Pattern Recognition 67 (2017) 287–298

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Person re-identification by multiple instance metric learning with impostor rejection

Xiaokai Liu, Hongyu Wang*, Jie Wang, Xiaorui Ma

School of Information and Communication Engineering, Dalian University of Technology, Dalian, 116024, PR China

ARTICLE INFO

Article history:
Received 24 October 2016
Revised 4 January 2017
Accepted 9 February 2017
Available online 20 February 2017

Keywords:
Person re-identification
Graphical model
Multiple instance metric learning
Impostor rejection

ABSTRACT

Due to its ability to eliminate the visual ambiguities in single-shot algorithms, video-based person re-identification has received an increasing focus in computer vision. Visual ambiguities caused by variations in view angle, lighting, and occlusions make the re-identification problem extremely challenging. To overcome the ambiguities, most previous approaches often extract robust feature representations or learn a sophisticated feature transformation. However, most of these approaches ignore the effect of the impostors arising from annotation or tracking process. In this case, impostors are regarded as genuine and applied in training process, leading to the model drift problem. In order to reduce the risk of model drifting, we propose to automatically discover impostors in a multiple instance metric learning framework. Specifically, we propose a *k*NN based confidence score to evaluate how much an impostor invades the interested target and utilize it as a prior in the framework. In the meanwhile, we integrate an impostor rejection mechanism in the multiple instance metric learning framework to automatically discover impostors, and learn the semantical similarity metrics with the refined training set. Experiments show that the proposed system performs favorably against the state-of-the-art algorithms on two challenging datasets (iLIDS-VID and PRID 2011). We have improved the rank 1 recognition rate on iLIDS-VID and PRID 2011 dataset by 1.0% and 1.2%, respectively.

1. Introduction

Person re-identification aims to match a probe person against a set of gallery persons over different non-overlapping camera views, without imposing any constraints on spatial or temporal continuity. One-shot algorithms [1–3] have been highly developed in the past few years. The state-of-the-art algorithms mostly develop invariant and discriminative feature representations [2–5] or exploiting view-to-view similarity transformation strategies [6–8]. However, the high dimensional variables, such as sharp illumination changes, severe view and pose changes, complex environment and heavy occlusions, make single-shot re-identification problem ill-posed and ambiguous. Therefore, video based algorithms have been developed to reduce the visual ambiguities by exploring highly discriminative appearance features [9] or space-time information [10].

Although great progress has been achieved, three main problems remain unsolved. First, lighting conditions are always complex, and may undergo rapid changes, as shown in Fig. 1(a). Average operation will diminish the ability to distinguish from others, and 'best' strategies would be affected by inferior examples. In this case, how to pick up discriminative fragments for training process? Second, when a person undergoes heavy occlusion, the occluders tend to be regarded as interested targets. Take the situation in Fig. 1(b) for example, the person in black may be taken as the target if chosen as the most discriminative fragment, and matched with another person in black from another camera with high probability. In this case, how to reduce the negative impact of such impostor images? Last, in recent researches, Mahalanobis metric learning [8,11] is proved to be effective in improving the re-identification performance. In the video-based re-identification problem, all the labels are given in bag level, which means we only know the persons' IDs, but not the real matched fragments. However, the metric learning approaches need instance level labels to learn the linear transformations. Therefore, directly applying metric learning algorithms to get a view-to-view metric is infeasible. In this case, how to obtain a proper Mahalanobis metric in the video based situation?

In this paper, we aim to address the aforementioned problems using a multiple instance metric learning framework to automatically select discriminative fragments, discover impostors and learn the linear transformation from source view to target view. To this end, we first construct a tree-structured graphical model to ex-

[*] Corresponding author.
E-mail address: whyu@dlut.edu.cn (H. Wang).

http://dx.doi.org/10.1016/j.patcog.2017.02.015
0031-3203/© 2017 Elsevier Ltd. All rights reserved.

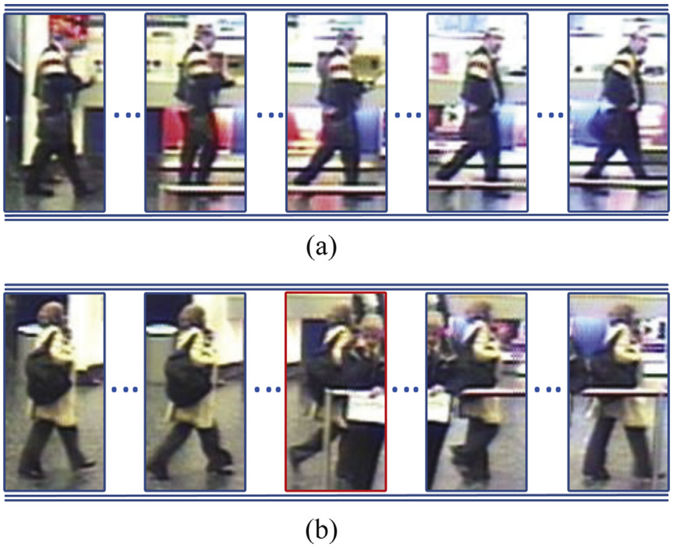


Fig. 1. Sample images from two example videos in the iLIDS-VID dataset. Video (a) undergoes rapid illumination changes during the video capture period. In video (b), the woman in khaki is heavily occluded by the man in black, marked in a red box. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

ploit the intrinsic structure of the input video and generate bags of fragment hypotheses, in which we are intended to include all potentially discriminative fragments. In order to construct an optimal tree in factorial growth searching space, we adopt data-driven state transition proposals and formulate the generation of tree-structured model in a simulated tempering framework, which are efficient in avoiding being trapped in a local minimum by heating up the distribution repeatedly. Considering the ambiguities arising from heavy occlusions and rapidly changing lighting conditions, first we apply a *k*-nearest neighbor (*k*NN) based anomaly score to evaluate the appearance variation along a video fragment sequence. Impostors are assigned less weights to reduce their impact to metric learning. Second, we propose a jump cost measurement to heuristically adjust the labels of the instances. Then both *k*NN based anomaly score and jump cost measurement are integrated into a unified multiple instance multiple label logistic discriminant metric learning (MiMl-DML) [12] framework. MiMl-DML iteratively optimizes by impostor estimation and updates of the logistic discriminant modeled metric, thereby obtains instance level labels and semantical similarity metrics with refined training samples.

Main contributions of our study are summarized below:

- We come up with a measure of possible impostors, jump cost, so that we can evaluate the abnormal degree of an impostor and identify ones arising from annotation or tracking process.
- Taking into account the impostor rejection mechanism, we propose a unified MiMl-DML-IR framework. MiMl-DML-IR iterates between updates of the metric and selection of putative impostors from positive pairs of bags, thereby automatically discovering impostors and obtaining transition metrics with refined samples.
- We propose a novel approach to generate a set of potentially discriminative video fragments based on a tree-structured graphical model. For efficiency, we formulate the task of optimizing the tree-structure in a simulated tempering framework. Data-driven state transition proposals are proposed to help the algorithm converge rapidly across both state and temperature space.

2. Related work

2.1. Multi-shot/video based re-identification

Numerous features [13–15] have been proposed to obtain a discriminative appearance descriptor for the multi-shot re-identification problem. In order to integrate complementary global and local statistical human descriptions, Bazzani et al. [13] extract a highly informative signature histogram plus epitome, which focuses on overall chromatic content. Bak et al. [14] combine information from sequential images and obtain the mean Riemannian covariance grid descriptor. Bedagkar-Gala and Shah [15] combine the characteristic appearance and the appearance variations statistics to enhance the feature description.

Training based algorithms are also developed to obtain an appearance model from image sets [16] or learn a locally aligned feature transformation [17]. Liu et al. [18] develop a deep non-linear metric learning approach based on neighborhood component analysis and deep belief network to overcome the limitations of traditional linear metric learning. For one person, if a longer video sequential is collected (generally tens to hundreds), behavioral biometrics such as gait [19,20] can be used for matching, while it is unpractical for human re-identification task due to the resolution or frame rate constraints of typical cameras. Wang et al. [10] utilize a reliable space-time features to exploit intrinsic motion properties for pedestrians. Karanam et al. [21] propose to train a viewpoint invariant dictionary to discriminatively encode feature descriptors representing different people. By abstracting patches on different sales and exploring the relationship between those patches, Pribadi et al. [22] propose a sparse tree-structured image representation to solve the re-identification problem.

2.2. Metric learning

In recent years, Mahalanobis metric learning has attracted a considerable interest for person re-identification. The main idea is to seek an optimal metric that reflects the visual view-to-view transitions, allowing for a more powerful classification. In [8], a large number of Mahalanobis metric learning algorithms have been evaluated and shown to be effective in re-identification problem, for example, linear discriminant metric learning (LDML), information theoretic metric learning (ITML), large margin nearest neighbor(LMNN), large margin nearest neighbor with rejection (LMNN-R), and keep it simple and straightforward metric learning (KISSME). Several task-oriented approaches are exploited to address the ill-posed problems arising from re-identification. By ignoring easy samples and focusing on hard samples, Hirzer et al. [23] propose an impostor-based LMNN, exploiting the natural constraints given by the person re-identification task. Hirzer et al. [11] develop a relaxed pairwise learned metric to reduce the computational effort. As an extension to the aforementioned linear metrics, Xiong et al. [24] evaluate four kernel-based distance learning approaches to improve re-identification ranking accuracy when the data space is under-sampled. Considering neighborhood structure manifold which exploits the relative relationship between the concerned samples and their neighbors in the feature space, Li et al. [25] propose a neighborhood structure metric learning algorithm to learn discriminative dissimilarities on such manifold.

Although metric learning has been proven to be effective in the field of computer vision, a large number of applications have difficulty in using the metric method due to the limitation of insufficient or incomplete data annotations. In order to address the problem of insufficient data, several approaches are proposed. Under the guidance of weak supervisory information, Wang [26] proposes a semi-supervised metric learning algorithm to train the data with pair-wise constraint. Bilenko et al. [27] introduce a metric-based

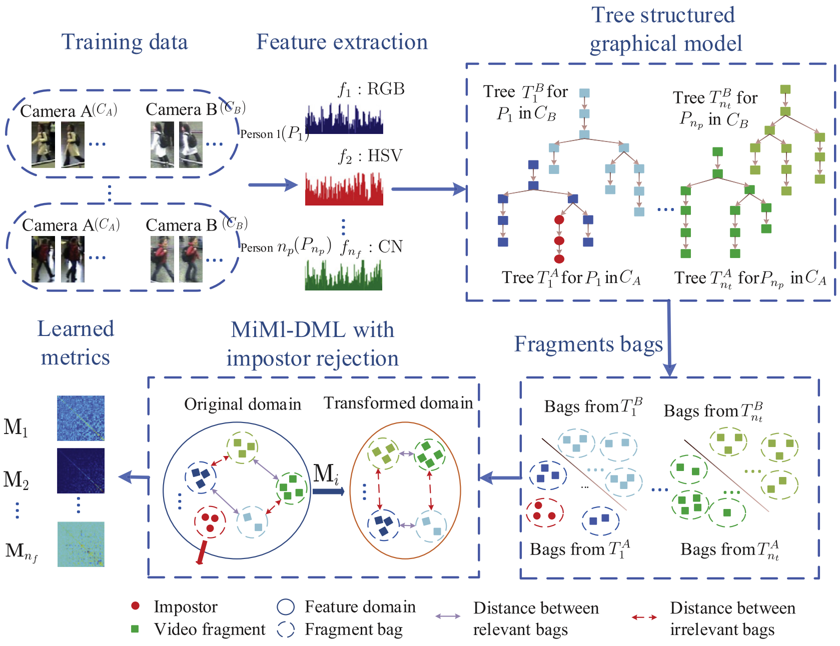


Fig. 2. An overview of the proposed metric training algorithm. Features from eight models are first extracted from each person, then a tree-structured graphical model is applied to generate a set of discriminative video fragments. Finally, metrics based on different features are learned by the MiMI-DML-IR algorithm.

semi-supervised clustering algorithm and learn individual metrics for each cluster. In many cases, we are unable to obtain the fully-annotated information, such as many multi-instance tasks. For example, drug activity recognition, image annotation [28], and face verification [12] are such cases. Xu et al. [28] propose a multiple-instance metric learning framework to obtain a distance metric under the multi-instance setting by minimizing the KL divergence between two multivariate Gaussians.

2.3. Tree-structured graphical model

In the computer vision community, linear structure model may encounter a dilemma due to the complexity of natural scenes. Tree-structured model is therefore proposed to exploit the intrinsic structures of the visual patterns in different tasks. Shi et al. [29] introduce a part-based tree-structured model for text recognition, by making use of the character-specific structure information and the local appearance information. Adams et al. [30] infer a tree-structured model from latently hierarchical data and apply it to hierarchical clustering of images and topic modeling of text data. Bart et al. [31] present a tree-shaped hierarchy to discover the visual taxonomies. Kwon et al. [32] formulate the event summarization and rare event detection problem in a tree-structured graph editing model.

3. Learning tree structured model by simulated tempering

A diagram that summarizes the proposed training algorithm is depicted in Fig. 2. After extracting feature descriptors, a tree-structured model is applied to generate a set of discriminative video fragments (Section 3). Then metrics based on different features are learned by the MiMI-DML with impostor rejection approach (Section 4).

Along the video sequence, adverse conditions like varying lighting conditions, severe pose changes and heavy occlusions, make the sequence multi-modal, locally similar while differs in a long run, as shown in Fig. 1. In this section, we aim to construct a fragment pool, including potentially discriminative video fragments in different lengths. In this case, multiple branches can increase the chance of discovering discriminative video fragments.

3.1. Problem definition

Let $\mathbf{G} = (\mathcal{V}, \mathcal{E})$ be a tree-structured graph, where $\mathcal{V} = \{v_1, \ldots, v_N\}$ denotes a set of nodes corresponding to N frames and $\mathcal{E} = \{e_1, \ldots, e_M\}$ denotes a set of directed edges defining conditional dependency between frames. Let $f(v_i)$ denotes the feature descriptor generated by node v_i, and $f^B(v_i)$ denotes the average feature descriptor of the branch B_t, which the node v_i belongs to. Given an input video, we ideally aim to generate an optimal tree-structure $\hat{\mathbf{G}}$ over the state of the tree $\mathbf{x}$ based on individual edge reliabilities, so that (a) instances in the same branch are captured with similar pose, under similar lighting conditions and closely in time; (b) the number of the branches is limited. Request (a) makes sure that each branch represents a specific state of the person in one view and seeks to match the corresponding state in another view. Request (b) makes sure that the state of the person is limited and sufficient instances are included in each state. The quality of the tree-structure is defined by an energy function

$$E(\mathbf{v}) = \sum_{t=1}^{n_t} \sum_{(v_i,v_{i+1})\in B_t} ||f(v_i) - f(v_{i+1})||^2 + \lambda \sum_{(v_j,v_j)\in\mathcal{E}} ||f^B(v_i) - f^B(v_j)|| \qquad (1)$$

where (v_i, v_{i+1}) is a pair of adjacent nodes in branch B_t, and n_t is the number of the branches. The first term encourages the adjacent nodes in the same branch to be as similar as possible. The second term is a well known sparsity regularization and encourages several of the terms $||f^B(v_i) - f^B(v_j)||$ to be exactly zero. Equivalently, several of the nodes will belong to the same branch, efficiently reducing the number of branches. The regularization parameter λ is a user choice that will control the tradeoff between model fit and the number of branches. Suppose one node could only be linked with the node which is temporally ahead. The searching complexity is factorial level. Apparently, the problem is NP hard. In order to find the optimal solution in limited number of iterations, we apply a Wang-Landau adapted simulated tempering algorithm.

3.2. Wang–Landau adapted simulated tempering

The problem of finding the minimum of the energy E is equivalent to finding the maximum of the variant function $\pi(\mathbf{x}) = \exp\{-E(\mathbf{x})/T\}$(as long as $T > 0$). This distribution function becomes 'flatter' as the temperature T increases and becomes 'spikier' as T decreases towards 0. When $T \to 0$, all of the probability mass will be concentrated around the global minimum of $E(\mathbf{x})$. However, sample directly from $\pi_k(\mathbf{x})$ with a small T would encounter 'low probability barriers', and get stuck in a local minimum. So we apply Simulated Tempering (ST) algorithm to help bridge parts of the sample space which are separated by low probability (or high energy) by constructing a few 'companion chains'. The intuition behind ST is that by heating up the distribution repeatedly, the new sampler can escape from local modes and increase its chance to reach the main body of the distribution.

We first construct a family of distributions $\Pi = \{\pi_k(\mathbf{x})\ \ k \in K\}$, by varying a single parameter, the temperature T_k, in the target distribution π, that is $\pi_k(\mathbf{x}) \propto \exp\{-E(\mathbf{x})/T_k\}$. Let $T_1 > T_2 > \ldots > T_k > \ldots > T_K$ be a sequence of monotone decreasing temperatures in which T_1 is reasonably large and $\lim_{k\to\infty} T_k = 0$. The original target distribution π corresponds to the member of this family with the lowest temperature. Therefore, the joint distribution $p(\mathbf{x}, k)$ could be defined on the augmented space $(\mathbf{x}, k) \in \mathcal{X} \times \mathcal{K}$

$$p(\mathbf{x}, k) = w_k \exp\left(-E(\mathbf{x})/T_k\right) \tag{2}$$

where w_k are tunable parameters.

In ST algorithm, simulating from the joint $p(\mathbf{x}, k)$ is done by iterating through two transition operators alternately:

- With the current state $(\mathbf{x}^n, k^n)$, we draw a sample from the uniform distribution $u = U(0, 1)$;
- If $u \leq \alpha_0$, we draw a new sample $\mathbf{x}^{n+1}$ from a transition distribution $\Gamma_k(\mathbf{x}^n|\mathbf{x}^{n+1})$ and keep the temperature unchanged; If $u > \alpha_0$, we keep sample $\mathbf{x}$ unchanged as $\mathbf{x}^{n+1} = \mathbf{x}^n$ and make a transition $k^n \to k^{n+1}$, and accept the transition with probability

$$\alpha = \min\left(1, \frac{p(\mathbf{x}^{n+1}, k^{n+1})q(k^{n+1} \to k^n)w_{k^n}}{p(\mathbf{x}^{n+1}, k^n)q(k^n \to k^{n+1})w_{k^{n+1}}}\right) \tag{3}$$

where $q(k^t \to k^{t+1})$ is defined $q(i \to i+1) = q(i \to i-1) = 1/2$, and $q(1 \to K) = q(K \to 1) = 1$.

In ST algorithm, the weight parameters w_k should be tuned so that each tempered distribution in the system have a roughly equal chance to be visited. The Wang-Landau (WL) algorithm [33] is proposed to determine the weight dynamically. In WL algorithm, we partition the state space into K sets $\{k\} \cup \mathcal{X}$, each corresponding to a different temperature value. If the move into a different temperature value is rejected, the adaptive weight for the current partition will increase, thus exponentially increasing the probability of being accepted in the next move. The detailed algorithm is described in Algorithm 1, where $\mathbf{I}$ is an indicator function, and $\gamma > 0$ denotes the weight adapting factor.

3.3. Data-driven state transition proposal

Given the current state of the structured tree $\mathbf{x}^n$, the aim of the proposal distribution is to provide a guided sampling $\mathbf{x}^{n+1}$ from $\mathbf{x}^n$. With reference to the calculation of the acceptance rate in Algorithm 1, the probability of transition from $\mathbf{x}^n$ to $\mathbf{x}^{n+1}$ as well as the reverse step needs to be computed. We propose a data-driven proposal distribution that accomplish this task in an efficient manner so that the Simulated Tempering algorithm rapidly converges across both state and temperature space.

Algorithm 1 The Wang–Landau adapted Simulated Tempering Algorithm.

```
Initializing the every element of the adaptive weights $\{w_k\}_{k=1}^{K}$ equals $1/K$.
for n=1:$N_T$ (number of iterations) do
    Given current state $(\mathbf{x}^n, k^n)$, we sample $u = U(0, 1)$.
    if $u \leq \alpha_0$ then
        let $k^{n+1} = k$ and draw $\mathbf{x}^{n+1}$ from a transition distribution $\Gamma_k(\mathbf{x}^n|\mathbf{x}^{n+1})$.
    else if $u > \alpha_0$ then
        let $\mathbf{x}^{n+1} = \mathbf{x}^n$ and make transition $k^n \to k^{n+1}$, and accept the transition with probability
        $\alpha = \min\left(1, \frac{p(\mathbf{x}^{n+1}, k^{n+1})q(k^{n+1} \to k^n)w_{k^n}}{p(\mathbf{x}^{n+1}, k^n)q(k^n \to k^{n+1})w_{k^{n+1}}}\right)$    (4)
    end if
    Update adaptive weight
    $w_{k^{n+1}} = w_{k^n}(1 + \gamma\mathbf{I}(k^{n+1} \in \{i\})), i = 1, \ldots, K.$
end for
```

Differences of many adjacent frames are quite small. In order to decrease the computation complexity, we restrict those edges with high similarity to be fixed. We associate the images into tracklets when they are in the consecutive frames and similar enough in appearance and regard the edges between adjacent frames within the same tracklet as fixed. We merge the pair of nodes on both sides of the fixed edge into one super-node, in the following steps, we regard the super-node as normal node for simplicity. For those edges with uncertainty, they are regarded as variable, and the link relationship can be changed. To better describe the proposed algorithm, we specify the following definition. Denoting a node in time t by v_t, we define the node which is ahead of v_t as *father node*, denoted $N_F(v_t) = v_{t-1}$. For all the nodes ahead of v_t, we call them *ancestor nodes*, denoted $N_A(v_t) = \{v_i\}_{i=1}^{t-1}$. We constrain that any node could only be linked with its temporal ancestor nodes. The edge set is constructed by any node and its temporal ancestors, denoted by $\{(v_t, \phi(N_A(v_t)))\}_{t=1}^{T}$, where $\phi(N_A(\cdot))$ indicates any element in $N_A(\cdot)$. The weights of the edges e_i are defined by the similarities between each related node pair (v_{i1}, v_{i2}), denoted by $Sim(v_{i1}, v_{i2})$. We set the initial state of the tree structure $\mathbf{x}^0$ to the conventional chain model in temporal order. Given a current state of tree-structured model $\mathbf{x}^n$, a new tree sample $\mathbf{x}^{n+1}$ is proposed from proposal distribution $\Gamma(\mathbf{x}^{n+1}|\mathbf{x}^n)$. For simplicity, the proposal space is restricted to a single edge *delete-and-create* operation and defined as the probability to delete an original edge e_i by the probability to create a new edge e'_i

$$\Gamma(\mathbf{x}^{n+1}|\mathbf{x}^n; e_i, e'_i) = P_{delete}(\mathbf{x}^n, e_i) \cdot P_{create}(\mathbf{x}^{n+1}, e'_i) \tag{5}$$

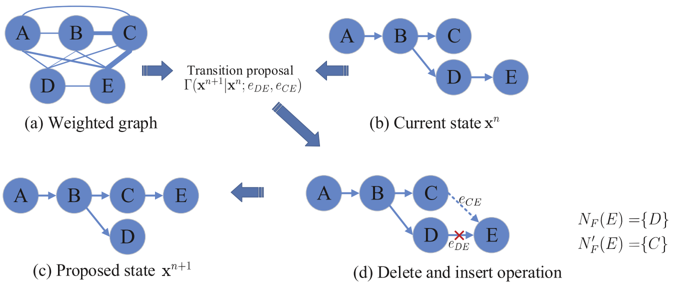
where e_i is the edge between node v_i and its original father node $N_F(v_i)$ and e'_i is that of v_i and its proposed new father node $N'_F(v_i)$. $P_{delete}(\mathbf{x}^n, e_i)$ denotes the dissimilar probability between v_i and its original father node $N_F(v_i)$

$$P_{delete}(\mathbf{x}^n, e_i) = \frac{\exp(-Sim(v_i, N_F(v_i)))}{\sum_{v_j \in \mathcal{E}} \exp(-Sim(v_i, v_j))} \tag{6}$$

And $P_{create}(\mathbf{x}^{n+1}, e'_i)$ is defined as the difference between similarity of v_i and the proposed new father node, and that of v_i and the original father node

$$P_{create}(\mathbf{x}^{n+1}, e'_i) = \frac{\exp(Sim(v_i, N_F(v_i)) - Sim(v_i, N'_F(v_i)))}{\sum_{v_j \in \mathcal{E}} \exp(Sim(v_i, N_F(v_i)) - Sim(v_i, v_j))} \tag{7}$$

If the similarity of v_i and $N_F(v_i)$ is larger than that of v_i and $N'_F(v_i)$, the probability to create the new edge is greatly increased. By deleting and adding edges based on Eq. (5), a new state of the tree



(a) Weighted graph (b) Current state x^n

(c) Proposed state x^{n+1} (d) Delete and insert operation

Fig. 3. Illustration of state transition. If we aim to move from the current state $\mathbf{x}^n$ (a) to the proposed state $\mathbf{x}^{n+1}$ (c), edge e_{DE} will be removed and edge e_{CE} will be inserted, as shown in (d). The transition proposal is derived from the weighted graph (a).

structure x^{n+1} is proposed. The procedure to propose a new tree sample x^{n+1} from x^n is illustrated in Fig. 3.

4. Mutiple instance metric learning

With the tree-structured graphical model, we get a set of fragments bags, each of which consist of similar instances from one person along the video sequence. Considering that the video sequence may be contaminated due to severe occlusion, we propose two schemes to address this problem. First, a *k*NN-based confidence score is introduced to evaluate the reliabilities of the instance, and applied in the Multiple Instance Multiple Label Logistic Discriminant Metric Learning (MiMI-DML) framework as a prior. Second, we use an impostor rejection mechanism in MiMI-DML framework to heuristically discover the impostors.

4.1. Impostor evaluation by K*-Nearest Neighbor*

k-Nearest Neighbor (*k*NN) algorithm has been proven to be effective in cluster removal and salience detection [34] area. We define a *k*NN-based confidence score to evaluate how much an impostor invades the interested target, and used it as a confidence prior in the multiple instance metric learning approach. We denote the neighbors of node v_i by $N(v_i) = \mathcal{V}/v_i$. In order to apply *k*NN for our impostor evaluation task, we utilize the distance of the *k*th nearest neighbor to define the anomaly score of a node belonging to an impostor within a bag

$$\mathcal{S}_{knn}(v_i) = D_k(v_i, N(v_i)) \tag{8}$$

where D_k denotes the distance of the *k*th nearest neighbor of v_i in its neighbor set $N(v_i)$. Then the confidence coefficient can be expressed by

$$w(v_i) = \exp(-\frac{||\mathcal{S}_{knn}(v_i)||_2^2}{2\sigma^2}) \tag{9}$$

where $\sigma^2 = \text{median}(\{||\mathcal{S}_{knn}(v_i)||\}_{v_i \in \mathcal{V}})$. The aim of the impostor evaluation is to measure the possibility of a vector being an impostor. In the re-identification problem, impostor is always caused by short-term occlusions. In this case, we assume that only a few nodes in the same video sequence could be similar to the impostor, so we set $k = N/3$ in the experiments.

4.2. MiMI-DML algorithm

In this section, we describe how Mahalanobis metrics can be learned from bag-level forms of supervision. As the instance-level real matches are unknown, we formulate a bag-level multiple instance metric learning strategy to jointly select the most discriminative pairs and learn an optimal Mahalanobis metric. A Mahalanobis distance $d_{\mathbf{M}}$ for instances $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^D$ is defined as

$$d(\mathbf{x}_1, \mathbf{x}_2; \mathbf{M}) = (\mathbf{x}_1 - \mathbf{x}_2)^\top \mathbf{M}(\mathbf{x}_1 - \mathbf{x}_2) \tag{10}$$

where **M** is a $D \times D$ symmetric positive semidefinite matrix. Let us denote a bag of examples as $\mathcal{X} = \{\mathbf{x}_i^1, \mathbf{x}_i^2, \dots \mathbf{x}_i^{N_i}\}$, where N_i is the number of examples in the bag indexed by *i*. To adjust to multiple instance setting, Eq. (10) is extended by selecting the one with minimum distance

$$d(\mathcal{X}_i, \mathcal{X}_j; \mathbf{M}) = \min_{\mathbf{x}_{i'} \in \mathcal{X}_i, \mathbf{x}_{j'} \in \mathcal{X}_j} d(\mathbf{x}_{i'}, \mathbf{x}_{j'}; \mathbf{M}) \tag{11}$$

Due to the influence of the impostors, the instance level labels are actually unknown. We have two choices of obtaining the discriminative metric: (a) Only bag level labels are used and only the most discriminative samples are applied to train the metric. (b) Instance level labels are estimated and most samples are utilized to train the metric. In the case of (a), as only the most discriminative pairs are used, the obtained metric is robust to noise in the data. However, if only one instance is used for each bag, quite a lot of samples are lost. In person re-identification problem, the number of positive samples is quite limited. If we use (a), many pairs of positive examples are dropped, which may result in the decrease of the generalization ability. So we use multiple instance multiple label logistic discriminant metric learning (MiMI-DML) algorithm [12] to jointly evaluate the instance level labels and learn proper metrics under different feature spaces. One of the key issues for applying MiMI-DML to the re-identification problem is the insufficient prior labels. In [12], the disturbance faces and names are removed manually from the dataset, so no 'negative' labels which indicating the impostors are allowed. In this paper, we integrate the impostor rejection process in the MiMI-DML framework.

Suppose we have n_p persons, we regard each person as a separate class. We also add an additional dimension of the class labels to indicate the 'impostor' class. Given a label vector $y_i \in \{0, 1\}^{n_p+1}$, where $y_i^{(n)} = 1 (n \in [1, n_p])$ indicates that the instance *i* belongs to class *n*. So $y_i^\top y_j = 1 (i, j \in [1, n_p])$ indicates that the two instances have the same identity. If not, $y_i^\top y_j = 0 (n \in [1, n_p])$. And $y_i^{(n_p+1)} = 1$ indicates that the instance is an impostor.

MiMI-DML optimizes the problem by maximizing the concave log-likelihood $\mathcal{L}$ of a logistic discriminant model

$$\underset{\mathbf{M}, \mathbf{Y}, b}{\text{maximize}} \ \mathcal{L} = \sum_{i,j} w_{ij} ((y_i^\top y_j) \log p_{ij} + (1 - y_i^\top y_j) \log(1 - p_{ij})) \tag{12}$$

where p_{ij} indicates the possibility of instances $\mathbf{x}_1$, $\mathbf{x}_2$ having the same identity

$$p_{i,j} = p(y_i = y_j | \mathbf{x}_1, \mathbf{x}_2, M, b) = \sigma(b - d(\mathbf{x}_1, \mathbf{x}_2; \mathbf{M})) \tag{13}$$

where $\sigma(z) = (1 + exp(-z))^{-1}$ is the sigmoid function, and the bias *b* acts as a threshold on the distance value to decide the identification of a new data pair. We define the prior weight $w_{ij} = w(v_i) w(v_j)$. The purpose of designing w_{ij} is: for sample pairs with $y_i^\top y_j = 1$, only real matches are expected and for sample pairs with $y_i^\top y_j = 0$, only real dis-matched pairs are expected. In both cases, the impostor can impact the quality of the learned metric and the weight for sample pairs with any one likely being an impostor should be decreased.

4.3. MiMI-DML-IR algorithm

In this section, we elaborate how MiMI-DML with impostor rejection (MiMI-DML-IR) algorithm works. The objective function Eq. (12) could be optimized by iteratively alternating (i) optimizing the metric **M** for the fixed instance labels, and (ii) estimating label matrix **Y** for fixed metric **M** and bias *b*. For fixed **Y**, it is a precise convex optimization problem and can be solved using projected gradient descent. Suppose $\mathbf{M} = \mathbf{L}^\top \mathbf{L}$, where **L** is a $d_r \times d_f (d_r \ll d_f)$

matrix, which ensures that **M** is a positive semidefinite matrix of rank d. The gradient of $\mathcal{L}$ with respect to **L** is

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\mathbf{L}} &= \mathbf{L}\sum_{i,j} w_{ij}(y_i^\top y_j - p_{ij})(\mathbf{x}_i-\mathbf{x}_j)(\mathbf{x}_i-\mathbf{x}_j)^\top \\ &= 2\mathbf{L}\sum_i \mathbf{x}_i((\sum_j w_{ij}y_i^\top y_j - w_{ij}p_{ij})\mathbf{x}_i^\top - \sum_j w_{ij}(y_i^\top y_j - p_{ij})\mathbf{x}_j^\top) \\ &= 2\mathbf{LXHX}^\top\end{aligned} \tag{14}$$

where $\mathbf{H}=[h_{ij}]\in\mathbb{R}^{n_h\times n_h}$ with $h_{ii}=\sum_{j\neq i} w_{ij}(y_i^\top y_j - p_{ij})$ and $h_{ij}=w_{ij}(p_{ij}-y_i^\top y_j)$ for $j\neq i$.

For fixed **L** and b, the objective function can be rewritten as

$$\begin{aligned}\mathcal{L} &= \sum_{i,j} w_{ij}(y_i^\top y_j)(\log p_{ij} - \log(1-p_{ij})) + c \\ &= \sum_{ij} w_{ij}\delta_{ij}(y_i^\top y_j) + c\end{aligned} \tag{15}$$

where $c=\sum_{ij} w_{ij}\log(1-p_{ij})$, and $\delta_{ij}=b-d_{\mathbf{M}}(\mathbf{x}_i,\mathbf{x}_j)$ are all constants. In Eq. (15), the only non-constant terms are those for data points in the same class. So we rewrite the objective for a particular instantiation of **Y**

$$\underset{\mathbf{Y}}{\text{maximize}}\sum_{n=1}^{C}\sum_{i\in\mathcal{Y}_n}\sum_{j\in\mathcal{Y}_n} w_{ij}\delta_{ij} \tag{16}$$

where $\mathcal{Y}_n$ is the set of indices of instances that are assigned to class n, i.e. $\mathcal{Y}_n=\{i\,|y_i^{(n)}=1\}$. Eq. (16) reveals that the optimization problem is an equivalent of a constrained weighted-data clustering problem. The constraint comes from two aspects: (a) The similarities come from instances derived from different views but with the same identity; (b) The instance in one bag can only have two possible labels, one is the class number it initially assigned, and the other is the label indicating the impostor. With these constraints, the clustering problem could be further simplified to an impostor rejection problem.

Then we aim to detect impostors based on the following assumption: the environments of the same person presented in different views are quite different. For example, if a is occluded by b in view A, there is little chance for a to be occluded by b in view B. So we assume that the impostor instance in one view is not able to find the correspondent match, and the distance between the impostor instance and its best match is large. Before moving into the details of the impostor rejection process, we first introduce the definition of jump cost, which is applied to automatically identify the impostors.

Definition 1. Given the order statistics $x_1, x_2, \ldots, x_n$, which are from the population distribution $F(x)$ with sample size n, the point estimation of expectation μ for the k-th sample is $\hat{u}_k=(\sum_{i=1}^{k} x_i+(n-k)x_k)/n$. We define the jump cost of the kth sample as

$$J_k=\frac{\hat{\mu}_{k+1}}{\hat{\mu}_k}(1\le k\le n-1) \tag{17}$$

Thus jump cost measures the ratio of the point estimation of expectation between two adjacent points. The distances between instance $\mathbf{x}^{A,\,i}$ and the candidate matches in $\mathcal{X}^B=\{\mathbf{x}^{B,1},\mathbf{x}^{B,2},\ldots,\mathbf{x}^{B,n^B}\}$ under the metric **M** can be regarded as the sample observations from one population distribution with sample size n^B. After sorting the distances from small to large, those values related to the impostors must be in the last places. If no occlusion happens, the value of the jump cost at each point should be low, and thus the variance of the jump cost over all the points should also be low. Otherwise, if occlusion happens and parts of the images belong to the occluder, there must be a big jump in value at the point where occlusion happens.

Our MiMl-DML algorithm with impostor rejection is summarized in Algorithm 2. Starting from initialing the instance labels

Algorithm 2 MiMl-DML-IR Algorithm.

```
Input: Training examples from two views $\{\{\mathbf{x}_c^{A,i}\}_{i=1}^{n_c^A}\}_{c=1}^{n_p}$ and $\{\{\mathbf{x}_c^{B,i}\}_{i=1}^{n_c^B}\}_{c=1}^{n_p}$.
1:  Initializing the instance level labels as $y^{(c)}(\mathbf{x}_c^{A,i})=1$ and $y^{(c)}(\mathbf{x}_c^{B,i})=1$ for each instance in class $c$.
2:  for $t=1:T$ (number of iterations) do
3:    With newly updated $\mathbf{Y}$, updating $\mathbf{M}$ with Eq. (14).
4:    for $c=1:n_p$ (number of classes) do
5:      for $i=1:n_c^A$(number of instances in class $c$ from view $A$) do
6:        $\bar{\mathbf{x}}_c^{A,i}=\arg\min_{\mathbf{x}^{B,j}} d(\mathbf{x}^{A,i},\mathbf{x}^{B,j};\mathbf{M})$
7:      end for
8:      for $i=1:n_c^A$ do
9:        Sorting $\{\bar{\mathbf{x}}_c^{A,i}\}_{i=1}^{n_c^A}$ from small to large.
10:       Calculating jump cost at each point $\{J_k\}_{k=1}^{n_c^A}$.
11:       Finding the largest $J_k$ within safe region $r=N_c^A/2$, denoted $J^*=\max_{k=[n_c^A/2]}^{N_c^A} J_k$
12:       if $y^{(n_p+1)}(\mathbf{x}_c^{A,i})=1$ & $\frac{|J^*-\mu(J)|}{\mu(J)}\le\delta$ where $\mu(J)=\frac{1}{n_c^A-1}\sum_{k=1}^{N_c^A-1}J_k$ then
13:         $y^{(c)}(\mathbf{x}_c^{A,i})=1$
14:       else if $y^{(c)}(\mathbf{x}_c^{A,i})=1$ & $\frac{|J^*-\mu(J)|}{\mu(J)}\ge\delta$ then
15:         $y^{(n_p+1)}(\mathbf{x}_c^{A,i})=1$
16:       end if
17:     end for
18:     Do the same step 5–17 for Camera $B$
19:   end for
20: end for
Output: Optimal metric $\mathbf{M}$ and label $\mathbf{Y}$
```

with the indexed class labels, after iteration T, the optimal metric **M** and label **Y** is obtained by iteratively alternating optimizing. In each iteration, we first update the metric **M** with the newly updated labels **Y**. Then we utilize an impostor rejection process (step 4–19) to automatically identify the impostors in each class. The impostor rejection process applied in each class consists of three key parts as follows:

Part.1 (step 5–7) For a pair of positive bags, $\mathcal{X}^A=\{\mathbf{x}^{A,1},\mathbf{x}^{A,2},\ldots\mathbf{x}^{A,n^A}\}$ and $\mathcal{X}^B=\{\mathbf{x}^{B,1},\mathbf{x}^{B,2},\ldots\mathbf{x}^{B,n^B}\}$, where subscripts indicating the identities are left out because they have the same identity, we apply a due-way selection process to obtain the best match under the current metric **M**. For any instance $\mathbf{x}^{A,\,i}$ from camera A, we choose the one who has the smallest distance to $\mathbf{x}^{A,\,i}$ as the best match

$$\bar{\mathbf{x}}^{A,i}=\arg\min_{\mathbf{x}^{B,j}} d(\mathbf{x}^{A,i},\mathbf{x}^{B,j};\mathbf{M}) \tag{18}$$

and vice versa for instance $\mathbf{x}^{B,\,j}$.

Part.2 (step 8–11) After sorting the best match set in an ascending sequence, we calculate the jump cost at each point, denoted by $\{J_k\}_{k=1}^{n_c^A}$. Then with an empirical assumption that in a bag at most quarter of the instances could be the impostors, we check the later part of the sorted jump cost sequence to obtain the largest jump cost. We call the first part of the sequence as a safe region. Sometimes rapid illumination changes may make the distance between the positive pair large. The impostor may exist for a short time, and could return to positive after the occlusion. However, the changes cause by the illumination changes may last long and unlikely return to the original state. With the safe region setting, we are able to avoid regarding positive samples as impostors.

Part.3 (step 12–16) In this part, we adjust the instance labels by checking the rate of the relative variation $|J^* - \mu(J)|/\mu(J)$. If the rate of the relative variation for an assumed impostor is less than the threshold δ, the instance is supposed to be an genuine instance and the value at the point of the class index in the label vector is set equal to 1. And if the rate of the relative variation for an assumed genuine instance is greater than the threshold δ, the instance is supposed to be an impostor and last value in the label vector is set equal to 1.

5. Experimental results

We conduct extensive experiments on two image sequence datasets designed for re-identification, the PRIDS 2011 dataset [11] and the newly published iLIDS-VID dataset [10].

iLIDS-VID datset - The iLIDS-VID dataset [10] was collected from two non-overlapping camera views in the i-LIDS Multiple Camera Tracking Scenario (MCTS), which was captured at an airport arrival hall under a multi-camera CCTV network. There are 300 randomly selected persons and 2 video sequences from separate cameras for each of them. Each video sequence has variable lengths consisting of 23 to 192 image frames, with an average number of 73. This dataset is quite challenging due to clothing similarities among people, lighting and viewpoint variations across camera views, cluttered background and severe occlusions.

PRID 2011 datset - The PRID 2011 re-identification dataset [11] includes 475 person trajectories from one camera and 753 from the other one, with 245 persons appearing in both views. Each image sequence has variable length consisting of 100 to 150 image frames, depending on the walking speed of an individual. Compared with the iLIDS-VID dataset, it was captured in uncrowded outdoor scenes with relatively simple and clean background and rare occlusions. For fair comparison, we use the same setting as in [11]. So for each of the 245 persons in the probe set, there are 753 candidate individuals in the gallery set.

5.1. Feature representation and matching mechanism

Feature representation - For the feature representation, we follow the feature extraction scheme in [35] and apply commonly used histogram-based features. We are intended to include many low-level features that have invariant properties in some illumination change situations. Three types of features are extracted: photometric color feature (RGB, HSV, Lab), invariance color feature (transformed color distribution-TCD, normalized RGB-NormRGB, opponent-OPP, color moment-CM) and color names feature [36] (CN). We partition an image into six horizontal stripes. For each strip, all histograms based color descriptors are extracted. For each color model, the color descriptors are extracted from both the whole image and foreground image (using a max-margin segmentation method [37]) and concatenated to form a feature vector. We refer the reader to [35] for further details about the feature representation.

Matching mechanism - With the test video fragments obtained from two cameras, we first apply feature extraction algorithm on each image to get 8 features based on 8 different color models. Then we conduct the tree-structured model to generate a set of video fragments and obtain the fragments bags, each of which contains the instances from one branch of trees. Then with the Mahalanobis metrics trained in the MildML algorithm, we get the distance between two instances:

$$d(\mathbf{x}_1, \mathbf{x}_2; \{\mathbf{M}_i\}_{i=1}^{8}) = \sum_{i=1}^{8} (\mathbf{x}_1^i - \mathbf{x}_2^i)^\top \mathbf{M}_i (\mathbf{x}_1 - \mathbf{x}_2) \tag{19}$$

After getting the Mahalanobis distances between any instance pair, we regard the minimum distance among all the distances between instance pairs separately from two bags derived from different trees as the distance between the two bags. And also we regard the minimum distance among all the distances from the bag pairs separately derived from two trees as the distance between the two trees. After ranking the distances between the probe image and all the gallery images in an ascending order, we get the matching results for the probe image.

5.2. Settings

Parameter setting - After feature extraction, the dimensionality of the concatenated feature vector for each color model is reduced to d_f via principal component analysis (PCA). In all the experiments, we set the dimension d_f to 70. In the Wang–Laudau adapted Simulated Tempering algorithm, we set the parameter $\alpha_0 = 0.5$ to provide an equal chance to keep the current state and transform to a new state. The parameter γ denoting the weight updating rate is empirically set to 0.2. In the energy function $E(x)$, the trade-off parameter λ is set to be 0.1 after cross-validation on a small experimental data subset.

Evaluation settings - Following [10], the whole set of human sequence pairs of both datasets is randomly split into two subsets with equal size, one for training and the other for testing. The sequences from the first camera are used as the probe set while the ones from the other camera as the gallery set. For both datasets, the performances are shown in the average cumulative matching characteristics (CMC) curves over 10 trials.

5.3. Comparison to the state-of-the-art

In this section, our method is compared with three state-of-the-art muti-shot person re-identification approaches. Table 1 shows the results of the proposed algorithm with comparisons to three algorithms: DVR [10], DVDL [21] and STV3D+KISSME [38]. Overall, , the proposed method achieves 45.3% at rank-1, 73.2% at rank-5, 84.1% at rank-10 and 91.5% at rank-20 on the iLIDS-VID dataset, and 65.3% at rank-1, 81.1% at rank-5, 89.4% at rank-10 and 92.6% at rank-20 on the PRID 2011 dataset. The proposed algorithm performs comparably to the STFV3D + KISSME algorithm, but achieves significant improvements over DVR and DVDL. As shown in Table 1, our method outperforms the DVR method for more than 20% on rank-1 recognition rate. This is because that instead of selecting only one most discriminative fragment, we evaluate the instance level labels and use all the positive samples to train the Mahalanobis metric, which makes our method more discriminative.

5.4. Ablation study

We conduct ablation studies to understand how critical are the tree-structure based video fragment bags generation model(TSM) and MiMl-DML-IR algorithm for the performance of the re-identification task. We use the data obtained from the iLIDS-VID dataset to show the experimental results. All the results for the ablation study are reported in Table 2. We use two baseline approaches to demonstrate the effectiveness of the tree-structure based video fragments bag scheme. (i) Each video sequence is equally divided into several units, and each unit contains n_u images. For fair comparison, n_u is selected via cross-validation. (ii) Each video sequence is automatically clustered through the affinity propagation (AP) [39] approach. Compared with using the tree-structured model (TSM) (45.3% on rank-1 recognition), the rank-1 recognition rate drops to 21.9% when equal scheme applied and 34.5% when AP clustering applied.

Table 1
Performance of our method compared with state-of-the-art methods on iLIDS-VID and PRID 2011 dataset.

Dataset	iLIDS-VID				PRID 2011			
Rank R	R = 1	R = 5	R = 10	R = 20	R = 1	R = 5	R = 10	R = 20
DVR [10]	23.3	42.2	55.3	68.4	28.9	55.3	65.5	82.8
DVDL [21]	25.9	48.2	57.3	68.9	40.6	69.7	77.8	85.6
STFV3D + KISSME [38]	44.3	71.7	83.7	**91.7**	64.1	**87.3**	**89.9**	92.0
OURS	**45.3**	**73.2**	**84.1**	91.5	**65.3**	81.1	89.4	**92.6**

Table 2
Effectiveness evaluation of two main components in this paper: TSM and MiMl-DML-IR. TSM and MiMl-DML-IR are alternately disabled and substituted by baseline methods.

Dataset	iLIDS-VID			
Rank R	R = 1	R = 5	R = 10	R = 20
Equal + MiMl-DML-IR	21.9	32.5	54.1	67.2
AP [39] + MiMl-DML-IR	34.5	49.8	67.6	78.7
TSM + KISSME [40]	30.2	42.1	61.7	70.9
TSM + MildML [12]	38.1	54.6	72.4	82.4
TSM + MiMl-DML	40.4	61.2	76.8	84.1
OURS (TSM + MiMl-DML-IR)	**45.3**	**73.2**	**84.1**	**91.5**

Next, we conduct experiments on the MiMl-DML-IR subset. Two representative metric learning approaches, KISS Metric (KISSME) learning [40] and Multiple Instance Logistic Discriminant Metric Learning (MildML) [12], are applied to evaluate the effectiveness of our MiMl-DML-IR algorithm. KISSME algorithm is proved to be quite effective in one shot person re-identification [8,35] task. In the experiments, we randomly select two instances respectively from two relevant bags (generated by the same person in different views) as positive sample pair, and those from irrelevant bags (generated by different persons in different views) as negative sample pair. To reduce the impact of potential impostors and add robustness to the metric learning process, 5 instances are extracted from each bag and both positive pairs and negative pairs are generated by the permutation and combination mechanism. In the MildML algorithm, only bag level labels are estimated and only the samples with the minimum distance are applied to train the metric. As impostor rejection is a key component of the proposed approach, we also conduct an experiment to evaluate the effectiveness of the impostor rejection module. We disable the impostor rejection module. No extra bit is reserved for the label of the impostors and any instance in the bag could be labeled as positive. The ablation study in terms of impostor rejection is named 'TSM + MiMl-DML' in Table 2. We can see that, compared with using MiMl-DML-IR, the rank-1 recognition rate drops to 30.2% with KISSME algorithm, 38.1% with MildML algorithm, and 40.4% with impostor rejection module disabled. The experimental results indicate that the proposed tree-structured model and MiMl-DML-IR algorithm play their due roles in person re-identification, while TSM plays a more important role in the overall contribution.

Equal scheme classifies the images by force, and images in each unit may present diverse states. AP clustering classifies the images with appearance similarities without any temporal information employed. So using either equal scheme or AP clustering algorithm, it is not able to select the most discriminative state for each positive pair. The proposed TSM encourages the images which are adjacent on temporal order to be in the same bag. Therefore, it can not only classify the images with similar states and adjacent temporal orders in one bag, but also classify the impostor images in separate bag, laying the foundations for automatically selecting the most discriminative bags, and rejecting the impostor images for MiMl-DML-IR algorithm.

KISSME method indiscriminately regards the image pairs from two relevant bags as positive sample, taking no account of whether they are the best matches. In MildML algorithm, only bag-level labels are estimated. In person re-identification problem, the number of positive samples is quite limited. So only with bag-level labels, many pairs of positive samples are dropped, resulting in the decrease of the generalization ability. Without impostor rejection module, the MiMl-DML algorithm may mistakenly take the impostors as the most discriminative instance, leading to the decrease of the performance. The proposed MiMl-DML-IR algorithm can automatically discover impostor images, and obtain transition metrics by selecting most discriminative sample pairs.

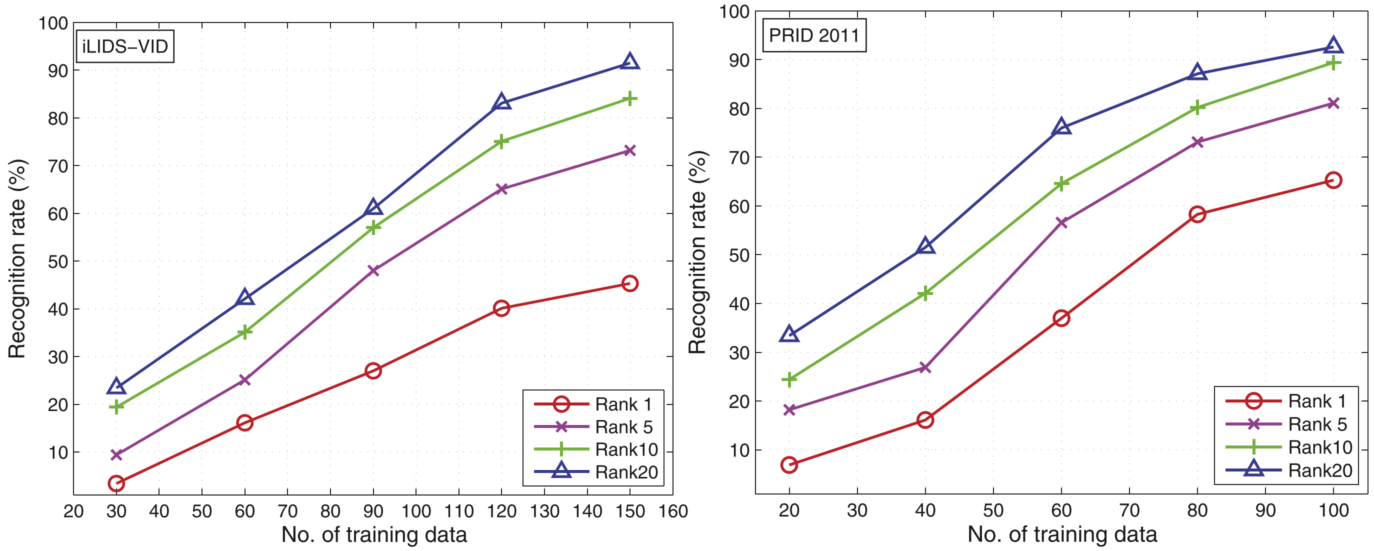
5.5. Evaluation on sensitivity and color descriptors

In this section, we first present the sensitivity tests of several critical parameters and then we conduct experiments using the proposed method on individual color features to evaluate the performance of each color descriptor.

Sensitivity to the number of training samples - In person re-identification, approaches with less training examples are preferred, as in camera network, collecting label data from each camera pair is time consuming. We conduct experiments to discover the relationship between the matching rate and the number of the training data. Fig. 4 shows the recognition rate of rank 1 to rank 20 using various amount of training data on both datasets. On the iLIDS-VID dataset, as the amount of training data rising from 30 to 120, the increase of the recognition rate is notable and then begins to level off since 120. The rank 1 matching rate achieves 40.1% when training with 120 sample pairs, only 5.2% less than training with 150 sample pairs. And on the PRID 2011 dataset, the turning point appears around 80. The rank 1 matching rate achieves 58.3% when training with 80 sample pairs, about 7.0% less than training with 100 sample pairs. This indicates that to some extent, the proposed algorithm relaxes the restriction to the number of labeled data.

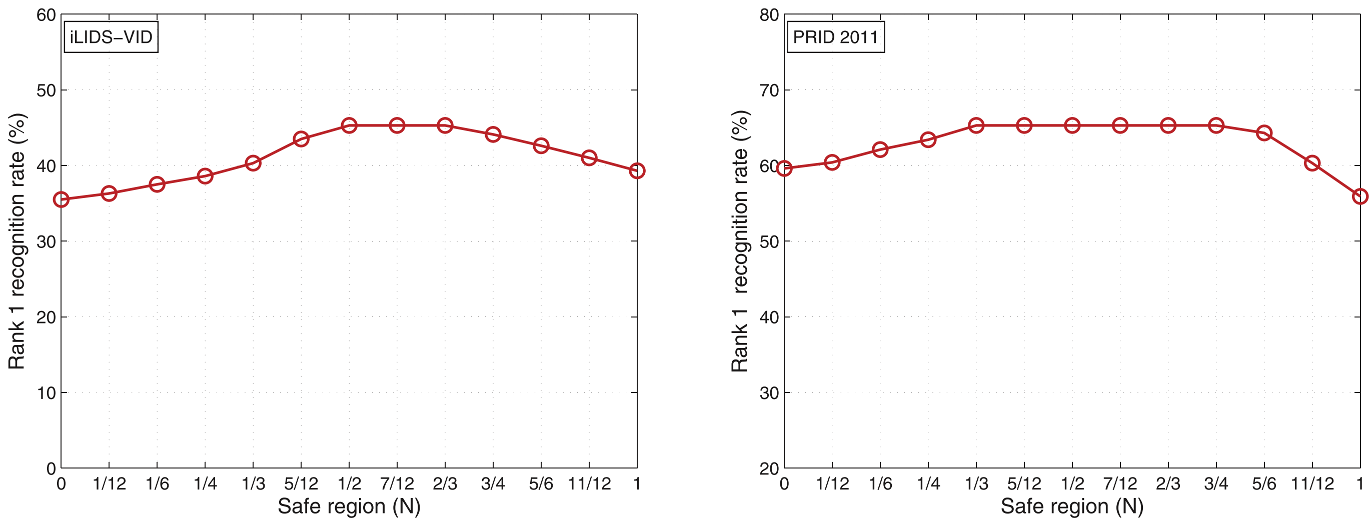
Sensitivity to safe region Safe region r is introduced in the impostor rejection process to differentiate impostor data from positive examples with obvious appearance changes caused by illumination variation. We conduct experiments on both datasets to determine the value of r and evaluate how much our method can benefit from the safe region setting. The parameter r vary from 0 to N with step $N/12$, where N is the number of the instances in each bag. Zero indicates that no safe region is taken into consideration, and in this case positive examples with significant appearance changes tend to be regarded as impostors. N means an ultimate state, in which the safe region setting is disabled. As shown in Fig. 5, the rank 1 recognition rate gradually declines when r approaches '0' and 'N'. For the iLIDS-VID dataset, when varying r from $N/2$ to $2N/3$, the best rank 1 recognition rate could be achieved. And for the PRID 2011 dataset, the model performs well with r varying from $N/3$ to $3N/4$. There is a sharp drop when r varies from $5N/6$ to N, because with a large r setting, some impostors are regarded as positive examples by mistake.

Evaluation of individual color model To evaluate the performance of different color descriptors on the proposed algorithm,



(a) (b)

Fig. 4. Recognition rate of rank 1 to rank 20 using various amount of training data on the iLIDS dataset (a) and the PRID 2011 dataset (b).



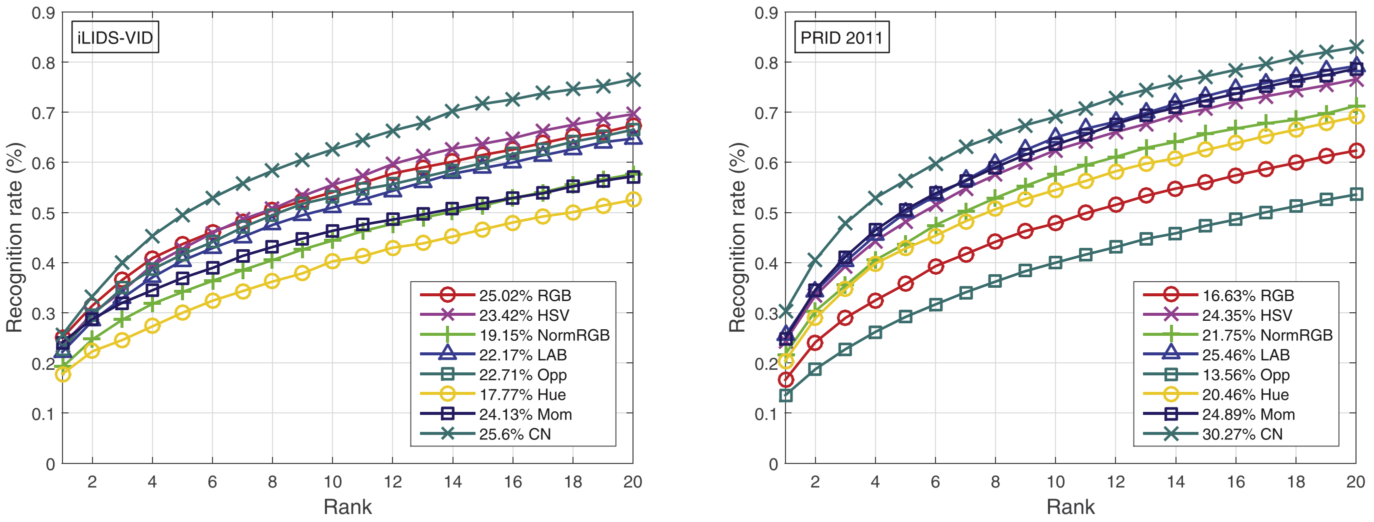
(a) (b)

Fig. 5. Rank 1 recognition rate using various safe region settings on the iLIDS dataset (a) and the PRID 2011 dataset (b).

we carry out experiments using individual color features. The results are illustrated in the CMC curves in Fig. 6 for both iLIDS-VID and PRID 2011 datasets. Because the re-identification performance of different color rankers are scene specific, all the features except the CN feature perform differently in the two scenes. The CN-based model performs the best in both datasets: 25.6% in the iLIDS-VID dataset and 30.27% in the PRID 2011 dataset. For the Mom feature, both O1 and O2 are invariant to light intensity shift (O3 represents intensity information) and thus they are effective to account for appearance changes in the PRID 2011 dataset. Overall, the algorithm based on HSV, LAB and CN features performs constantly well on both datasets with different lighting conditions.

5.6. Computational complexity

In this section, we compare the computational complexity of the two main components in the proposed algorithm with that in DVR [10]. In the first part of generating candidate video fragments pools, DVR requires computing the optic flow of each frame. The time complexity of the computation of the optic flow is $\mathcal{O}(Nn_{pix}^3)$, where N is the number of the frames and n_{pix} is the number of the pixels in the lower body. In the proposed algorithm, the tree structure is learned by simulated tempering algorithm, and the time complexity for this algorithm is $\mathcal{O}(N_t n_{frag})$, where N_t is the number of the iterations and n_{frag} is the number of the fragments. In ILIDS-VID dataset, $n_{pix} \approx 4000$, and $n_{frag} \approx 10$–20, so the proposed algorithm is much better than DVR. In the second part of selecting and ranking the fragment pairs, DVR formulates the ranking step in a non-constrained primal problem, and solves it by a linear conjugate gradient method. The time complexity of conjugate gradient is $\mathcal{O}(d_{H3}\sqrt{\kappa})$, where d_{H3} is the dimension of the HOG3D feature, and κ is the condition number. The proposed MiMI-DML algorithm has the time complexity $\mathcal{O}(n_h(n_h + d_f)d_r)$, where n_h is the number of the image pairs from related bags, d_f is the dimension of the histogram feature used in this paper and d_r is the rank of the matrix, which are all introduced in Section 4.3. In the practical application, $d_{H3} = 9600$, κ is set to approach 100 in order to get a desired convergence [41], $d_f = 70$, $n_h \approx 50$, and $d_r \ll d_f$. Therefore, the rank-



(a) (b)

Fig. 6. The CMC performance comparison of using different color rankers on the iLIDS-VID (a) and PRID 2011 datasets (b). Rank-1 recognition rate is marked in front of the ranker name.

ing algorithms from the two compared methods are roughly at the same level in computational complexity.

All the experiments are performed on an Intel 3.2GHZ PC with 8G memory, and the algorithms are implemented in Matlab. For the iLIDS-VID dataset, the time used to train the metrics is 2.1 min, and the time using in testing for each person is 10± 1.5 s. For the PRID 2011 dataset, the time used to train the metrics is 2.8 min, and the time used in testing for each person is 13± 1.9 s.

6. Conclusion and future work

In this paper, we present a novel multiple instance metric learning framework to solve the model drift problem caused by impostor images in video-based person re-identification. Impostors and discriminative metrics are iteratively discovered and learned to obtain an optimal solution for the logistic discriminant energy function. Experimental results on two benchmark datasets demonstrate the proposed algorithm performs favorably against the state-of-the-art methods. Future research includes jointly learning multiple distance metrics with multiple complementary features, so that more discriminative information can be exploited than those learned from individual features.

Acknowledgment

X. Liu, H. Wang, J. Wang, and X. Ma are supported in part by National Natural Science Foundation of China under grant 61671102, 61671103 and 61401059.

References

[1] D. Gray, H. Tao, Viewpoint invariant pedestrian recognition with an ensemble of localized features, in: European Conference on Computer Vision, Springer Berlin Heidelberg, Marseille, 2008, pp. 262–275.

[2] M. Farenzena, L. Bazzani, A. Perina, V. Murino, M. Cristani, Person re-identification by symmetry-driven accumulation of local features, in: IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, 2010, pp. 2360–2367.

[3] R. Zhao, W. Ouyang, X. Wang, Person re-identification by salience matching, in: Proceedings of the IEEE International Conference on Computer Vision, Sydney, 2013, pp. 2528–2535.

[4] B. Ma, Y. Su, F. Jurie, Local descriptors encoded by fisher vectors for person re-identification, in: European Conference on Computer Vision, Springer Berlin Heidelberg, Florence, 2012, pp. 413–422.

[5] I. Kviatkovsky, A. Adam, E. Rivlin, Color invariants for person reidentification, IEEE Trans. Pattern Anal. Mach. Intell. 35 (7) (2013) 1622–1634.

[6] S.G. W. S. Zheng, T. Xiang, Re-identification by relative distance comparison, IEEE Trans. Pattern Anal. Mach. Intell. 35 (3) (2012) 653–668.

[7] C. Loy, C. Liu, S. Gong, Person re-identification by manifold ranking, in: IEEE International Conference on Image Processing, Bangalore, 2013, pp. 3567–3571.

[8] P.M. Roth, M. Hirzer, M. Köstinger, C. Beleznai, H. Bischof, Mahalanobis distance learning for person re-identification, In Person Re-Identification(2014) 247–267.

[9] N. Gheissari, T.B. Sebastian, P.H. Tu, J. Rittscher, R. Hartley, Person reidentification using spatiotemporal appearance, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, New York, 2006, pp. 1528–1535.

[10] T. Wang, S. Gong, X. Zhu, S. Wang, Person re-identification by video ranking, in: Person Re-identification by Video Ranking, Springer International Publishing, Zurich, 2014, pp. 688–703.

[11] M. Hirzer, P. Roth, M. Köstinger, H. Bischof, Relaxed pairwise learned metric for person re-identification, in: European Conference on Computer Vision, Springer Berlin Heidelberg, Firenze, 2012, pp. 780–793.

[12] M. Guillaumin, C. Schmid, Multiple instance metric learning from automatically labeled bags of faces, in: European Conference on Computer Vision, Springer Berlin Heidelberg, heraklion, 2010, pp. 634–647.

[13] L. Bazzani, M. Cristani, A. Perina, M. Farenzena, V. Murino, Multiple-shot person re-identification by HPE signature, in: IEEE International Conference on Pattern Recognition, Istanbul, 2010, pp. 1413–1416.

[14] S. Bąk, E. Corvee, F. Bremond, M. Thonnat, Multiple-shot human re-identification by mean Riemannian covariance grid, in: IEEE International Conference on Advanced Video and Signal Based Surveillance, Klagenfurt, 2011, pp. 179–184.

[15] A. Bedagkar-Gala, S.K. Shah, Part-based spatio-temporal model for multi-person re-identification, Pattern Recognit. Lett. 33 (14) (2012) 1908–1915.

[16] V. Takala, Y. Cai, M. Pietikäinen, Boosting clusters of samples for sequence matching in camera networks, in: International Conference on Pattern Recognition, Istanbul, 2010, pp. 400–403.

[17] W. Li, X. Wang, Locally aligned feature transforms across views, Portland, 2013, pp. 3594–3601.

[18] H. Liu, B. Ma, L. Qin, J. Pang, C. Zhang, Q. Huang, Set-label modeling and deep metric learning on person re-identification, Neurocomputing 151 (2015) 1283–1292.

[19] R.C. M. S. Nixon T. Tan, Human identification based on gait, Springer Science & Business Media, 2010.

[20] Z.L.I.R.V.P.G. S. Sarkar P. J. Phillips, K.W. Bowyer, The humanID gait challenge problem: Data sets, performance, and analysis, IEEE Trans. Pattern Anal. Mach. Intell. 27 (2) (2005) 162–177.

[21] S. Karanam, Y. Li, R.J. Radke, Person re-identification with discriminatively trained viewpoint invariant dictionaries, in: Proceedings of the IEEE International Conference on Computer Vision, Santiago, 2015, pp. 4516–4524.

[22] H.C.R. Pribadi, H.-K. Pao, Sparse tree structured representation for re-identification, Pattern Recognit. 60 (2016) 394–404.

[23] M. Hirzer, P.M. Roth, H. Bischof, Person re-identification by efficient impostor-based metric learning, in: IEEE Advanced Video and Signal-based Surveillance, Beijing, 2012, pp. 203–208.

[24] F. Xiong, M. Gou, O. Camps, M. Sznaier, Person re-identification using kernel-based metric learning methods, in: European Conference on Computer Vision, Springer International Publishing, 2014, pp. 1–16.

[25] W. Li, Y. Wu, J. Li, Re-identification by neighborhood structure metric learning, Pattern Recognit. 61 (2016) 327–338.

[26] F. Wang, Semisupervised metric learning by maximizing constraint margin, IEEE Trans. Syst., Man, Cybern., Part B 41 (4) (2011) 931–939.

[27] M. Bilenko, S. Basu, R.J. Mooney, Integrating constraints and metric learning in semi-supervised clustering, in: International Conference on Machine Learning, Banff, 2004, pp. 81–88.

[28] Y. Xu, W. Ping, A.T. Campbell, Multi-instance metric learning, in: Proceedings of IEEE International Conference on Data Mining, ICDM, 2011, pp. 874–883.

[29] C. Shi, C. Wang, B. Xiao, S. Gao, J. Hu, End-to-end scene text recognition using tree-structured models, Pattern Recognit. 47 (9) (2014) 2853–2866.

[30] R. Adams, Z. Ghahramani, M.I. Jordan, Tree-Structured stick breaking for hierarchical data, Advances in neural information processing systems, 2010.

[31] E. Bart, I. Porteous, P. Perona, M. Welling, Unsupervised learning of visual taxonomies, in: IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, 2008.

[32] J. Kwon, K.M. Lee, A unified framework for event summarization and rare event detection, in: IEEE Conference on Computer Vision and Pattern Recognition, Providence, 2012.

[33] F. Wang, D.P. Landau, Efficient, multiple-range random walk algorithm to calculate the density of states, Phys. Rev. Lett. 86 (10) (2001) 2050–2053.

[34] W.O. R. Zhao, X. Wang, Unsupervised salience learning for person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition, Portland, 2013.

[35] X. Liu, H. Wang, Y. Wu, J. Yang, M.-H. Yang, An ensemble color model for human re-identification, in: Proceedings of IEEE Winter Conference on Applications of Computer Vision, Hawaii, 2015, pp. 868–875.

[36] J. Van De Weijer, C. Schmid, J. Verbeek, Learning color names from real-world images, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, Minneapolis, 2007, pp. 1–8.

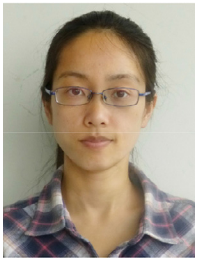
[37] J. Yang, S. Sáfár, M.-H. Yang, Max-margin Boltzmann machines for object segmentation, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Columbus, 2014, pp. 320–327.

[38] K. Liu, B. Ma, W. Zhang, R. Huang, A spatio-temporal appearance representation for video-based pedestrian re-identification, in: Proceedings of the IEEE International Conference on Computer Vision, Paris, 2016, pp. 3810–3818.

[39] B.J. Frey, D. Dueck, Clustering by passing messages between data points, Science 315 (5814) (2007) 972–976.

[40] M. Kostinger, M. Hirzer, Large scale metric learning from equivalence constraints, in: IEEE Conference on Computer Vision and Pattern Recognition, Springer, Rhode Island, 2012, pp. 2288–2295.

[41] J.R. Shewchuk, An introduction to the conjugate gradient method without the agonizing pain, Science 49 (CS-94-125) (1994) 64.



Xiaokai Liu received the B.E. degree in 2010 from School of Information Engineering, Dalian Maritime University (DMU), PR China. Now she is a doctoral candidate in the School of Information and Communication Engineering of Dalian University of Technology (DUT), PR China. Her research interests include person reidentification and machine learning.

Hongyu Wang received the B.S. degree from Jilin University of Technology in 1990, and M.S. degree from Graduate School of Chinese Academy of Sciences in 1993, both in Electronic Engineering. He received the Ph.D. in Precision Instrument and Optoelectronics Engineering from Tianjin University in 1997.

He is currently a Professor at Dalian University of Technology. His research interests include algorithmic, optimization, and performance issues in wireless ad hoc, mesh and sensor networks.

Jie Wang (M'12) received his B.S. degree from Dalian University of Technology, Dalian, China, in 2003, M.S. degree from Beijing University of Aeronautics and Astronautics, Beijing, China, in 2006, and Ph.D degree from Dalian University of Technology, Dalian, China, in 2011, all in Electronic Engineering.

He is currently an Associate Professor at Dalian University of Technology. His research interests include wireless localization and tracking, wireless sensor networks, and cognitive radio networks.

Xiaorui Ma received the B.E. degree in 2008 from School of Mathematics and Statistics, Lanzhou University (LZU), P.R. China. Now she is a doctoral candidate in the School of Information and Communication Engineering of Dalian University of Technology (DUT), P.R. China. Her research interests include remote sensing image classification and machine learning.