reflected in MCML. Correlative measurement indicates there is a direct correlation between intra-class and inter-class distances. Optimizing the correlative measurement will be conducive to ranking because it matches the requirement of a good ranking: given a query, samples within a corpus from the same class as the query have smaller distances than samples from different classes.

With an ideal MCML metric, all intra-class distances will be zero and all inter-class distances will be infinite. Thus, the inter-class distances will definitely be larger than the intra-class distances, so the optimization of correlative measurement can also be achieved. Nevertheless, in real cases, the data distribution is complex due to small between-class but large within-class variations, especially in the case of an extremely small sample class size, as discussed in this paper. There is a possibility that MCML will sacrifice correlative measurement to attain the minimization of its own objective function, which is expressed as the sum of the discrepancies between intra-class distances and zero, and the discrepancies between inter-class distances and infinite. In other words, MCML may sacrifice the relationship that ensures intra-class distance is smaller than inter-class distance to optimize the accumulation of intra-class compactness and inter-class separation.

To illustrate this, we give an example of MCML, where the independent measurement impairs the correlative measurement. An illustration is shown in Fig. 1.

Consider the layout of three points in a two-dimensional space; the scaling of axes by the metric is simplified in two orthogonal dimensions, with (u, v) denoting the scaling parameters. The two points i and j are in the same class, marked by the same color, and point k is from a different class, marked by a different color. With i assigned as the query, we can see that the intra-class distance is initially smaller than the inter-class distance.

As mentioned in Ref. 19, the objective function of the optimization framework described by Eq. (7) is convex in matrix A. Equation (7) can be rewritten as

$$\arg\min_{A} \sum_{i} P_0(j|i) \log \frac{P_0(j|i)}{P^A(j|i)}. \tag{12}$$

In order to seek A to optimize Eq. (12), the constant part can be ignored and the part containing A should be maintained. $P_0(j|i)\log P_0(j|i)/P^A(j|i) = P_0(j|i)(\log P_0[j|i) - \log P^A(j|i)] = P_0(j|i)\log P_0(j|i) - P_0(j|i)\log P^A(j|i)$, which shows that these terms are an additive constant and a positive multiplicative constant w.r.t. $-\log P^A(j|i)$ and thus can be ignored. Therefore, minimizing Eq. (12) is equivalent to minimizing $f(A)$ as below:

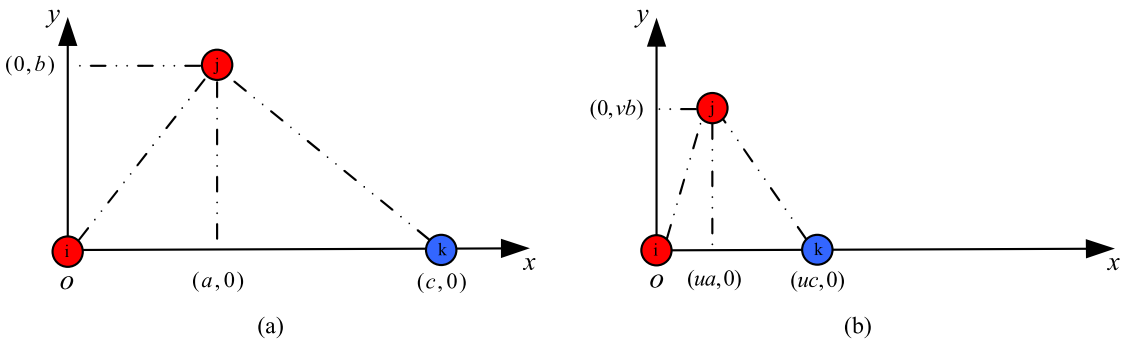


Fig. 1 Exemplar to show independent measurement impairs correlative measurement by MCML. Sample classes are distinguished by color.

$$f(A) = -\sum_{i,j;y_i=y_j} \ln P^A(j|i) = \sum_{i,j;y_i=y_j} d_{ij}^A + \sum_i \ln Z_i. \quad (13)$$

In each step, the metric scaling parameters will be adjusted to simulate the decrease of Eq. (13). If the example appears to be minimizing the sum of independent measurements at the expense of the correlative measurements, the following system of inequalities will have a nonempty solution set for u and v.

$$\begin{cases} a^2+b^2+a^2+b^2+\ln(e^{-(a^2+b^2)}+e^{-c^2})+ \\ \ln(e^{-(a^2+b^2)}+e^{-[(c-a)^2+b^2]}) > (ua)^2+(vb)^2+(ua)^2+ \\ (vb)^2+\ln(e^{-[(ua)^2+(vb)^2]}+e^{-[(uc-ua)^2+(vb)^2]})+ \\ \ln(e^{-[(ua)^2+(vb)^2]}+e^{-(uc)^2}); \\ a^2+b^2<c^2; \\ a^2+b^2<(c-a)^2+b^2; \\ (ua)^2+(vb)^2 \geq (uc)^2; \\ u \geq 0; \\ v \geq 0; \\ v+u>0. \end{cases} \quad (14)$$

In this system, the first inequality means that in each convergence step, after scaling, the value of the objective function is smaller than in the previous step. The second and third inequalities describe the initial distance relationship among the three points, and the fourth describes the impairment of the correlative measurement after scaling by the metric. The fifth, sixth, and seventh inequalities are the constraints of u and v. u and v cannot be zero simultaneously because this would imply that the metric matrix of MCML is the zero matrix, at which point MCML loses its meaning.

To prove the existence of a nonempty solution set for u and v, we do not require all of the analytical solutions for the system of inequalities. Instead, we focus on the case $(ua)^2+(vb)^2=(uc)^2$, because this is the critical condition of the system of inequalities. We can then obtain $v>0$, and thus,

$$\left[(u,v)|0<u<\sqrt{\ln\frac{(1+e^{-c^2+a^2+b^2})(1+e^{-c^2+2ac})-2}{2e^{-c^2+2ac}}}, v = \sqrt{\frac{(uc)^2-(ua)^2}{b^2}}\right].$$

In order to ensure the solution set is not empty, we set

$$\frac{(1+e^{-c^2+a^2+b^2})(1+e^{-c^2+2ac})-2}{2e^{-c^2+2ac}} > 1,$$

and then acquire $e^{c^2-2ac}+e^{a^2+b^2-2ac}+e^{-c^2+a^2+b^2}>3$. According to the property of inequalities that the arithmetic mean is greater than the geometric mean, and considering other inequalities, we use the amplification and minification method to obtain the conditions for a, b, and c as $\sqrt{a^2+b^2}<c<(a^2+b^2/2a)$, where $0<a<(b/\sqrt{3})$. This proves the fact that the sum of independent measurements may be minimized at the expense of the correlative measurements. In other words, MCML may impair the ranking to optimize its own objective function. Therefore, MCML is not equivalent to MLR, and cannot replace MLR. If we force MCML to perform the ranking directly, it is not guaranteed that we will obtain the optimum ranking results, though there is some potential for this. Hence, unlike "MCML + MLR", "MLR + MCML" is likely to be no better than MLR itself, i.e., "MLR + MCML $\preceq_{\text{perf}}$ MLR,".

Therefore, the proposed MCML + MLR is the most reasonable choice, and cannot be replaced by other model forms such as MLR + MCML, MLR + MLR, and MCML + MCML.

Conclusively, MCML and MLR have their own strengths. MCML optimizes the metric for classification and MLR optimizes the metric for ranking, and one cannot replace the other's role. Nevertheless, in a sense, it is this difference between MCML and MLR that offers the space for their complementarity and cooperation. MCML + MLR is the best choice for exploring this complementarity.

4 Experiments and Results

4.1 *Dataset Description*

We demonstrate the effectiveness of our method across several public benchmark datasets: VIPeR,[20] CAVIAR4REID,[8] ETHZ,[21] iLIDS,[22] iLIDS-MA,[23] and iLIDS-AA.[23] Each dataset has its own characteristics and challenges.

The VIPeR dataset consists of 1264 images for 632 unique pedestrians. Each pedestrian image pair has been taken from arbitrary viewpoints under varying illumination conditions. Complicated variations of viewpoint, illumination, and pose make VIPeR one of the most challenging datasets for person reidentification.

The CAVIAR4REID dataset[8] has been extracted from the CAVIAR dataset,[24,25] which consists of several sequences filmed in the entrance lobby of the INRIA Labs and in a

shopping center in Lisbon. For CAVIAR4REID, there are 72 pedestrians, 50 of which have two camera views while the remaining 22 have only one camera view. It includes people walking alone, meeting with others, window shopping, entering and exiting shops. Person images vary a lot in terms of pose, light, viewpoint, resolution, and so on. These variations make the person reidentification issue more difficult.

The ETHZ dataset is composed of three video sequences of crowded street scenes captured by two moving cameras mounted on a chariot. We utilize three subsets extracted by Schwartz and Davis for person reidentification.[26] ETHZ has smaller variations of pose and viewpoint, yet more occlusions than VIPeR. SEQ1, denoted by ETHZ1, has 83 pedestrians within 4857 images; SEQ2, denoted by ETHZ2, has 35 pedestrians within 1936 images; SEQ3, denoted by ETHZ3, has 28 pedestrians within 1762 images.

The iLIDS MCTS dataset is a publicly available video dataset captured in a busy airport arrival hall under a multicamera CCTV network. It is a real scenario. i-LIDS contains 479 images of 119 pedestrians extracted by Zheng et al. for testing their context-based pedestrian reidentification method.[22] Images from the i-LIDS dataset, derived from nonoverlapping cameras, are subject to large illumination changes and occlusions (not present in VIPeR). i-LIDS-MA[23] contains 40 individuals extracted from two cameras, with 46 frames from both cameras annotated manually for each individual. i-LIDS-AA[23] is made up of 100 individuals seen from both cameras, obtained by a HOG-based human detector and tracker. Undoubtedly, the noisy detection and tracking results make the task of person reidentification more challenging. Sample images of each dataset are shown in Fig. 2.

We use all of the people in each dataset. We normalize all images to 128×48 pixels, the same size as the images in VIPeR, then randomly halve each dataset into training data and testing data. We repeat this 10 times for cross-validation and average the results for evaluation. The experimental results are illustrated by CMC curves. For each person, we randomly select two images, one as the query image and the other as the corpus image, in both the training and testing stages. Each time, we use the same selected data for comparing the methods.

4.2 *Feature Representation*

According to current research, local color statistical descriptors perform remarkably well for person reidentification. As each dataset has its own characteristics, we propose the most suitable feature representation for each set on an individual basis.

VIPeR is distinguished by its large viewpoint, illumination, and pose variations. Hence, it is reasonable to explore the local color statistical descriptors by considering the human-body-structure information. We recommend the concatenation of densely sampled color histograms (DCHs)[9] and a weighted HSV color histogram (wHSV),[2] denoted by DCHs + wHSV.[14] The DCHs are able to deal with illumination variations and occlusion due to their dense sampling of the cells that contain the local color statistical description of human body appearance. wHSV has the merit of dealing with viewpoint and pose changes because it not only makes use of global color information but also takes symmetric and asymmetric properties of the human body into consideration.

CAVIAR4REID has large pose, illumination, and viewpoint variations, which are similar to VIPeR. Accordingly, the feature representation is recommended to be DCHs + wHSV as well.

Fig. 2 Sample images from eight public benchmark datasets. Each column represents the matching pair of the same person.

ETHZ does not have such marked viewpoint variations, but changes in illumination and walking pose are apparent. Therefore, it is necessary to capture the characteristics of a person's appearance by considering local color statistical descriptors from different spaces. Multispace color histograms (MSCHs)[11] are recommended, as they combine the DCHs of multiple color spaces, including RGB, YCbCr, and HSV. Although convertible to one another, these spaces describe appearance information from different angles. These variety of spaces enhance the descriptive power of DCHs to handle illumination and walking pose changes.

For i-LIDS, i-LIDS-MA, and i-LIDS-AA, though sharing the problems of VIPeR and ETHZ, the crux is significantly more accessory occlusions and larger similarity of clothes between different people. Hence, we can utilize color and texture information together. We combine DCHs with a Schmid-filter-bank (DCHs + SFB) to capture both the color and texture information of a person's appearance, where SFB[27] captures the texture information using 13 separate rotationally invariant filters.

4.3 *Result Analysis*

We compare the proposed CML method with both the related methods and the state-of-the-art to demonstrate the effectiveness and superiority of our model. These methods include SDALF, OMRR, MCML, MLR + MCML, and a model called "relative distance comparison" (RDC) proposed by Zheng Wei-Shi et al.[11]

As a representative method that can deal with S versus S person reidentification, the SDALF compared in our experiments is not exactly the original one. Original SDALF concatenates three features into one signature: wHSV, maximally stable color regions (MSCR), and recurrent high-structured patches (RHSP). Here, SDALF is generalized to indicate a direct matching by the most suitable feature representation measured with the Bhattacharyya metric. Although SDALF is not a learning-based approach, it uses the same features as those in our method, and thus this pure matching based method is valuable for demonstrating the effectiveness of our learning based algorithm.

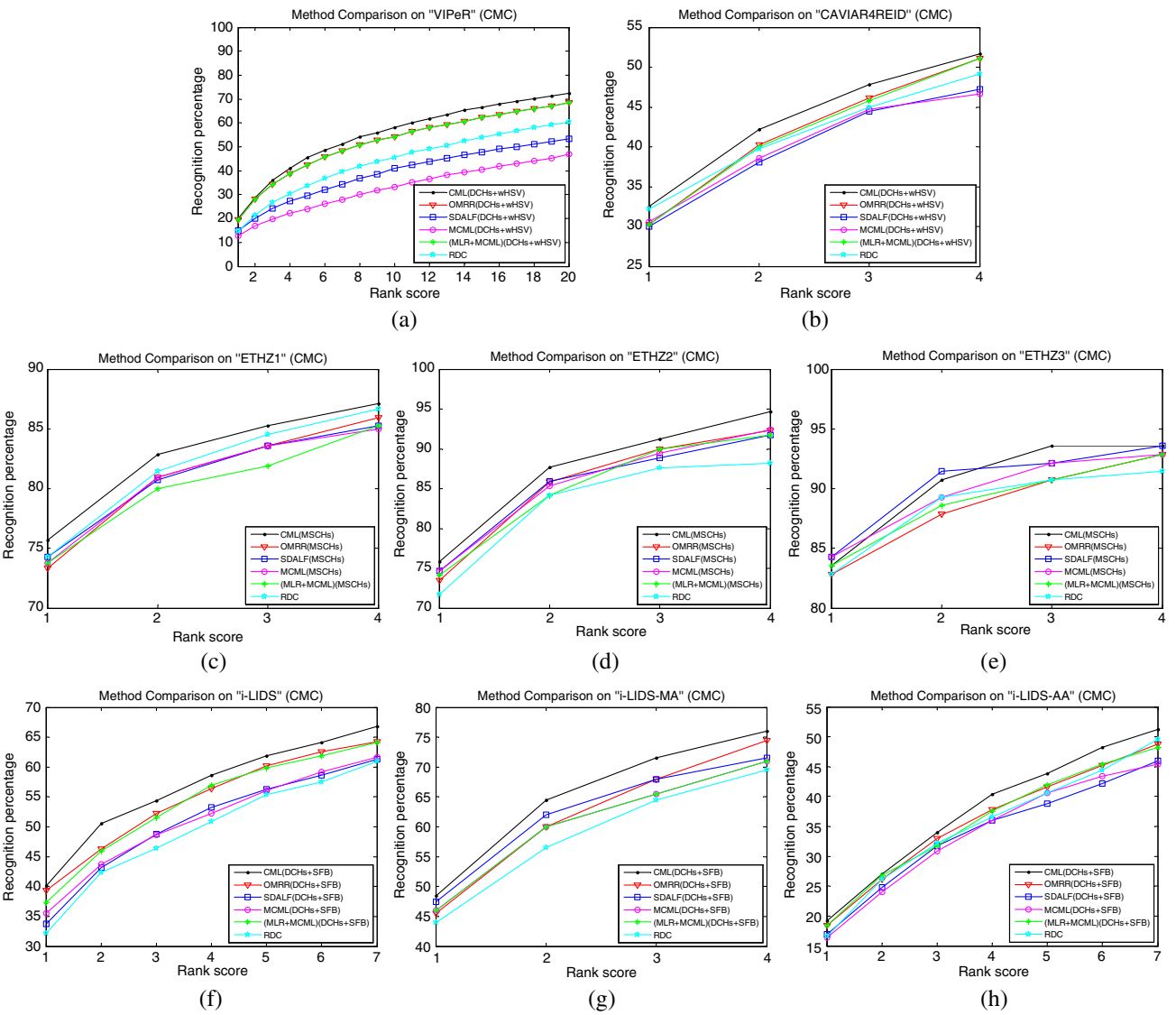


Fig. 3 CMC performance comparison on the eight public benchmark datasets: (a) VIPeR, (b) CAVIAR4REID, (c) ETHZ1, (d) ETHZ2, (e) ETHZ3, (f) i-LIDS, (g) i-LIDS-MA, and (h) i-LIDS-AA.

OMRR is an application of MLR to person reidentification. OMRR can not only represent MLR but it is also a capable metric learning method that has achieved remarkable results for single-shot person reidentification.

MCML is a typical metric learning method, and is implemented here for further validation of the superiority of our CML method.

Moreover, the MLR + MCML alternative is also adopted to validate the reasoning behind our proposed CML modeling formulation.

RDC is a state-of-the-art method that focuses on learning a reliable metric based on the novel relative distance comparison modeling and has achieved encouraging results compared with other metric learning methods for person reidentification. Here, we use the original code of RDC model provided by the authors on our datasets for comparison.[11]

Although OMRR and RDC can work on single-shot person reidentification, they actually use multiple images per-person in the training and testing stages if possible[15,11] (except on the VIPeR dataset), whereas the proposed method CML focuses on S versus S, using only two images per person for both training and testing. Fewer training samples will make the problem more challenging.

There are also some other good methods for M versus M person reidentification. For example, Slawomir Bak et al. have proposed a new method called "learning to match appearances by correlations in a covariance metric space". They extracted the effective covariance descriptor based on multiple-shot images, then further applied learning for feature selection, and thus achieved good results. This method works effectively on the M versus M problem, but likely has limitations for the S versus S problem. Without enough images, it is unreliable to perform correlation-based feature selection, and the measurement between covariance descriptors will lose effect as well. In fact, Slawomir Bak et al. have mentioned in their paper that their method belongs to the group of multiple-shot approaches. That is why the authors have not tested it on the VIPeR dataset. However, the proposed method CML is toward the S versus S case. It is unfair to compare it with M versus M methods.

Overall, the experimental results in Fig. 3 show that CML significantly prevails over other methods. In greater detail, we can see that "MCML $\leq_{\text{perf}}$ MLR (OMRR)" and "MLR + MCML $\leq_{\text{perf}}$ MLR (OMRR)" in terms of ranking for all datasets except ETHZ3. This verifies our argument in Sec. 3 that MCML cannot replace MLR for ranking. For ETHZ3, "MCML $\geq_{\text{perf}}$ MLR (OMRR)" and "MLR + MCML $\geq_{\text{perf}}$ MLR (OMRR)", and CML is even slightly outperformed by SDALF with the same original feature representation in the Rank-1 CMC score. This is unsurprising because ETHZ3 is the most specific case. It has the fewest people, only 28 altogether. When we use half of these to carry out training, too few samples inevitably leads to a high probability of over-training. Even though the proposed CML model is less over-fitting than any other learning-based methods compared here. Furthermore, especially for VIPeR, which is the largest and the most difficult dataset for S versus S person reidentification, CML evidently outperforms the other methods.

Conclusively, the experimental results not only justify the model formulation for CML but also show its superiority over the techniques compared here.

5 Conclusion

In this paper, we have proposed the CML method for one of the most difficult cases of person reidentification, S versus S. We recast the original problem into a ranking optimization problem, and then utilized the complementarity of two metric learning models to optimize a linear projection of the original feature space. This reduced the empirical risk caused by the small sample class size, which to some extent contributed to the construction of a more satisfactory space for ranking. Experimental results have verified the superiority of our method. Possible future work includes the application of CML to human detection and tracking.

Acknowledgments

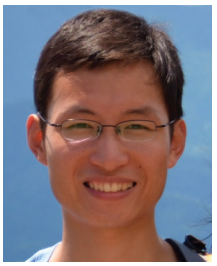
This work was supported by "R&D Program for Implementation of Anti-Crime and Anti-Terrorism Technologies for a Safe and Secure Society", Special Coordination Fund for Promoting Science and Technology of the Ministry of Education, Culture, Sports, Science and Technology, the Japanese Government.

References

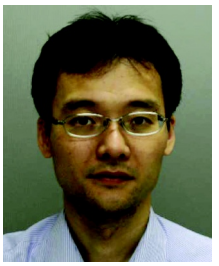
1. G. Doretto et al., "Appearance-based person reidentification in camera networks: problem overview and current approaches," *J. Ambient Intell. Humaniz. Comput.* **2**(2), 127–151 (2011).
2. M. Farenzena et al., "Person reidentification by symmetry-driven accumulation of local features," in *Proc. IEEE Conf. on Comput. Vis and Pattern Recognit.*, pp. 2360–2367, IEEE Computer Society, San Francisco, California (2010).
3. B. Slawomir, C. Etienne, B. Francois, and T. Monique, "Person reidentification using haar-based and dcd-based signature," in *Proc. 7th IEEE Int. Conf. on Adv. Video and Signal Based Surveillance*, pp. 1–8, IEEE Computer Society, Boston, Massachusetts (2010).
4. B. Loris et al., "Multiple-shot person reidentification by HPE signature," in *Proc. 20th IEEE Int. Conf. on Pattern Recognit.*, pp. 1413–1416, IEEE Computer Society, Istanbul, Turkey (2010).
5. S. Bak et al., "Learning to match appearances by correlations in a covariance metric space," in *Proc. 12th European Conf. on Comput. Vis.*, Vol. 1, pp. 1–14, Springer, Florence, Italy (2012).
6. S. Bak et al., "Multiple-shot human reidentification by mean riemannian covariance grid," in *Proc. 8th IEEE Int. Conf. on Adv. Video and Signal-Based Surveillance*, pp. 179–184, IEEE Computer Society, Klagenfurt, Austria (2011).
7. Y. Wu et al., "Set based discriminative ranking for recognition," in *Proc. 12th European Conf. on Comput. Vis.*, Vol. 7574, pp. 497–510, Springer-Verlag, Berlin Heidelberg (2012).
8. D. S. Cheng et al., "Custom pictorial structures for reidentification," in *Proc. British Mach. Vis. Conf.*, pp. 68.1–68.11, BMVA Press, Dundee, United Kingdom (2011).
9. M. Dikmen et al., "Pedestrian recognition with a learned metric," in *Proc. 10th Asian Conf. on Comput. Vis.*, pp. 501–512, Springer-Verlag, Berlin, Heidelberg (2010).
10. K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *J. Mach. Learn. Res.* **10**, 207–244 (2009).
11. W. S. Zheng, S. Gong, and T. Xiang, "Reidentification by relative distance comparison," *IEEE Trans. Pattern Anal. Mach. Intell.* (2012) (In Press).
12. B. Prosser et al., "Person reidentification by support vector ranking," in *Proc. British Mach. Vis. Conf.*, pp. 21.1–21.11, BMVA Press, Aberystwyth, Wales, United Kingdom (2010).
13. T. Joachims, T. Finley, and C. N. Yu, "Cutting-plane training of structural SVMs," *Mach. Learn.* **77**(1), 27–59 (2009).
14. B. McFee and G. Lanckriet, "Metric learning to rank," in *Proc. 27th Int. Conf. on Mach. Learning*, pp. 775–782, Omnipress, Haifa, Israel (2010).
15. Y. Wu et al., "Optimizing mean reciprocal rank for person reidentification," in *Proc. 8th IEEE Int. Conf. on Adv. Video and Signal-Based Surveillance*, pp. 408–413, IEEE Computer Society, Klagenfurt, Austria (2011).
16. J. V. Davis et al., "Information-theoretic metric learning," in *Proc. 24th Int. Conf. on Mach. Learning*, pp. 209–216, ACM, New York (2007).
17. H. V. Nguyen and L. Bai, "Cosine similarity metric learning for face verification," in *Proc. 10th Asica Conf. on Comput. Vis.*, Vol. 6493, pp. 709–720, Springer, Berlin Heidelberg (2010).
18. S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, New York (2004).

19. A. Globerson and S. Roweis, "Metric learning by collapsing classes," *Advances in Neural Information Processing Systems*, Vol. 18, pp. 451–458, MIT Press, London, United Kingdom (2006).
20. D. Gray and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *Proc. 10th IEEE Int. Workshop on Performance Evaluation for Tracking and Surveillance*, pp. 41–47, IEEE Computer Society, Rio de Janeiro, Brazil (2007).
21. A. Ess, B. Leibe, and L. V. Gool, "Depth and appearance for mobile scene analysis," in *Proc. 11th IEEE Int. Conf. on Comput. Vis.*, pp. 1–8, IEEE Computer Society, Rio de Janeiro, Brazil (2007).
22. W. S. Zheng, S. Gong, and T. Xiang, "Associating groups of people," in *Proc. British Mach. Vis. Conf.*, pp. 23.1–23.11, BMVA Press, London, United Kingdom (2009).
23. B. Slawomir et al., "Boosted human reidentification using Riemannian manifolds," *Image Vis. Comput.* **30**(6–7), 443–452 (2012).
24. C. H. Kuo, C. Huang, and R. Nevatia, "Multi-target tracking by on-line learned discriminative appearance models," in *Proc. 23rd Conf. on Comput. Vis. and Pattern Recognit.*, pp. 685–692, IEEE Computer Society, San Francisco, California (2010).
25. B. Slawomir et al., "Multi-target tracking by discriminative analysis on riemannian manifold," in *Proc. 19th IEEE Int. Conf. on Image Process.*, Vol. 1, pp. 1–4,IEEE Computer Society, Florida (2012).
26. W. R. Schwartz and L. S. Davis, "Learning discriminative appearance-based models using partial least squares," in *Proc. XXII Brazilian Symposium on Comput. Graphics and Image Process.*, pp. 322–329, IEEE Computer Society, Washington, DC (2009).
27. J. Zhang, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: a comprehensive study," *Int. J. Comput. Vis.* **73**(2), 213–238 (2007).

Wei Li is a PhD candidate in Department of Intelligence Science and Technology at Kyoto University currently. He received a BS degree in Measuring and Control Technology and Instrumentations and a MS degree in Instrument Science and Technology from Southeast University in 2007 and 2010, respectively. His research interests include computer vision, pattern recognition and convex optimization.



Yang Wu is currently a post-doctoral researcher of Academic Center for Computing and Media Studies, Kyoto University. He received a BS degree in information engineering and a PhD degree in pattern recognition and intelligent systems from Xi'an Jiaotong University in 2004 and 2010, respectively. From September 2007 to December 2008, he was a visiting student in the General Robotics, Automation, Sensing and Perception lab at University of Pennsylvania. His research is in the fields of computer vision and pattern recognition, with particular interests in the detection, tracking and recognition of humans and also generic objects. He is also interested in image/video search and retrieval, along with machine learning techniques.



Masayuki Mukunoki received the BS, MS and PhD degrees in information engineering from Kyoto University. He is now an associate professor in the Academic Center for Computing and Media Studies and a faculty member in the Graduate School of Informatics, in Kyoto University. His research interests include computer vision, video media processing, lecture video analysis, and human activity sensing with camera.

Michihiko Minoh is a professor at Academic Center for Computing and Media Studies, Kyoto University, Japan. He received the BEng, MEng, and DEng degrees in information science from Kyoto University, in 1978, 1980, and 1983, respectively. He served as director of ACCMS from April 2006 to March 2010 and concurrently served as vice director in the Kyoto University Presidents Office from October 2008 to September 2010. Since October 2010, he has been vice-president, chief information officer at Kyoto University, and director-general at Institute for Information Management and Communication, Kyoto University. His research interests include a variety area of Image Processing, Artificial Intelligence and Multimedia Applications, particularly, model centered framework for the computer system to help visual communication among humans and information media structure for human communication. He is a member of Information Processing Society of Japan, Institute of Electronics, Information and Communication Engineers of Japan, the IEEE Computer Society and Communication Society, and ACM.