Multiple-Shot Person Re-Identification by Pairwise Multiple Instance Learning

Chunxiao LIU†a, Nonmember, Guijin WANG†b, Member, and Xinggang LIN†c, Nonmember

1. Introduction

Person re-identification refers to matching images of the same person captured in different locations or by different cameras. It has wide applications in surveillance across a camera network, such as person retrieval and global understanding of human behavior. Combining visual information from multiple images of a person would help the recognition. However, as shown in Fig. 1(a) human appearance can be inconsistent during tracking due to variations in illumination and viewpoint. Moreover, corrupted images due to occlusion may be contained in the collected sequences. Another problem with person re-identification is the visual ambiguity owing to the fact that people can wear similar clothes (see Fig. 1(b)(c)). In some cases appearances of the same person between two cameras vary more significantly than those of two different persons, due to different camera settings and environments. Thus, it is challenging to take advantage of information from multiple images to do the recognition.

Existing work on multiple-shot re-identification can be categorized into two types: non-learning based and learning based methods. In the non-learning based methods, the objective is to generate a set of distinctive signatures (each for a single image) [1], [2] or a single signature (with features accumulated from multiple images) for each person [3], [4]. The former scheme is sensitive to corrupted images. The latter overcomes that problem; however, accumulating features from multiple images degrades the effectiveness of feature due to appearance inconstancy. In the learning based methods, the objective is to train an appearance model for each person. Discriminative learning methods, partial least squares (PLS) [5] and boosting [6], have been explored to learn an appearance model by combining weighted features. However, these methods assume that the images of a person are correctly labeled without occlusion or false detections.

Recently the ranking/metric learning based methods [7]–[11] have gained much improvement in handling visual ambiguity in person re-identification by exploring a set of pairwise comparison, which inspire our work in this paper. Instead of always aiming at high matching rate at rank 1, they favor to provide a ranking list in which the correct match is more likely to be identified at a high rank. However, these supervised metric learning methods require labeled image pairs from two cameras for training, which is troublesome to collect in practice. In contrast, we focus on the robust appearance learning for each individual but not the metric learning, and thus do not require any pairwise labelled training set from two cameras.

In this paper, we propose a novel pairwise comparison based multiple instance learning (PMIL) method for multi-shot person re-identification, with the aim of learning a representative and robust model for each person. Our contributions are as follows. First, we employ the multiple instance concept to represent an image sequence. It does not assume that all the images are correctly tracked, which well suits the real problem. Second, based on the above representation, we propose a novel pairwise comparison based optimization for multiple instance learning. Our approach differs from most existing learning methods in that, rather than seeking a large separation margin between two classes, it favors finding a robust model to order the classes. Experiments on two public datasets show that our approach is more capable of handling corrupted images and visual ambiguity.

SUMMARY Learning an appearance model for person re-identification from multiple images is challenging due to the corrupted images caused by occlusion or false detection. Furthermore, different persons may wear similar clothes, making appearance feature less discriminative. In this paper, we first introduce the concept of multiple instance to handle corrupted images. Then a novel pairwise comparison based multiple instance learning framework is proposed to deal with visual ambiguity, by selecting robust features through pairwise comparison. We demonstrate the effectiveness of our method on two public datasets.

key words: person re-identification, pairwise, multiple-shot, multiple instance learning
in the person re-identification application.

2. Pairwise Multiple Instance Learning

2.1 Multiple Instance Learning

Multiple instance learning (MIL) is a machine learning method representing training sets by labeled bags that are composed of unlabeled instances. In binary classification, a bag is labeled positive if at least one instance in that bag is positive, while a bag is labeled negative if all the instances in it are negative. As to appearance learning, generally an image sequence of a person extracted by a tracking method contains several corrupted images caused by occlusion or false detection. Thus, it is appropriate to represent the training samples in a bag form and pass the sample ambiguity to the learning model. The goal of MIL in this paper is to find a typical model that can well represent the ‘true’ positive instances in the positive bags, and then use it for recognition.

In [12] a general framework for solving MIL problem is proposed by maximizing Diverse Density (DD) function:  

\[ D(x_i) = \prod_i p_i(1 - p_i)^{1 - \gamma}, \]

where \( p_i \) is the bag’s label. First, we formulate the probability comparison model by:  

\[ P(p(x_i) > p(x_j)) = (1 + \exp(-\beta[p(x_i) - p(x_j)]))^{-1}. \]  

(1)

We assume that the pairwise comparisons between all the positive and negative bags are independent since each person can have his/her own wearing style and dress differently. Then our learning objective is to optimize the relative rank likelihood:

\[ \max \prod_i \prod_j P(p(x_i) > p(x_j)), \]  

(2)

where \( \forall i, j : y_i = 1, y_j = -1 \). To facilitate the following derivations, log-likelihood \( L \) of the criteria 2 is employed:

\[ L = - \sum_i \sum_j \log(1 + \exp(-\beta[p(x_i^+) - p(x_j^-)])). \]  

(3)

For simplicity we use superscript + and − to indicate positive and negative bag. This criteria shares the similar spirit with [7], [8], [10], [11], in terms of treating person re-identification as a ranking problem to tackle visual ambiguity. However, we integrate pairwise comparison into appearance learning while they focus on metric learning between two cameras.

2.3 PMIL Boosting

With this new criteria 3, we solve the MIL problem in the framework of boosting. Recall that in boosting each instance is classified by a linear combination of \( T \) weak classifiers, denoted by \( c(\cdot) = \sum_{t=1}^T h_t(\cdot) \). Consider \( h(\cdot) \) has real value output. To this end, the appearance learning can be viewed as training a classifier \( c(\cdot) \), but using bag samples. Let \( x_{ik} \) denotes the \( k \)-th instance in the \( i \)-th bag. We define the probability \( p(x_{ik}) \) by the standard logistic function with respect to \( c(x_{ik}) \):

\[ p(x_{ik}) = (1 + \exp(-c(x_{ik})))^{-1}. \]  

(4)

To express the contribution of instances to a bag of being positive, Noisy-OR (NOR) model [12] is used as follows:

\[ p(x_i) = 1 - \prod_{k=1}^w (1 - p(x_{ik})), \]  

(5)

which can produce high response to a positive bag even if there exists only one positive instance.

In each iteration we select the weak classifier that exhibits best performance on the current distribution of sample weight, which is defined by:

\[ \hat{h} = \arg \max_h \sum_{i,k} w_{ik} h(x_{ik}), \]  

(6)

where \( w_{ij} \) is the instance weight. Following the AnyBoost approach, reweighting an instance is derived as the derivative of the optimization function \( L \) with respect to a change in the classifier response of the instance. For instances in positive bags, weight updating is computed by:

\[ w_{ik}^+ = \frac{\partial L}{\partial c(x_{ik}^+)} = \frac{\partial L}{\partial p(x_{ik}^+)} \cdot \frac{\partial p(x_{ik}^+)}{\partial c(x_{ik}^+)} \]  

\[ = (1 - p(x_{ik}^+))p(x_{ik}^+) \sum_j W_{i-j}, \]  

(7)

where \( \forall j, y_j = -1 \) and \( W_{i-j} = \beta(1 + \exp(-\beta[p(x_i^+) - p(x_{ij}^-)]))^{-1} \). From Eq. 7 we can see that weight \( w_{ik}^+ \) consists of three elements: \( 1 - p(x_{ik}^+) \), \( p(x_{ik}^+) \), and \( \sum_j W_{i-j} \). The interpretation is straightforward and meaningful: the first term denotes the contribution from its associated bag; the second term denotes the instance’s probability; the third term denotes the contribution from all the bags with opposite labels. By \( i \leftarrow j \) we illustrate that the feedback is given from bag \( j \) to \( i \). The first and third terms are bag specific which means that instances in the same bag share the same contribution from the bag. The second term is instance specific indicating the importance of an instances within a bag. Thus as learning proceeds, if the probability of a bag increases then
the weights of instances inside the bag should be reduced; if the contribution from its opposite labeled bag increases, which implies this bag is relatively easy to rank according to pairwise comparison, then the weights of instances inside this bag should be increased; while within the bag, instance with high scores will dominate the subsequent iteration as shown by the second term.

Weights $w^-$ for the instances in the negative bags can be derived in the similar manner:

$$w^--(1-p(x_j^i)p(x_j^k)) \sum_i W_{i\rightarrow j},$$

where $v_i, y_i = 1$ and $W_{i\rightarrow j} = W_{j\rightarrow i}$. It means that the interaction between positive bag and negative bag is symmetric. Note that the weights here are signed agreeing with the bag label.

Compared to [13] which optimizes DD function, our reweighting equation has an additional quantity $W$ due to the constraints from relative comparisons between positive bags and negative bags. It steers the feature selection process to pick up features that has better performance in ranking the positive bags, rather than giving positive bags high scores.

3. Appearance Learning Based on PMIL

Let $S$ denotes the target person for appearance learning. The output of appearance learning is a strong classifier $c(\cdot)$ which can evaluate the similarity between a probe image and the person $S$. We overview the learning procedure as follows: we first prepare bag form samples, followed by generating a feature pool. Positive bags are constructed by multiple images of $S$. Images of the other persons in the training dataset form the negative bags, with one person corresponding to a negative bag. In our implementation, we use all the images of $S$ to build one positive bag, and for each negative bag we randomly sample $N_{neg}$ (typical value 2) images of the corresponding person. For each feature in the feature pool we initialize a weak classifier by SVM with linear kernel [14], considering its robustness to insufficient amount of training data (e.g. the typical number of positive samples in training a weak classifier is 10 in our experiments). After the above initialization, PMIL developed in Sect. 2.3 is used to learn an appearance model $c(\cdot)$.

The feature pool $\{f_p(r)\}_p$ is a set of color features defined on local regions. Specifically, we randomly initialize the central location and scale (between $5 \times 5$ and $30 \times 30$ in our experiments) of the regions by uniformly sampling on the person appearance to produce candidates for the learning process. Other region detection methods, such as SIFT [15], could be options for generating a set of regions. As to color feature, we use the spatial pyramid color representation (SPCR) proposed in [16]. It partitions the region spatially into subregions in a coarse-to-fine manner using spatial pyramid, and aggregate the color histograms over these subregions to form a holistic representation. Please refer to [16] for details.

4. Experiments

We evaluated our method for the multiple-shot vs. multiple-shot (MvM) case on the public ETHZ and CAVIAR+REID datasets. In the process of person re-identification, we first learn an appearance model as described in Sect. 3 for each person in the gallery set. Then the similarity between a probe person and each gallery person is computed by Eq. 5. The performance is shown in terms of recognition rate, by the Cumulative Matching Characteristic (CMC) curve which represents the correct matching rate in the top $K$ ranks.

Experimental setup - The parameters are set as follows: all the images are resized to $64 \times 32$; we fix the size of feature pool to 30, while finally selected 10 weak classifiers through appearance learning; the scale of SPCR feature is set to 2, in total 80 histogram bins for each region; The $\beta$ in Eq. 1 and the penalty factor $C$ in linear SVM are set respectively to 3 and 10 after cross-validation on ETHZ Sequence 1. All those parameters are kept fixed for all the datasets.

ETHZ dataset - This dataset [5] carries challenges such as false human detection, occlusion, intra camera illumination changes and scale changes, which are inevitable in practical automatic person re-identification and also our main concern in this paper. It consists of three sequences: Seq.1 contains 4857 images of 83 persons; Seq.2 contains 1936 images of 35 persons; Seq.3 contains 1762 images of 28 persons. To avoid that the images in probe set and gallery set may overlap in time, we randomly pick up 10 frames from the beginning for gallery set and from the end for probe set, which is the similar strategy used in [1] and [6].

We run the MvM ($M=10$) experiments 20 times and reported the average CMC curves in Fig. 2 (a-c). Comparisons were made against PLS [5], SDAFL [1], and LCP [6] which need no labelled pairwise training set from two cameras and presented state-of-the-art results on this dataset. To
In this paper, we present a novel pairwise multiple instance learning method for multiple-shot person re-identification, in view of corrupted images in an image sequence and visual ambiguity of human appearances. Future work will concentrate on two parts: (1) developing robust feature or feature comparison method to deal with illumination changes; (2) developing a hierarchical representation which incorporates multiple-view and multiple-shot images to tackle pose and viewpoint variations.

5. Conclusion

In this paper, we present a novel pairwise multiple instance learning method with MIL boosting proposed in [13]. PMIL is effective in handling corrupted images (see Fig. 3 (a)) and visual ambiguity (see Fig. 3 (b)).

CIAVIAR4REID - It contains 50 persons manually captured from a less controlled shopping center scenario by two cameras. Although occluded and false positive images are removed for each person, it presents much challenges caused by low resolutions, illumination changes, pose variations and visual ambiguity. For each person, we randomly selected 5 images from two cameras for the gallery and probe set, respectively. The averaged CMC curves over 20 trials are reported in Fig. 2 (d). We compare PMIL with AHPE [4] which produced state-of-the-art results on this dataset. As shown in Fig. 2 (d), PMIL shows superior results in top ranks, with a 16.51% increase for rank 1 and 6.33% increase for rank 5. AHPE is better from top 8. As shown in Fig. 3 (e)-(f), the most erroneous recognition of PMIL are due to significant color variations caused by different environments since we only employ color features. Incorporating other texture features, such as Gabor and Schmid filters [7], [8] may improve the performance. Visual ambiguity exists in this dataset due to a few people wearing black suits and drastic illumination changes (e.g. Fig. 3 (e)). To show the impact of pairwise learning scheme, we implemented a ‘MIL-based’ method which only replaced our learning model with MIL boosting proposed in [13]. PMIL obtains much better result than ‘MIL-based’ method, which is also consistent with the conclusion in [7], [8], [10] that the pairwise ranking is suitable in handling visual ambiguity.

References

[1] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, “Person re-identification by symmetry-driven accumulation of local features,” Proc. CVPR, pp.2360–2367, 2010.
[2] D. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, “Custom pictorial structures for re-identification,” Proc. BMVC, pp.68.1–68.11, 2011.
[3] S. Bak, E. Corvee, F. Bremond, and M. Thonnat, “Multiple-shot human re-identification by mean riemannian covariance grid,” Proc. 8th Advanced Video and Signal-Based Surveillance (AVSS), pp.179–184, 2011.
[4] 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, 2011.
[5] W. Schwartz and L. Davis, “Learning discriminative appearance-based models using partial least squares,” Proc. 22nd Brazilian Symposium on Computer Graphics and Image Processing, pp.322–329, 2009.
[6] S. Bak, E. Corvee, F. Bremond, and M. Thonnat, “Boosted human re-identification using riemannian manifolds,” Image and Vision Computing, vol.30, pp.443–452, 2011.
[7] B. Prosser, W. Zheng, S. Gong, T. Xiang, and Q. Mary, “Person re-identification by support vector ranking,” Proc. BMVC, pp.21.1–21.11, 2010.
[8] W. Zheng, S. Gong, and T. Xiang, “Person re-identification by probabilistic relative distance comparison,” Proc. CVPR, pp.649–656, 2011.
[9] D. Gray and H. Tao, “Viewpoint invariant pedestrian recognition with an ensemble of localized features,” Proc. ECCV, pp.262–275, 2008.
[10] M. Hirzer, P. Roth, M. Köstinger, and H. Bischof, “Relaxed pairwise learned metric for person re-identification,” Proc. ECCV, pp.780–793, 2012.
[11] S. Bak, G. Charpiat, E. Corvee, F. Brémond, and M. Thonnat, “Learning to match appearances by correlations in a covariance metric space,” Proc. ECCV, pp.806–820, 2012.
[12] M. O and L.P. T, “A framework for multiple-instance learning,” Proc. NIPS, pp.570–576, 1998.
[13] P. Viola, J.C. Platt, and C. Zhang, “Multiple instance boosting for object detection,” Proc. NIPS, pp.1417–1426, 2005.
[14] C.C. Chang and C.J. Lin, “LIBSVM: A library for support vector machines,” ACM Trans. Intelligent Systems and Technology, vol.2, pp.27:1–27:27, 2011.
[15] D. Lowe, “Object recognition from local scale-invariant features,” Proc. ICCV, pp.1150–1157, 1999.
[16] C. Liu, G. Wang, and X. Lin, “Person re-identification by spatial pyramid color representation and local region matching,” IEICE, Trans. Inf. & Syst., vol.E95-D, no.8, pp.2154–2157, Aug. 2012.