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.

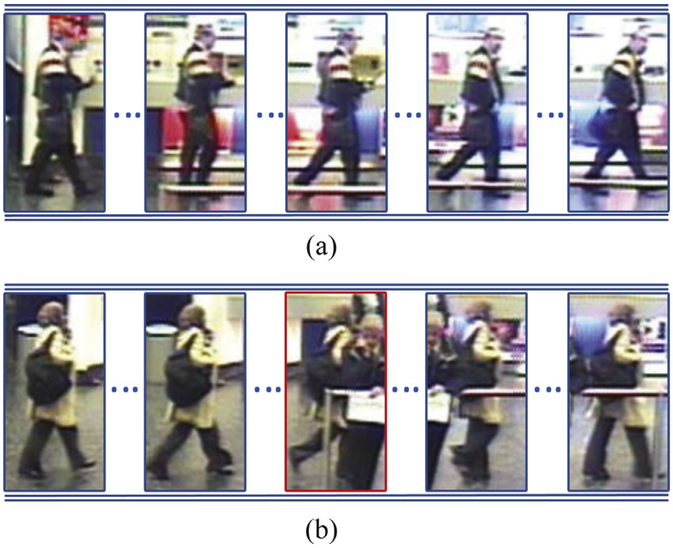

**Fig. 1.** Sample images from two example videos in the iLIDS-VID dataset. Video (a) undergoes rapid illumination changes during the video capture period. In video (b), the woman in khaki is heavily occluded by the man in black, marked in a red box. (For interpretation of the references to color in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