PAPER

People Re-Identification with Local Distance Comparison Using Learned Metric

Guanwen ZHANG[†], **Jien KATO**[†a)], ***Members***, **Yu WANG**[†], ***Nonmember***, ***and*** **Kenji MASE**[†], ***Fellow***

SUMMARY In this paper, we propose a novel approach for multiple-shot people re-identification. Due to high variance in camera view, light illumination, non-rigid deformation of posture and so on, there exists a crucial inter-/intra- variance issue, i.e., the same people may look considerably different, whereas different people may look extremely similar. This issue leads to an intractable, multimodal distribution of people appearance in feature space. To deal with such multimodal properties of data, we solve the re-identification problem under a local distance comparison framework, which significantly alleviates the difficulty induced by varying appearance of each individual. Furthermore, we build an energy-based loss function to measure the similarity between appearance instances, by calculating the distance between corresponding subsets in feature space. This loss function not only favors small distances that indicate high similarity between appearances of the same people, but also penalizes small distances or undesirable overlaps between subsets, which reflect high similarity between appearances of different people. In this way, effective people re-identification can be achieved in a robust manner against the inter-/intra- variance issue. The performance of our approach has been evaluated by applying it to the public benchmark datasets ETHZ and CAVIAR4REID. Experimental results show significant improvements over previous reports.

key words: *multiple-shot re-identification, local distance comparison, multimodal distribution*

1. Introduction

People re-identification refers to the problem that, recognize people when he/she leaves one camera view and enters another camera view, or recognize people when he/she reappears in a given camera view. This technology is crucial for inter-camera tracking as well as for understanding people behavior within a camera network. It is required by applications in various fields such as video surveillance and has received more and more attention as security cameras come into wide use.

Due to the low image resolution and the large distance between people and cameras in many cases of practical use, biological information, such as people's face or gait, is generally unavailable. In addition, because of discontinuity in the visual fields of non-overlap multiple cameras, continuous visual tracking and intra-camera motion information of people cannot be immediately utilized for re-identification [1]. Therefore, in the current literature, researches on people re-identification mainly focus on analyzing people appearance, with an acceptable assumption that people will not change their clothing during the observation period. The challenge in such an appearance-based people re-identification approach principally comes from appearance variants induced by the light illumination, camera views and non-rigid deformation of posture. This makes the intra-camera variance sometimes becomes even larger than the inter-camera variance, namely, same people could look considerably different in the videos captured by different cameras, whereas different people could look extremely similar in the videos captured by the same camera (see Fig. 1). This problem is known as inter-variance and intra-variance issue (inter-/intra-variance issue, in short).

Given a probe and a gallery of collected person images, to find the best matching of the query people from a certain amount of candidates, two steps are indispensable for appearance-based methods [2]: (1) seek a stable feature representation that models a discriminative signature of people appearance, and (2) measure similarity between such signatures (or models) with some optimal criterions. During this processing, according to the number of images utilized to create people appearance models, the alternatives are single-shot-based methods (using only one single image) or multiple-shot-based methods (using a set of images). The research on single-shot-based re-identification has been well investigated [2]–[9]. On the other hand, as many successful tracking algorithms are put to practical use, the application of multiple-shot people re-identification naturally arises and attracts more and more attention [10]–[12].

This paper proposes a novel approach for multiple-shot people re-identification. Compared to the single-shot case, multiple-shot-based methods can utilize more convincing information from multiple images, which is promising for re-identifying people with high accuracy. On the other hand, images within the image set of the same people manifest

Fig. 1 Multiple shots of two persons, with each row showing one. Each person's appearance varies considerably between the images owing to differences in posture, illumination, and resolution.

Manuscript received November 27, 2013.
Manuscript revised May 9, 2014.

[†]The authors are with the Graduate School of Information Science, Nagoya University, Nagoya-shi, 464–8601 Japan.
a) E-mail: jien@is.nagoya-u.ac.jp

DOI: 10.1587/transinf.2013EDP7424

large variance in camera views, light illumination, non-rigid deformation and so on. This poses a big challenge to exploit unstable semantic information of these images that traditional single-shot-based approaches do not need to take account of. More precisely, in a multiple-shot-based re-identification scenario, we face an intractable multimodal people appearance distribution in a high dimensional feature space. In such a distribution, appearance instances of the same people tend to form several separated clusters, each with some specific semantic meaning associated with particular feature properties such as color or silhouette. We call these clusters as *subsets* of people appearance. The subsets with the same semantic meaning for different people may be closer than those with different semantic meaning for the same people (see Fig. 2). This situation leads to the idea to conduct re-identification by comparing subsets using the local distance and furthermore to define the similarity of people on the basis of the local distance among subsets.

Based on above discussion, in this paper, we formulate the multiple-shot re-identification problem under a local distance comparison framework, and moreover construct an energy-based loss function for the local distance comparison where the local distance is calculated in k-nearest neighbor (k-NN) way. The loss function defined in this way takes into account two aspects: it favors small distances that indicate high similarity between appearances of the same people, and also penalizes small distances or undesirable overlaps between subsets, which reflect high similarity between appearances of different people. Benefit from such a local distance comparison framework and the loss function, our approach is robust to the above-mentioned inherent inter-/intra-variance issue in the multiple-shot people re-identification.

The principal contribution of this paper is two-fold: (1) we formulate the multiple-shot people re-identification under a local distance comparison framework to adapt to the multimodal property of people appearance distribution, (2) we propose a local-distance-based loss function that is able to greatly improve the re-identification accuracy, compared to the existing methods without paying specific attention to the multimodal properties of data.

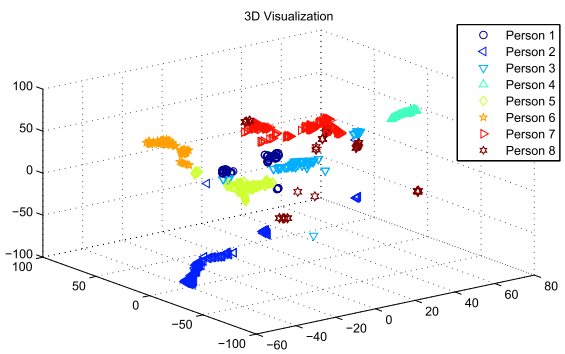


Fig. 2 Visualization of appearance instances corresponding to 8 people from ETHZ Seq. 1 [13]. Each instance is characterized by a feature vector. To visualize the multimodal distribution in the high-dimensional space, the feature dimension is projected into 3D by the t-SNE [14]. Different people are denoted by different colors and different makers.

The rest of the paper is organized as follows. Section 2 reviews related works on people re-identification and shows the originality of our approach. Section 3 describes our proposed method, the core of this paper, by first introducing a Mahalanobis metric learning algorithm (3.1) and then defining the local-distance-based loss function (3.2). Experimental results are discussed in Sect. 4. Finally, we present our conclusions and future perspectives in Sect. 5.

2. Related Work

According to above-mentioned two steps for solving the re-identification problem, existing researches can be roughly divided into two groups. Most researchers, as in the first group, try to seek stable feature representations, and have made a lot of efforts to extract and represent discriminative signatures or models of people appearances. On the other hand, fewer researchers focus on measuring the similarity between feature representations. They use either learning algorithms or comparison methods to find out the optimal measurement that is most likely to give correct matching results.

In the first group, various invariant global and local descriptors are extracted over images to model a discriminative signature of people appearance. For example, Wang et al. [4] introduce a shape and appearance context based representation. They utilize relative information of human body parts to model a spatial co-occurrence distribution of people appearance. The re-identification is performed by comparing such appearance models. To realize more accurate comparison, very detailed information about human body parts is integrated into appearance models later. For instance, an approach called symmetry-driven accumulation of local features (SDALF) is proposed by Farenzena et al. [7]. After separating human body into two parts along the horizontal axis, they compare the corresponding features of each part between people, by weighting the features in proportion to their distance from the symmetry axis in the vertical direction. Such symmetrical information is also used in the work of Bazzani et al. They extend their previous method called histogram plus epitome (HPE) [11] into a new method called asymmetry-based HPE (AHPE) [10]. More complex parts information is also used in the work of Cheng et al. [9], where exact parts such as chest, head, thighs and legs are first detected by pictorial structures (PS), then Bhattacharyya distance between features of the parts, with a weight different for different parts, is calculated by custom pictorial structures (CPS) model.

Although a lot of efforts toward feature representation approaches have been devoted in this manner, there are serious limitations on them. Firstly, they highly rely on detection or extraction results of people appearance. The poor results, induced by complex background or severe noise, obstruct to extract reliable and informative features. Secondly, detecting or extracting people appearance requires

a great deal of extra time consumption, which makes re-identification impossible for online applications. More essentially, since these approaches do not pay enough attention to camera view, they often yield the features that work well on videos captured from one camera view, but work poorly on those captured from other camera views. Therefore, it is very difficult to achieve impressive performance of people re-identification by simply comparing people appearance models, particularly in case large inter camera variance exiting.

In the second group, approaches are also appearance-based but specially focus on measuring the similarity between feature representations by using learning algorithm or comparison methods. For example, Schwart et al. [15] extract a great deal of color, texture and edge information to build a high dimensional feature representation. They then use partial least squares (PLS) to weight these features according to their ability to discriminate the signatures of people in gallery, and finally project the signatures into a low dimensional, separable subspace. Following the same idea of feature selection, Gray and Tao [5] propose a boosting approach based on Adaboost to select the optimal feature representation for people matching. From another viewpoint, Prosser et al. [6] treat the pairwise relationship of people appearance as relative ranking. They use RankSVM to explore the separability between true match and false match in a high dimensional feature space, and finally find out a stable feature representation. This idea is extended later by Zheng et al. [3]. They propose a probabilistic relative distance comparison (PRDC) model based on the probabilistic relation of distance between true match and false match to switch the re-identification into a distance learning problem. They further extend their PRDC to an ensemble relative distance comparison (RDC) model by improving the scalability and tractability of the original one [2]. In addition, under a framework of large margin nearest neighbor (LMNN), Dikmen et al. [8] learn a distance metric with reject constraints to make larger margin between instances with different labels. The learned distance metric is then generalized to target task to perform distance comparing. Recently, some researches such as the work of Li et al. [16] focus on transfer learning, in order to deal with overfitting problem caused by small training data. In Li et al.'s work [17], they sample several images as the "third party" images from another dataset, similar with but different from the probe and gallery sets. The images in the probe and gallery sets are expressed as a collaborative representation based on the "third party" images using a sparse coding method. They aim to discover an intermediate feature representation that bridges the gap between the query and the gallery sets.

All above approaches in the second group actually try to project the people appearance instances to a feature space where the appearance instances of the same people are close while those of different people are far away. However, since in practical use, there exists a lot of variance in appearance for the same people, such an attempt is extremely difficult. Moreover, these existing approaches almost focus on single-shot re-identification. In case of multiple-shot people re-identification where the intra-variance is even larger than inter-variance, the learned distance metric/relation by these approaches on a specific training dataset will hardly work well under various conditions.

Actually, in the literature, there are also some multiple-shot approaches. These approaches are mainly simple extension from single-shot approaches by such as averaging or re-weighting the signatures or models obtained from each image. Compared with single-shot approaches, they have the ability to supplement some of missing data but do not have the capacity to treat the intra-/inter-variance issue.

By contrast, our proposed approach follows the nature of multiple-shot people re-identification problem. We treat the re-identification task as a local distance comparison problem with a multimodal appearance distribution. This strategy enables us to effectively deal with the inter-/intra-variance issue. That is, our approach can work well on not only arbitrary appearance of the same people, but also the similar appearance of different people, by utilizing rich information from multiple images of probe and gallery sets.

3. Methods

Proposed approach consists of two phases: offline metric learning and online re-identification. In the offline metric learning phase, we use the LMNN algorithm to learn a Mahalanobis metric, while in the online re-identification phase, we use our loss function with the learned metric to conduct local distance comparison. The whole procedure is summarized in Fig. 3.

In the following sections, we assume people appearances have already been successfully detected and cropped from videos by bounding boxes, to simplify the problem and concentrate on the main issue. For each bounding box image, we use a feature vector to represent the people appearance instance. On the basis of such feature representation, we discuss our offline learning and online re-identification respectively.

3.1 Distance Metric Learning

In this section, we introduce the distance metric learning

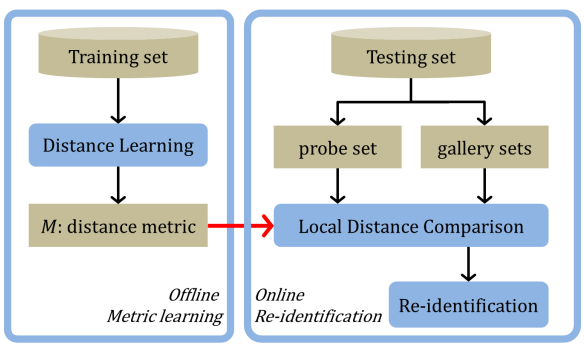


Fig. 3 Outline of proposed approach.

method. Since finding the nearest neighbors of an appearance instance depends directly on the distance metric of local subsets, selecting a proper metric learning method is critical. In existing metric learning methods [18]–[21], large margin nearest neighbor (LMNN) [22] is the state-of-the-art one for Mahalanobis metric learning. Since the large margin framework and local linear constraint of LMNN are perfectly consistent with our needs, we choose the LMNN algorithm as our distance metric learning method in the offline phase. The LMNN algorithm is briefly introduced below.

Let $D = \{(x_i, t_i)\}_{i=1}^{n}$ denote a training dataset composed of n pairs of data, with each pair consisting of a feature vector x_i and its label t_i. We measure the similarity of two feature vectors x_i and x_j by a Mahalanobis distance function:

$$d_M(x_i, x_j) = (x_i - x_j)^T M(x_i - x_j), \tag{1}$$

where M is a symmetric, positive definite matrix that completely parameterizes the distance function. The objective of distance metric learning is to learn M from D with the condition that in target neighbors, if $t_i = t_j$, $d_M(x_i, x_j)$ gets a small value, and otherwise $d_M(x_i, x_j)$ gets a large value.

The LMNN algorithm is an excellent algorithm for such a learning task. It learns M in two steps: (1) for each data point x_i, select its k-nearest data points that have the same label t_i as that of the target neighbors, and (2) estimate M by minimizing the following cost function:

$$L(M) = \sum_{i,j\rightsquigarrow i} d_M(x_i, x_j) + \xi \sum_{i,j\rightsquigarrow i,l} (1 - y_{il})[1 + d_M(x_i, x_j) - d_M(x_i, x_l)]_+, \tag{2}$$

where $j \rightsquigarrow i$ is used to indicate that x_j is a target neighbor of x_i, i.e., x_j is one x_i's the nearest neighbors with the same class label. The indicator variable $y_{il} = 1$ if and only if x_i and x_l have the same class label, $y_{il} = 0$ otherwise. The first term in Eq. (2) penalizes a large distance between data point x_i and its target neighbors x_j. On the other hand, the second term penalizes a small distance between x_i and all imposter points x_l, which are defined as points with different class labels. ξ is a predefined positive constant, and $[a]_+ = max(a, 0)$ is the standard hinge loss function that makes the cost function convex. Given the target neighbor membership, M can be obtained by using the semi-definitive programming (SDP) algorithm. We initialize the M in Eq. (2) with Euclidean distance. In this case, the neighbors are calculated as Euclidean distance in the original feature space.

In LMNN, the cost function only penalizes large distances between training data points and their target neighbors. Compared to other methods that attempt to minimize the distances between training data points and all other data points with the same labels, it can obtain a good solution more efficiently. More essentially, LMNN is the most suitable method for distance metric learning in this work, because our approach deals with a multimodal data distribution.

LMNN is very efficient when feature vectors are low-dimensional (in the range 50-100), but becomes computationally expensive with increasing dimension (more than 200) [8], [22]. However, since metric learning is performed in the offline phase and this processing would not affect the time required for online re-identification, this cost is acceptable.

3.2 Loss Function

The people re-identification can be formulated as a problem to find people (probe) in a gallery that contains different instances of a large amount of seen-before people. Let $C_i = \{x_{i1}, x_{i2}, \ldots, x_{iN_i}\}$ denote the instance set of person i consisting of N_i images. We call this instance set *image set* or *set*, and re-identification is performed by comparing two sets, one for probe and the other for the potential target (in short, target) in the gallery. As described in Sect. 1, an image set is likely to form several separate *subsets* in feature space.

Our loss function is specially designed to deal with this type of problems. Our basic idea originates in simple intuitions: (1) if image sets being compared match, the total distance between the corresponding local subsets will be ideally minimal, and (2) if image sets being compared match, the corresponding local subsets will ideally have minimal overlap with the subsets of other people. Inspired by Weinberger et al. [22] and Chopra et al. [23], we construct our loss function in an energy-based way, which consists of two terms:

$$L(p, t) = L_d(p, t) + \gamma L_o(p, t). \tag{3}$$

Here, p and t stand for the probe and target sets. The first term measures the total distance between local subsets of the compared people, and the second term measures the loss induced by an overlap with the subsets of other people. We want to find t^* that minimizes $L(p, t)$, by using this equation. γ ($\in [0, 1]$) is introduced to balance the weights of the two terms.

3.2.1 The First Term: $L_d(p, t)$

This term is used to calculate the distance between the probe set and target set. A smaller value suggests higher similarity between the instances of the two sets, indicating that the two sets are more likely to belong to the same people.

We divide the probe set indicated by C_p into two parts, C_{pA} and C_{pB}. For each instance in C_p, if it can be classified as a target indicated by C_t via the k-NN classification rule, we let it belong to C_{pB}; otherwise, it belongs to C_{pA}. Because C_{pB} is more likely to belong to C_t, from the perspective of k-NN rule, it is reasonable to exclude C_{pB} and regard C_{pB} as a part of C_t when we measure the distance between the probe and target sets. We use the notation C_t^+ to denote the union $C_t^+ = C_t \cup C_{pB}$. The first term, namely, the

distance between C_p and C_t, is thus defined as the distance between C_{pA} and C_t^+:

$$L_d(p,t) = \sum_{x \in C_{pA}} \sum_{y \in C_t^+} h_d(x,y) d_M(x,y), \tag{4}$$

where $h_d(x,y)$ is an indicator function such that $h_d(x,y) = 1$ if and only if y is among the k-NN to x, and $h_d(x,y) = 0$ otherwise. This term can be obtained by first finding the k-nearest instances in C_t^+ for each instance in C_{pA}, and then summing up the distances between all pairs. Here, the distance is calculated by Eq. (1) (Mahalanobis distance), and M is the distance metric learned in the offline learning phase.

We emphasize that in Eq. (4), we compute the distance in the k-NN manner by considering various situations between subset pairs (overlapping with each other or being far from each other). This is because we think it is unsuitable to simply compute pairwise distance, namely, the total distance of each instance in probe image set with all the instances in the target image set.

3.2.2 The Second Term: $L_o(p,t)$

If C_p and C_t match, the second term penalizes the loss induced by the overlap between the new set (i.e., $C_p \cup C_t$) and other sets in the gallery. To simplify the discussion, we regard all instances in the gallery, except C_t, as a new set denoted by C_e.

We firstly define the loss of an instance from a local subset. Each instance x has k-nearest neighbors N_x, which share the same class label with x. N_x can be mathematically defined as: $N_x = \{x_1, x_2, \ldots, x_k\}$, subject to $d_M(x_i,x) - d_M(x_j,x) < 0$, $\forall x_i \in N_x, \forall x_j \notin N_x$, and x_j has the same class label as x. The distances are calculated by using the distance metric learned in offline phase. If there exists another instance z with a different label that is closer to x than any instance y in N_x, this instance is called *invader* of the subset. The loss induced by z is defined as

$$I(x,z) = \sum_{y \in N_x} [d_M(x,y) - d_M(x,z)]_+, \tag{5}$$

where $[a]_+ = max(a,0)$ denotes the standard hinge loss. Note that here, we still use the local distance calculated in the k-NN manner. Similar as in Eq. (4), Eq. (5) is also measured by the learned distance metric.

As mentioned in the discussion on L_d, we still focus on the instances in C_{pA}, and we treat the instances in C_{pB} as a part of C_t. We consider the loss from two aspects: (1) the invasion from C_e to C_{pA}, and (2) the invasion from C_{pA} to C_e. They are different from each other in general. On the basis of this consideration as well as the definition in Eq. (5), we formulate $L_o(p,t)$ as follows:

$$\gamma L_o(p,t) = \sum_{x \in C_{pA}} \sum_{z \in C_e} \{\gamma_1 I(x,z) + \gamma_2 I(z,x)\}, \tag{6}$$

where γ is decomposed into γ_1 and γ_2 ($\in [0,1]$) to balance the weights of the two terms. Again, if C_p and C_t match, the new set $C_p \cup C_t$ should have a small overlap with C_e. Consequently, the loss will be small. The meanings of $L_d(p,t)$ and $L_o(p,t)$ are illustrated in Fig. 4.

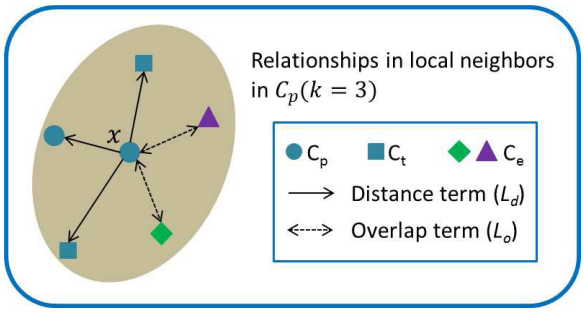


Fig. 4 Illustration of L_d and L_o terms. In some local neighbors of C_p, L_d measures the similarity between the regarded instance x ($\in C_p$) and its local neighbors y ($\in C_t$), while L_o penalizes an invasion z that is closer to x than any x's neighbor instance y.

Algorithm 1: Local Distance Comparison

Input : probe set: C_p
gallery: $G = \{C_i\}_{i=0}^N$
distance metric: M
parameter: $\gamma_1, \gamma_2; k$
Output: similarity ranking list: R

begin
  for ***all*** $x \in C_p$ **do**
    compute $x's$ k-nearest neighbors N_x' within G
  for ***all*** $C_t \in G$ **do**
    for ***all*** $x \in C_p$ **do**
      if $N_x' \subset C_t$ **then**
        $C_{pB} := C_{pB} + x$
      else
        $C_{pA} := C_{pA} + x$
      $C_t^+ := C_t + C_{pB}$
    $L_d \leftarrow 0$
    for ***all*** $x \in C_{pA}$ **do**
      compute $x's$ k-nearest neighbors N_x within C_t^+
      $L_d := L_d + \sum_{y \in N_x} d_M(x,y)$
    $C_e \leftarrow G - C_t$
    $L_{o,1} \leftarrow 0$
    for ***all*** $z \in C_e$ **do**
      for ***all*** $x \in C_{pA}$ **do**
        $L_{o,1} := L_{o,1} + \sum_{y \in N_x} [d_M(x,y) - d_M(x,z)]_+$
    for ***all*** $x \in C_e$ **do**
      compute $x's$ k-nearest neighbors N_x^* within C_e
    $L_{o,2} \leftarrow 0$
    for ***all*** $z \in C_{pA}$ **do**
      for ***all*** $x \in C_e$ **do**
        $L_{o,2} := L_{o,2} + \sum_{y \in N_x^*} [d_M(x,y) - d_M(x,z)]_+$
    $Loss \leftarrow L_d + \gamma_1 \cdot L_{o,1} + \gamma_2 \cdot L_{o,2}$
    store $Loss$ and t in R
  sort R according to $Loss$ in ascending order

3.2.3 Multiple-shot Re-identification

Given a probe set C_p, the multiple-shot re-identification is accomplished by matching the C_p to the gallery $G = \{C_i\}_{i=1}^{N}$. The similarity between C_p and each set $C_t (\in G)$ is evaluated by loss function Eq. (3), which further consists of two steps: calculating the local distance between C_p and C_t by Eq. (4), and calculating the invasion between $C_p \cup C_t$ and all other sets in C_e by Eq. (6). Small loss reflects the high similarity.

The output of the re-identification is expected to be a ranking list in descending order of the similarity. The instances at the top of the list give the identifications that the probe most likely to be. The pseudocode of the re-identification method is presented in Algorithm 1.

4. Experiments

4.1 Dataset

We evaluate the proposed method by applying it to two public multiple-shot datasets: ETHZ of Schwartz et al. [15] and CAVIAR4REID of Cheng et al. [9]. These datasets cover different genres and include different people postures, under various illumination conditions, with various degrees of occlusion and camera resolution. Therefore, they are very challenging for people re-identification.

Video records in the ETHZ dataset have been captured from moving cameras in streets [13]. This dataset consists of Seq. 1 (83 people with 4,857 images), Seq. 2 (35 people with 1,936 images), and Seq. 3 (28 people with 1,762 images). This dataset was originally used for pedestrian detection. Owing to the wide range of illumination and occlusion levels, it has been widely employed in re-identification researches [7], [9]–[11], [15]. The availability of numerous image shots for each person is useful for machine learning.

On the other hand, CAVIAR4REID is built especially for people re-identification tasks. All images are extracted from the CAVIAR [24] dataset, which is composed of video records captured in a shopping center. It consists of 72 people in 1,220 images, in which 50 people are captured in two camera views, and remaining 22 people are captured in a single camera view. The images in this dataset are selected by maximizing the variance with respect to the resolution, illumination, occlusion and posture [9].

4.2 Feature Representation

Similarly with the work done by Gray et al. [5], Zheng et al. [2], [3] and Dikmen et al. [8], we use a histogram of mixed color and texture features for people appearance representation. We divide an image into 15 image regions, three in the vertical direction and five in the horizontal direction. For each region, color and texture features are extracted and integrated into a histogram.

As to color, 5 channels of RGB and HSV (without V channel) are selected, and each channel is represented by 16 bins. For the remaining V channel, we construct a filter bank that consists of 24 Gabor filters (4 scale and 6 direction) and convolute with it. The response of each Gabor filter is quantified into 8 bins, and is further summarized into a histogram as texture features. In this way, we build a feature vector with 4,080 dimensions. The integration of such low-level color and texture cues can be regarded as a generic representation of people appearance.

Since the high-dimensional features contain noise as well as redundancy and are expensive in computation, it is necessary to reduce the dimension by using the method such as principal component analysis (PCA). However, there is an inherent trade-off between dimension reduction and the performance of the re-identification method, because the relationship among appearance instances in local subsets will change when they are projected in a lower-dimensional feature space [22]. In the experiments, we observe that when we reduce the feature vector into 200 dimensions, we can achieve a good balance between re-identification accuracy and computation speed. That motivates us to use 200-dimensional feature vectors reduced by PCA in experiments.

4.3 Experimental Method

In the experiments, according to the number of people in the dataset, we divide each dataset into halves: one is testing dataset for online re-identification and the other is training dataset for offline distance metric learning.

In online people re-identification phase, from a testing dataset, we randomly select N images for each people to generate the probe, and also select N images for each people to generate the gallery. In particular, for each sequence in ETHZ dataset, we select N frames of the head of each person's image sequence as the probe and select different N frames from the rest as the gallery. On the other hand, for CAVIAR4REID dataset, we select images taken from different camera views as the probe and gallery respectively. Re-identification experiments are conducted by finding a match for each people in probe from the gallery. It should be noticed that, the size of the gallery is the half number of people in the datasets.

Following the similar evaluation method used in [2]–[4], [7], [10], we choose cumulative matching characteristic (CMC) curve to show re-identification matching rates. It is a popular performance evaluation metric for re-identification that represents the expectation of finding the correct match in the top n candidates. We repeat the experiment 20 times and calculate the average re-identification rate.

In the local distance comparison by using k-NN distance, the choice of the k value is critical. A small value may lead to the disadvantage of high noise sensitivity, while a large value may make the boundaries of classes less distinct and also increase computation cost. As suggested by Duda et al. [25] and Maier et al. [26] , $k = \sqrt{n}$ reflects the nature of the data well. From the practical perspective, it is reasonable to have around 10 instances for each person.

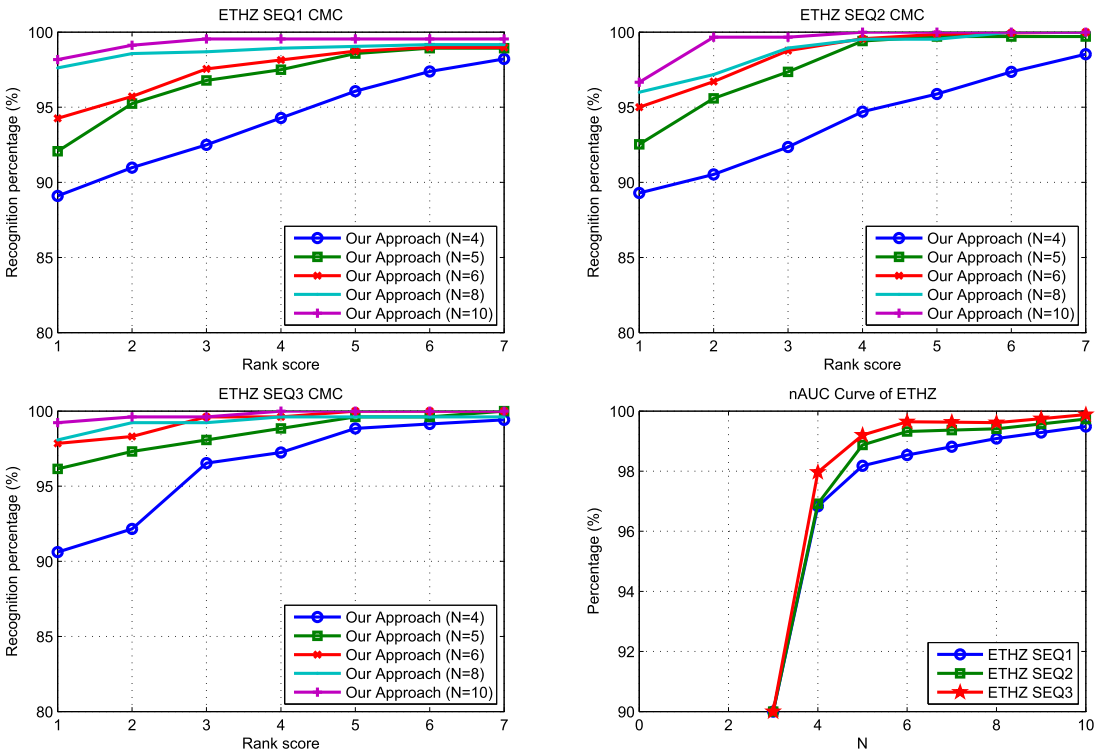


Fig. 5 Evaluation of CMC curve for Seq. 1 – Seq. 3 of ETHZ with different N values of the image set (first three figures), and the average AUC curve of CMC for Seq. 1 – Seq. 3 also with different N values (the last figure).

In the following experiments, we thus choose $k = 3$ in the re-identification. To study the impact of k to the proposed approach, we carry out further experiments and discussion about the effect of different k value in Sect. 4.4. The trade-off parameters in Eq. (6) are empirically set as $\gamma_1 = 0.5$ and $\gamma_2 = 0.5$ in the following experiments. The detailed discussion of the effectiveness of the different value of γ_1 and γ_2 is also presented in Sect. 4.4.

In the following experiments, we pay special attention to investigating: (1) how does the number of images N for each person in the probe or gallery set influence the re-identification results, (2) whether the k-NN distance used in our approach is superior to the pairwise distance, and (3) whether the proposed method is better adapted to multi-modal data distribution than exiting methods.

4.4 Experimental Results

We first analyze the re-identification rate of our approach for different numbers of images (N). Since the images of each people in ETHZ are numerous, we conduct our experiments with Seq. 1 – Seq. 3 of ETHZ with $N = \{4, 5, 6, 8, 10\}$ and plot the CMC in Fig. 5. From the results for three sequences, we can see that the re-identification rate increases as N increasing. However, when N becomes greater, the re-identification rate can no longer be significantly improved. In order to get an overview of the re-identification rate together with N, we use an normalized area under the curve (AUC) to show performances. We calculate the area under the CMC curve for each N value and normalize the results for these three sequences (Fig. 5). The analysis in Fig. 5 shows that $N = 6$ is a good choice which equilibrates N with the re-identification rate.

In our loss function, local distance comparison is conducted in k-NN manner for corresponding subsets of appearance instances. This k-NN distance comparison is more effective for a multimodal feature distribution compared to traditional comparison methods such as pairwise method, AHISD (Affine/convex Hull based Image set Distance) and CHISD, which measures the geometric distance between image sets. To prove this, we choose a pairwise method to compare with our proposed method. In pairwise comparison, for each appearance instance in the probe set, we sum up the distances between regarded instance and all instances in the target set and obtain the total distances as the loss. After comparing the probe set and the sets in the gallery, the minimal loss indicates the re-identification result.

We use $N = \{5, 10\}$ as the number of images for probe set and gallery sets, and choose ITML [21], RCA [19], Euclidean and the distance metric learned in offline training phase as the measure in the pairwise distance comparison. The CMC results for ETHZ datasets are shown in Fig. 6, where we compare pairwise distance with our local distance ($N = \{5\}$). In CAVIAR4REID, because there are 22 people with only 10 images, we choose $N = 5$ for pairwise comparison. From Fig. 6, we can see that the k-NN dis-

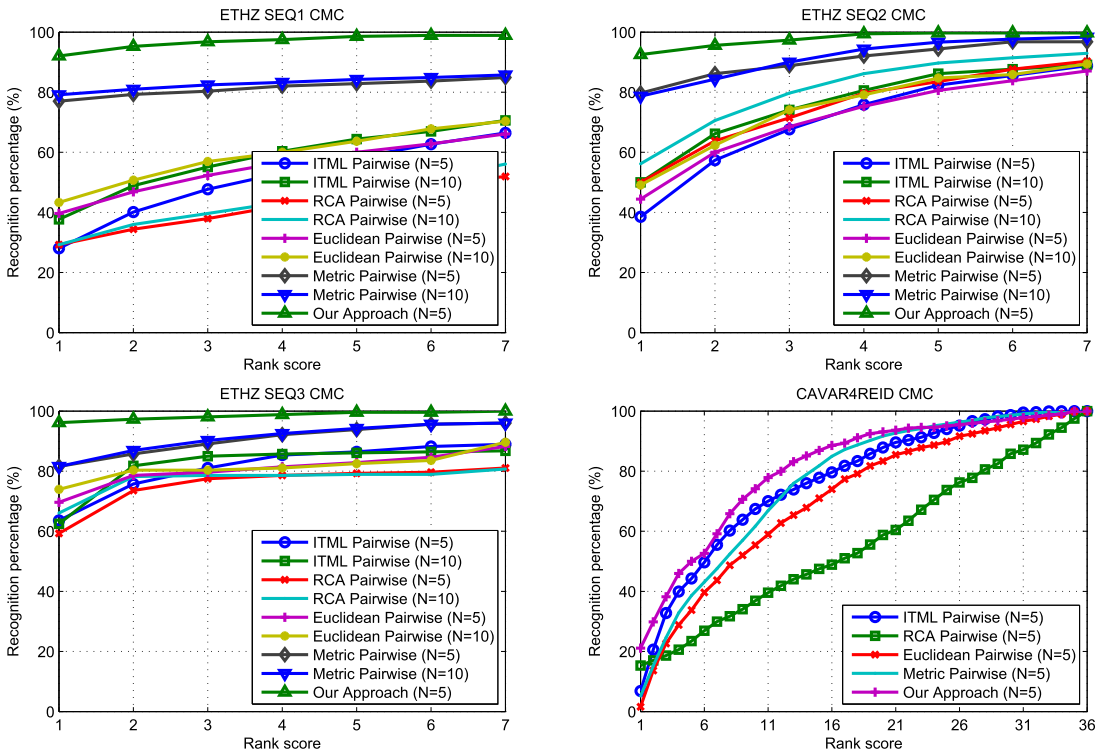


Fig. 6 Comparison between our approach (k-NN distance) and the pairwise distance of ITML, RCA, Euclidean and the distance metric learned in offline training phase with $N = \{5, 10\}$ on Seq. 1 – Seq. 3 of ETHZ (first three figures) and CAVIAR4REID (the last figure).

tance comparison method outperforms the pairwise method greatly. This proves our local distance comparison is superior to the pairwise distance comparison, and also supports our assumption of multimodal distribution properties for people appearance.

Finally, we compare our approach with SDALF [7] and PLS [15] to see the pros and cons of our approach. We also compare the proposed method with two learning based distances: ITML and RCA, and two non-learning based distances: Euclidean and Bhatachayya. The above four distances are used in our loss function. The comparison is shown in Fig. 7. Some selected results are shown in Fig. 10.

By using only the low level features, our approach is still competitive with SDALF, which encodes the high level spatial information into feature representation but does not pay attention to the inter-/intra-variance issue. PLS learns a people appearance model based on the gallery image sets. In the multiple-shot scenario, when the appearance instances of the same people are similar, the PLS works well (as in ETHZ dataset, Fig. 7). However, due to the nature of being sensitive to learning data, PLS cannot adapt to the situation where the appearance of same people varies greatly (as in CAVIAR4REID, Fig. 7).

Compared with Euclidean or Bhattacharyya distance, the performance of our approach shows clear improvement over these baseline methods. Particularly in the Seq. 1 – Seq. 3 of ETHZ dataset, rank-one performance of our approach outperforms at least 10% against Euclidean or Bhattacharyya distance. It also suggests the importance and effectiveness of distance metric learning. This enables us to be confident that under k-NN framework, the use of the distance metric learned by LMNN [22] is suitable for multimodal properties of people appearance. By contrast, ITML and RCA try to shrink distances between all appearance instances of the same people, while a large amount of complicated appearances introduced by using multiple shots actually further enlarge such difference or bias in both probe and gallery sets. Since the strategy of shrinking intra-variance is not suitable for the multimodal property of appearance distribution, they increase the re-identification error rate and eventually fail on these datasets.

4.5 Parameters Discussion

In this section, we study how different parameter settings could affect the performance of the proposed approach.

In the proposed approach, the most important parameter is the k in local distance comparison (i.e., k in Eq. (3) – (6)). To confirm its effect, we conduct the local distance comparison with different k values. The CMC curves of the proposed approach, with k setting to $\{1, 2, \ldots, 10\}$, are plotted in Fig. 8. To show an intuitive overview of k's effect, we also summarize the average area under curve (AUC) for each setting. From the result, we can observe that: though the effect of k shows slightly different trend on different datasets, when k is set to a small value (1 – 3), the proposed

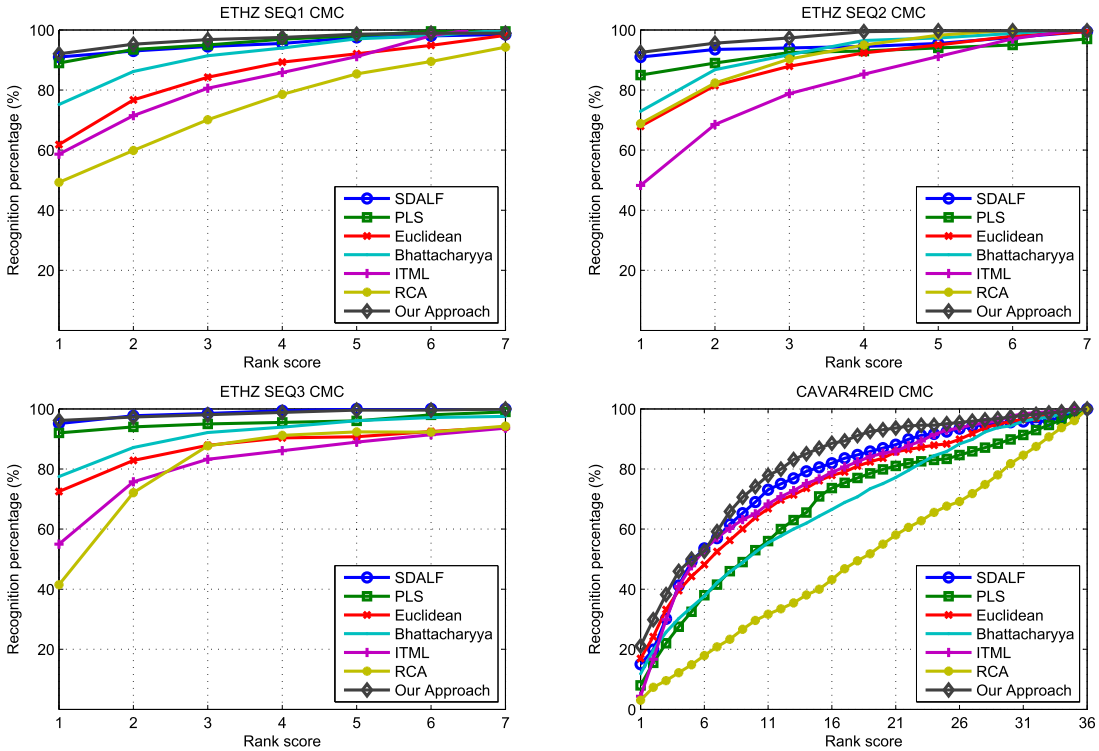


Fig. 7 Comparing SDALF, PLS approaches, local distances measured by Euclidean, Bhattacharyya, ITML and RCA metric to our approach (k-NN distance) with $N = \{5\}$ on Seq. 1 – Seq. 3 of ETHZ (first three figures) and CAVIAR4REID (the last figure).

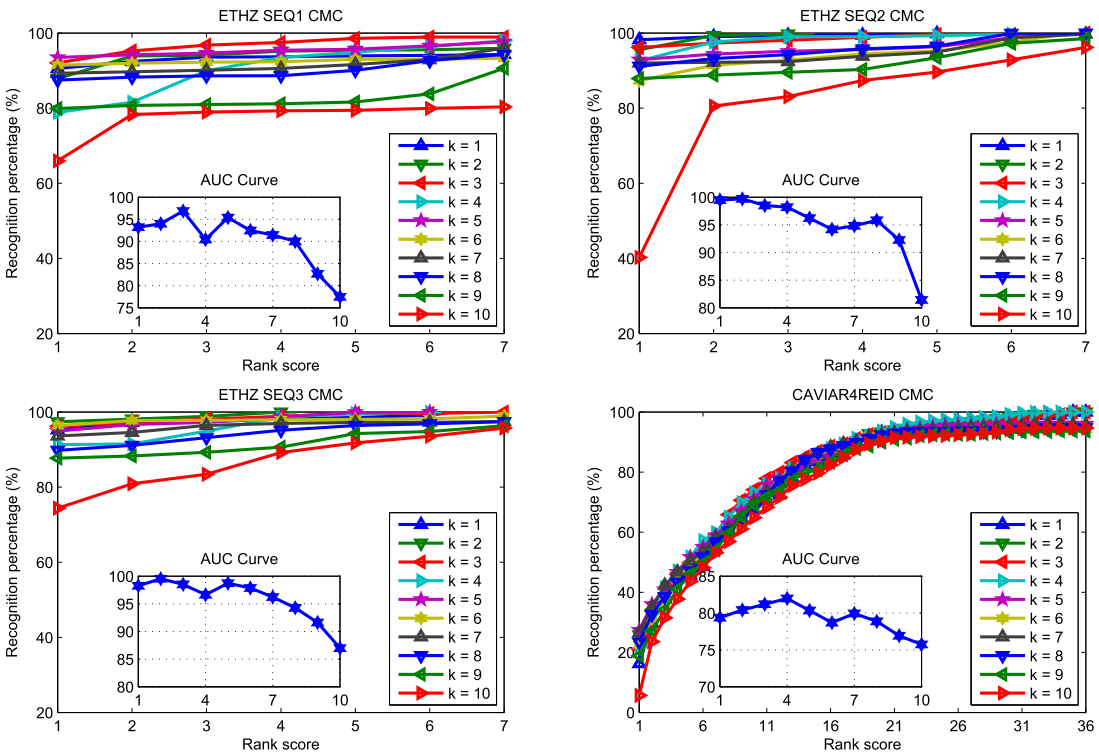


Fig. 8 Evaluation of CMC curve for Seq. 1 – Seq. 3 of ETHZ and CAVIAR4REID with different k values in local distance comparison, and the AUC curves of CMC with different k values are plotted to show overall performance.

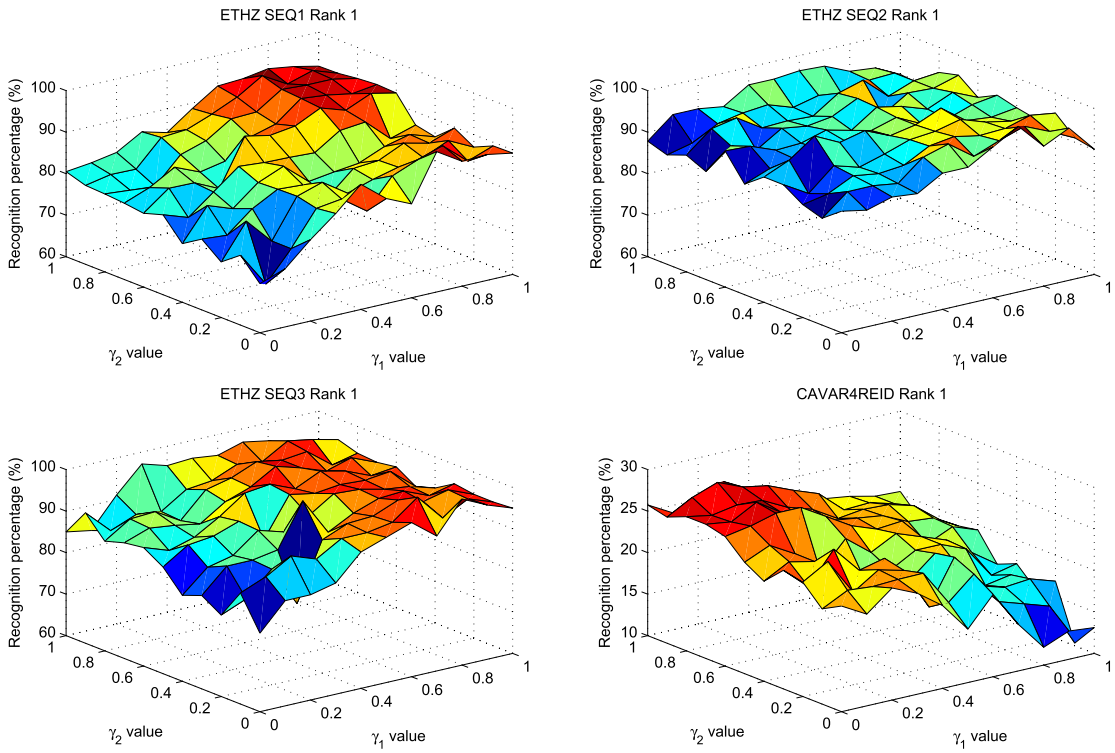


Fig. 9 Evaluation of rank 1 performance of local distance comparison for Seq. 1 – Seq. 3 of ETHZ and CAVIAR4REID with different γ_1 and γ_2 values in loss function.

approach achieves high performance overall. Such results reflect that the local neighbors are very useful for comparison. For the reason of different trends of k shown in AUCs, in our opinion, this is mainly due to the differences between datasets. It suggests that the optimal k should be determined by considering the dataset properties either. Figure 8 also shows an interesting phenomenon: after reaching its best performance, AUC suffers a judder and then shows a signification dip. This occurs because as the k value increases, the local neighbors measured in Eq. (3) are not always the 'true' local neighbors; they are sometimes from different local subsets. This makes the loss function, which is designed under the assumption of local aggregation, lose its power.

There are two other important parameters, λ_1 and λ_2, in the proposed approach. These two parameters control the relative importance of invasion in the loss function. In order to evaluate their effect to the approach, we conduct another experiment by simultaneously changing their values from 0 to 1 with a step of 0.1. The rank 1 results of the proposed approach on the four datasets are reported in Fig. 9. From the results, we can observe that the rank 1 rates show high performance when λ_1 and λ_2 are large in Seq. 1 – Seq. 3 of ETHZ; however for the CAVIAR4REID dataset, the rank 1 is better when λ_1 is small and λ_2 is large. The two invasions, from C_{pA} to C_e and from C_e to C_{pA}, vary with different datasets. The optimal parameters could balance the two terms to deal with the invasions in the local neighbors, and consequently achieve better performance. A efficient approach to evaluate γ_1 and γ_2 is using cross validation method on the training data.

5. Conclusion

In this paper, we proposed a novel approach for multiple-shot people re-identification under local distance comparison framework. Unlike single-shot re-identification problem, due to much more change of illumination, camera view, non-rigid body deformation and so on, we face to deal with a critical inter-variance and intra-variance issue introduced by using multiple images of people appearances.

To deal with this issue, we regard the multiple-shot people re-identification task as a local distance comparison problem, which is solved by comparing the subsets associated with the same semantic meaning on a multimodal appearance distribution, based on a novel local distance-based loss function. The proposed loss function is particularly designed to adapt to the multimodal properties of people appearance distribution, which not only favors small distances that indicate high similarity between appearances of the same people, but also penalizes small distances or undesirable overlaps between subsets, which reflect high similarity between appearances of different people. The re-identification result of the probe is given by finding the minimal loss with the potential target in the gallery.

The proposed approach has been evaluated on four public benchmark datasets that are widely used in currently re-identification literature. Experimental results show that our approach achieve great improvements compared

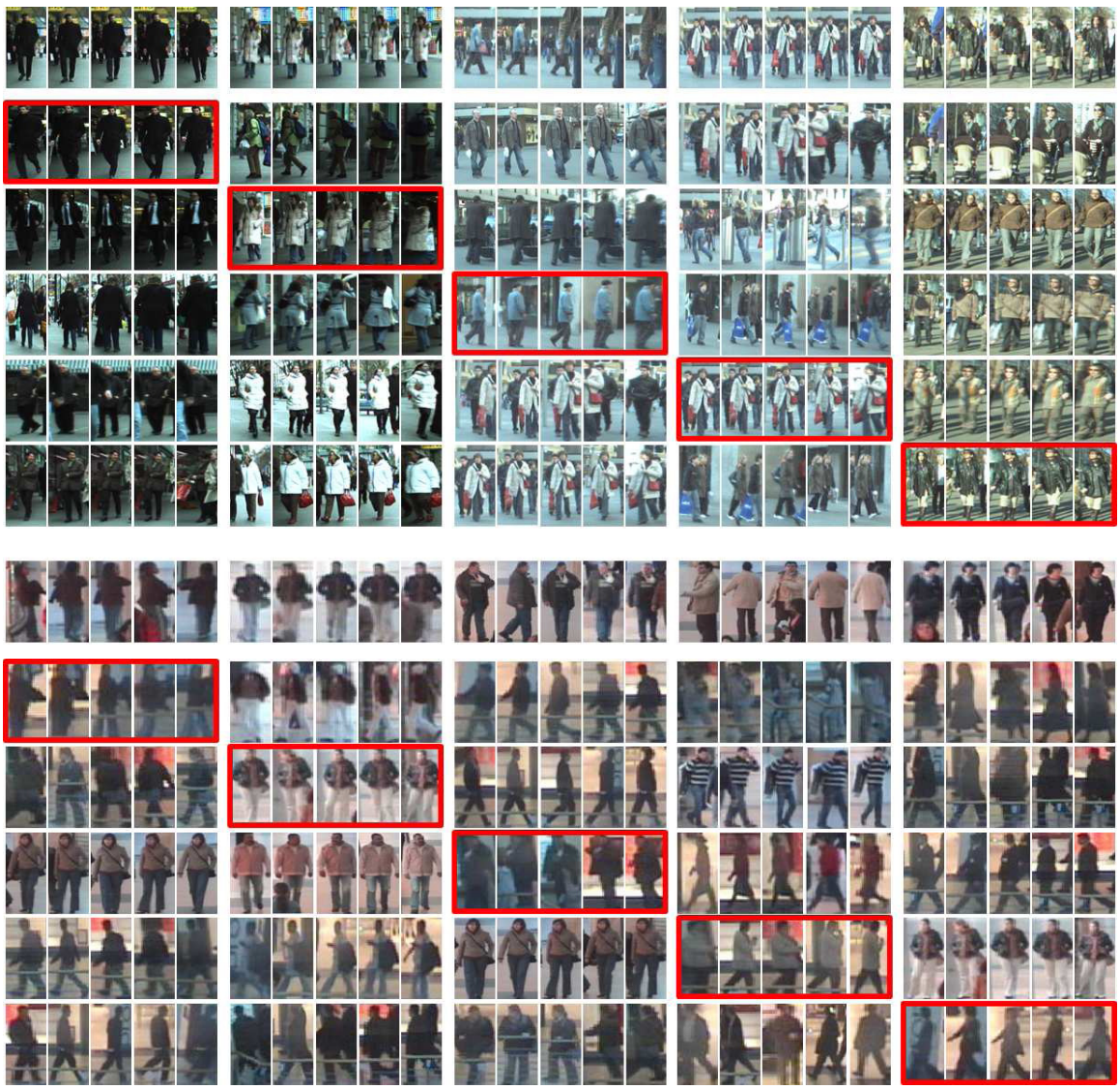


Fig. 10 Selected results of re-identification obtained by using our approach ($N = 5$) for ETHZ and CAVIAR4REID. The top row shows probe persons, while the bottom part shows the top 5 matching results. Corrects are highlighted by red.

with existing methods. With detailed analysis of the results, we come to the conclusion that the strategy of local distance comparison in multimodal appearance distribution can effectively alleviate so-called inter-variance and intra-variance issue. In the future work, we intend to customize our approach for use in online re-identification applications.

Acknowledgments

This research is supported by the Strategic Information and Communications R&D Promotion Programme No. 131306004, the Grant-in-Aid for Scientific Research (B) No.26280057 and the Grant-in-Aid for Challenging Exploratory Research No. 26540081. Yu Wang is supported by the JSPS Postdoctoral Fellowship for Foreign Researchers.

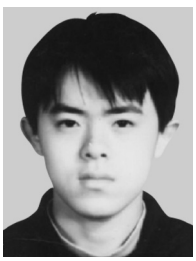
References

[1] G. Doretto, T. Sebastian, P. Tu, and J. Rittscher, "Appearance-based person reidentification in camera networks: problem overview and current approaches," J. Ambient Intelligence and Humanized Computing, vol.2, no.2, pp.127–151, 2010.
[2] W.S. Zheng, S. Gong, and T. Xiang, "Re-identification by relative distance comparison," IEEE Trans. Pattern Anal. Mach. Intell., vol.99, no.3, pp.653–668, March 2013.
[3] W.S. Zheng, S. Gong, and T. Xiang, "Person re-identification by probabilistic relative distance comparison," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp.649–656, June 2011.
[4] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu, "Shape and appearance context modeling," Proc. Int'l Conf. Computer Vision, pp.1–8, 2007.
[5] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," Proc. European Conf. Computer Vision, pp.262–275, 2008.

[6] B. Prosser, W.S. Zheng, S. Gong, and T. Xiang, "Person re-identification by support vector ranking," Proc. British Machine Vision Conf., pp.21.1–21.11, 2010.
[7] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp.2360–2367, June 2010.
[8] M. Dikmen, E. Akbas, T.S. Huang, and N. Ahuja, "Pedestrian recognition with a learned metric," Proc. Asia Conf. Computer Vision, pp.501–512, 2010.
[9] D.S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, "Custom pictorial structures for re-identification," Proc. British Machine Vision Conf., pp.68.1–68.11, Sept. 2011.
[10] L. Bazzani, M. Cristani, A. Perina, and V. Murino, "Multiple-shot person re-identification by chromatic and epitomic analyses," Pattern Recognit. Lett., vol.33, no.7, pp.898–903, May 2012.
[11] L. Bazzani, M. Cristani, A. Perina, M. Farenzena, and V. Murino, "Multiple-shot person re-identification by HPE signature," Proc. Int'l Conf. Pattern Recognition, pp.1413–1416, Aug. 2010.
[12] D.N.T. Cong, C. Achard, L. Khoudour, and L. Douadi, "Video sequences association for people re-identification across multiple non-overlapping cameras," Proc. Int'l Conf. Image Analysis and Processing, pp.179–189, Aug. 2009.
[13] A. Ess, B. Leibe, and L. van Gool, "Depth and appearance for mobile scene analysis," Proc. Int'l Conf. Computer Vision, pp.1–8, Oct. 2007.
[14] L.V.D. Maaten and G. Hinton, "Visualizing data using t-SNE," J. Machine Learning Research, vol.9, pp.2579–2605, Nov. 2008.
[15] W.R. Schwartz and L.S. Davis, "Learning discriminative appearance-based models using partial least squares," Proc. 2009 XXII Brazilian Symposium on Computer Graphics and Image Processing, pp.322–329, 2009.
[16] W. Li, R. Zhao, and X. Wang, "Human reidentification with transferred metric learning," Proc. Asia Conf. Computer Vision, pp.31–44, Nov. 2012.
[17] W. Li, Y. Wu, M. Mukunoki, and M. Minoh, "Bi-level relative information analysis for multiple-shot person re-identification," IEICE, Trans. Inf. & Syste. vol.E96-D, no.11, pp.2450–2461, Nov. 2013.
[18] E.P. Xing, A.Y. Ng, M.I. Jordan, and S. Russell, "Distance metric learning, with application to clustering with side-information," Proc. Advances in Neural Information Processing Systems 15, pp.505–512, 2002.
[19] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall, "Learning a Mahalanobis metric from equivalence constraints," J. Machine Learning Research, vol.6, pp.937–965, Dec. 2005.
[20] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov, "Neighbourhood components analysis," Proc. Advances in Neural Information Processing System 17, pp.513–520, 2004.
[21] J.V. Davis, B. Kulis, P. Jain, S. Sra, and I.S. Dhillon, "Information-theoretic metric learning," Proc. Int'l Conf. Machine Learning, pp.209–216, 2007.
[22] K.Q. Weinberger and L.K. Saul, "Distance metric learning for large margin nearest neighbor classification," J. Machine Learning Research, vol.10, pp.207–244, June 2009.
[23] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," Proc. IEEE Conf. Computer Vision and Pattern Recognition, vol.1, pp.539–546, June 2005.
[24] E.F.C. project/IST 2001 37540
[25] R.O. Duda, P.E. Hart, and D.G. Stork, Pattern Classification, 2nd ed., Wiley & Sons, New York, NY, USA, 2001.
[26] M. Maier, M. Hein, and U. von Luxburg, "Optimal construction of k-nearest-neighbor graphs for identifying noisy clusters," Theor. Comput. Sci., vol.410, no.19, pp.1749–1764, 2009.

Guanwen Zhang received the B.S. and M.S. degree in Computer Science and Technology from Northwestern Polytechnical University, China, in 2007 and 2010 respectively. He is currently a Ph.D. student at the Graduate School of Information Science, Nagoya University, Japan. His research interests include computer vision, pattern recognition, and machine learning.

Jien Kato received the M.E. and Ph.D. degrees in Information Engineering from Nagoya University in 1990 and 1993, respectively. She is currently Associate Professor at the Graduate School of Information Science, Nagoya University. Her research interests include computer vision, machine learning, multi-sensor perceptual computing, and their applications. She is a member of IEICE, IPSJ, and a senior member of IEEE.



Yu Wang received the MS degree in Information Science and PhD degree in Engineering, from Nagoya University, in 2010 and 2013, respectively. He is currently a postdoctoral researcher with the Graduate School of Information Science, Nagoya University. His research interests are object recognition and visual event categorization. He is a member of the IEEE.

Kenji Mase received the B.S. degree in Electrical Engineering and the M.S. and Ph.D. degrees in Information Engineering from Nagoya University in 1979, 1981, and 1992, respectively. He became a professor of Nagoya University in August 2002. He is now with the Graduate School of Information Science, Nagoya University. He joined the Nippon Telegraph and Telephone Corporation (NTT) in 1981 and had been with the NTT Human Interface Laboratories. He was a visiting researcher at the Media Laboratory, MIT, in 1988-1989. He has been with the Advanced Telecommunications Research Institute (ATR) in 1995-2002. His research interests include gesture recognition, computer graphics, artificial intelligence and their applications for computer-aided communications. He is a member of IPSJ, JSAI, VRSJ, HISJ, and ACM, a senior member of IEEE Computer Society, and a fellow of IEICE.