MULTI-VIEW IMPLICIT TRANSFER FOR PERSON RE-IDENTIFICATION

Wei Xu, Yijun Li, Chen Gong and Jie Yang

Institution of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, China

ABSTRACT

Implicit camera transfer (ICT), which models the multi-valued mappings between two specific and stationary cameras, is a descent solution for the person re-identification problem of the surveillance system. It has the properties of simplicity, computational efficiency and well utilizing negative training data. But it neglects the complementary relation between the descriptors of various views. And different appearance people have various most discriminative views among all the views, which are under diverse mappings. To tackle with this constraint, we model the multi-values mapping from different view independently, and fuse these transferring results of each view by LPBoost. Experimental results demonstrate that our scheme not only inherits most of the advantages (some sacrifice in speed, but still can run in real time for the same testing case in the ICT paper) of ICT but also obtains more discriminative mappings than ICT. In addition, our solution gains competitive performance on 2 challenging datasets.

Index Terms— person re-identification, multi-view fusing, implicit camera transfer

1. INTRODUCTION

Person re-identification, one of the main issues in video surveillance application, is to match persons observed from non-overlapping camera views based on image appearance. However, significant variations on the viewpoints, poses, illumination and appearance for the observed person make this problem pretty tough.

Unlike many recent works focus on general person re-identification problem, in which the purpose is to associate a person in any new place, we consider natural surveillance systems in which the cameras are stationary and specific.

Previous appearance-based solutions generally belong to one of these three groups: the invariant features schemes; the metric learning methods and the transformation ways. The first group usually uses hand crafted features [1, 2, 3, 4], such as color histogram, maximally stable color regions (MSCR), recurrent high-structured patches (RHSP), texture, shape and combination of them. And common distance metric is directly adopted to these features for matching or several discriminative classifiers could be employed on these representations to improve the performance. These classifiers can be based on Rank SVM [5], Partial Least Squares (PLS), Adaboost [6, 7] or multi-feature learning [8]. Recently learned invariant features [9, 10, 11] method of which performance is desirable, also draws more and more attention.

The second group often seeks for a metric in which the examples of the same people are close and those of the different people are far. Among them Logistic Discriminant Metric Learning (LDML) [12], Pairwise Constrained Component Analysis (PCCA) [13], Large Margin Nearest Neighbors (LMNN) [14], Information Theoretic Metric Learning (ITML)[15], and Keep It Simple and Straightforward Metric Learning (KISSME) [16] have achieved impressive results.

The last group tries to learn a transformation that transfers the descriptors of people from one camera view to those of another. Javeds work [17] shows that the transformation between different domain color histograms lies in a low-dimensional space, and assumes the existence of a metric with decent characters. Avraham proposes the implicit approaches [18], which models camera transfer by a binary relation R whose members are pairs (a, b) which represent the same person viewed from different cameras respectively. This relation obtained by training a binary non-linear SVM classifier with concatenated vectors (first one from camera A, the second one from camera B) equals to a multi-valued mapping function. Well utilizing massive number of the negative examples (pairs of descriptors whose members are associated with two different people and two different cameras) results in much more powerful transformation for this model. In addition, with the advantages of eliminating the background by recognizing the background associate with each camera, ICT could achieve decent performance with simple rough descriptors. While different appearance people have various most discriminative view features (color, shape, texture, etc), which are under diverse mappings. In order to obtain more discriminative mappings, we propose a MICT (short for multi-view implicit transfer) scheme to model these mappings.

In section 2 we describe the MICT algorithm, and then experiments are presented in section 3. At last we make the conclusion in section4.

This research is partly supported by NSFC, China (No: 61273258). Corresponding author: jieyang@sjtu.edu.cn, xuronaweizida@163.com.

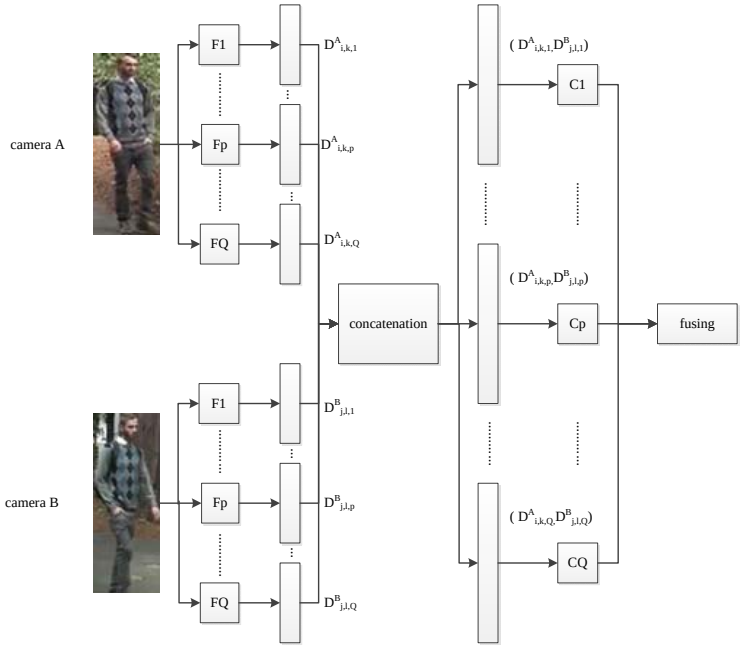

Fig. 1. Framework of our MICT solution for person re-identification problem, where F1 represents feature extractor 1, C1 represents classifier 1.

2. OUR APPROACH

We describe the MICT algorithm in this section. For two stationary cameras A and B with non-overlapping views, we try to match persons observed from different camera. Let $D^A_{i,k,p}$ denotes the view p descriptor of person is kth frame captured by camera A, and let $D^B_{j,l,p}$ denotes the view p descriptor of person js lth frame captured by camera B. Given Image I and J, we have Q pairs $(D^A_{1,k,1}, D^B_{1,l,1}), \ldots, (D^A_{i,k,Q}, D^B_{j,l,Q})$, we try to distinguish between positive pairs with same identity $(i = j)$, and negative pairs with different identity. MICT trains the binary non-linear SVM classifier of each view separately using concatenated negative and positive pairs of vectors, then fuse classifier result of each view by LPBoost. When new querying pairs come, this trained model could be applied for classifying. The frame work of our approach is illustrated in figure 1. The detailed illustration of MICT is as follows.

2.1. MICT training

The input:
A set $\{D^A_{i,k,p} | i = 1, \ldots, n; k = 1, \ldots, m^A_i; p = 1, \ldots, Q\}$ of vectors representing examples of n people by camera A.
A set $\{D^B_{i,k,p} | i = 1, \ldots, n; k = 1, \ldots, m^B_i; p = 1, \ldots, Q\}$ of vectors representing examples of n people by camera B.
Vectors concatenation:
Q view sets of positive instances $\{[D^A_{i,k,p} || D^B_{i,l,p}] | i = 1, \ldots, n; k = 1, \ldots, m^A_i; l = 1, \ldots, m^B_i; p = 1, \ldots, Q\}$.
Q view sets of negative instances $\{[D^A_{i,k,p} || D^B_{j,l,p}] | i \neq j; i, j = 1, \ldots, n; k = 1, \ldots, m^A_i; l = 1, \ldots, m^B_i; p = 1, \ldots, Q\}$.
Parameter learning:
Use cross validation to obtain optimal c_p and γ_p of each views nonlinear SVM, where $p = 1, \ldots, Q$ details refer section 3.3.
Fusing:
Use LPBoost to get the weight w_p for each view, where $p = 1, \ldots, Q$. details refer section 2.2.
Output
Optimal c_p and γ_p parameters for each views nonlinear SVM. w_p, fusing weight for view p where $p = 1, \ldots, Q$.

2.2. LPBoost for view fusing

The goal of fusing is to learn the optimize w_p using the following linear program (LP).

$$\begin{aligned} &\min_{w,\xi} -\rho + \frac{1}{M}\sum_{i=1}^{M}\xi_i \\ &sb.t. y_i \sum_{p=1}^{Q} w_p f_p(x_i) + \xi_i \geq \rho, i = 1, \ldots, M \\ &\sum_{p=1}^{Q} w_p = 1, w_p \geq 0, p = 1, \ldots, Q \end{aligned} \quad (1)$$

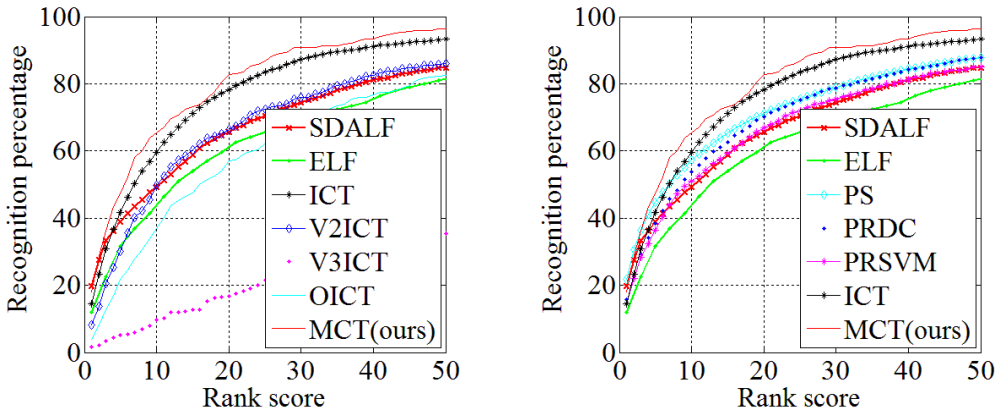

Fig. 2. CMC curves for VIPeR dataset, where V2ICT represents carrying out ICT with the features of view 2.

With $\{\xi_1, \ldots, \xi_M\}$ being slack variables, ρ is the margin parameter. $f_p(x_i)$ represents the classification result for the view p of sample x_i.

2.3. Testing

The input:
A set $\{D^A_{I,k,p}|k = 1, \ldots, m^A_I; p = 1, \ldots, Q\}$ of vectors representing examples of people I by camera A.
A set $\{D^B_{J,k,p}|k = 1, \ldots, m^B_J; p = 1, \ldots, Q\}$ of vectors representing examples of people J by camera B.
Output:

$$y == \sum_{k=1}^{m^A_I} \sum_{l=1}^{m^B_J} \sum_{p=1}^{Q} \frac{w_p y_{k,l,p}}{m^A_I m^B_J} \tag{2}$$

Where $y_{k,l,p}$ is the decision value obtain from classifier p

3. EXPERIMENT

3.1. Dataset

We evaluated our work on two public dataset, VIPeR dataset [19] and CAVIAR4REID dataset [20].

The VIPeR dataset, one of the most challenge and popular person re-identification datasets, contains 632 pedestrian pairs, each of which has two images captured in outdoor academic environment. Its challenging due to the significant variations on illuminations, poses, viewpoints for the observed person. Most of the images show viewpoints large than 90 degrees and all the images are normalized to 128*48 for experiments.

CAVIAR4REID dataset includes 50 pedestrians captured by two different cameras. For each person in each camera there are 10 available appearances. All images are resized to 3296 pixels.

3.2. Experimental settings

In fair comparison with ICT [18], we use the same color features, classifiers and evaluation method as ICT [18].
Features:
View 1 (color): the same as ICT [18],we use 150 dimensions (5 horizontal strips for each bounding box surrounding the people, each strip is represented by a HSV histogram with 10 bins for each color components) feature vector for each image.
View 2 (texture): In fair comparison with ICT [18], we also adapt 150 dimensions (5 horizontal strips for each bounding box surrounding the people, each strip is represented by a Gabor histogram with 10 bins for each color components H, S, V) feature vector for each image
View 3 (shape): Also in fair comparison with ICT [18], we employ 150 dimensions (5 horizontal strips for each bounding box surrounding the people, each strip is represented by a HOG histogram with 10 bins for each color components H, S, V) feature vector for each image.
Classifiers:
RBF kernel binary SVM
Evaluation method:
Average Cumulative Match Characteristic (CMC) curves; rank (i), percentage of true matches within the first i ranked examples; the CMC-expectation measure, mean rank of the true match; and the nAUC (normalized Area Under Curve).

3.3. Result and discussion

For the experiment on VIPeR dataset, we randomly split the set of 632 images pairs into equal testing and training sets 10 times, perform cross validations with k=30, negative examples numbers for each person, during the training process, which is the same as ICT [18].
Different measures represent various aspects of the algorithms performance. Lower ranks are very important in easy case,

method	expectation	Rank(1)	Rank(10)	Rank(20)	nAUC
SDALF	25.5	19.9	49.4	65.7	92.2
ELF	28.9	12	44	61	91.2
PS	21.2	**21.8**	57.2	71.2	93.6
PRDC	21.5	15.7	53.9	70.1	93.5
PRSVM	27.9	14.6	50.9	66.8	91.4
ICT	15.9	14.4	59.7	78.3	95.3
V2ICT	24.3	8.5	49.7	66.7	92.6
V3ICT	92.1	1.3	9.5	18.4	71.1
OICT	28.9	4.1	37.0	56.6	91.2
MICT(OURS)	**13.0**	17.7	**64.6**	**81.7**	**96.2**

Table 1. Testing results on the VIPeR dataset

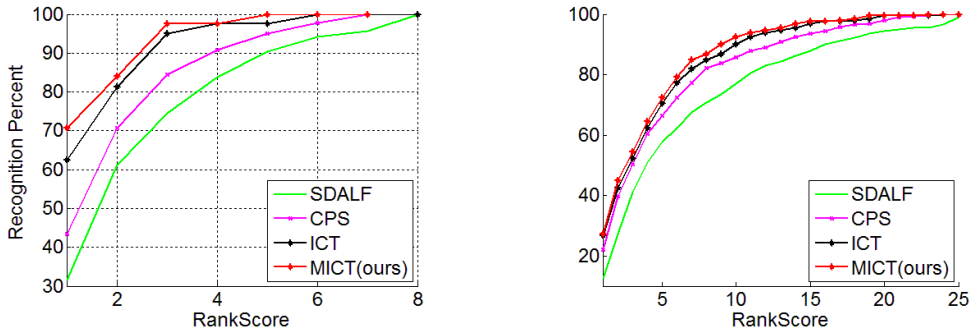

Fig. 3. CMC curves for CAVIAR4REID dataset. Left: Training with 42 people and 8-person test set. Right: Training with 25 people and 25-person test set.

while the CMC- expectation shows the average performance of the algorithm, which is priority in the difficult case since it associates with the average human effort in the real-life applications.

Mappings of different views lie in diverse local domain: From Figure 2 (a), using ICT with combination descriptor of various views (OICT) results in inferior performance to those of each only single view (ICT,V2ICT) with the same number of training data. In addition, as the performance of view 3 is too weak, we dont consider it in MICT for all the experiments.

Different views have complementary relation with each other: Figure 2 (left) and Table 1 reflect that view1 ICT is better than that of view2 and view 3, which is mainly attribute to the coarse texture image of outdoor surveillance system and little shape variance between humans makes color information become most discriminative features for most cases. We could also see that MICT boost the performance significantly, which implies that various view features capture different and complementary properties of the image. From Figure 2 (right) and Table1, we could view that MICT does not achieve the best rank 1 performance, but performs best for all ranks 8 and up.

For the CAVIAR4REID dataset, we adapt all the same setting as those of ICT [18], but employ MICT for training and testing. Figure 3 illustrate that MICT achieve the best performance. In addition, inherit from ICT, the feature extraction and classifier calls of MICT are very simple can run in real time.

4. CONCLUSION

In this paper, we extend implicit transfer for person re-identification problem of the surveillance system to multi-view. Apart from the simple and fast carrying out properties inherited from ICT, MICT is an effective and efficient way to fuse the complementary view descriptors, which lies in dissimilar local domain, for improving the re-identification rate. Experiments on two challenging datasets show the competitive performance of our work. Yet there are still several questions needed to be answered, such as how to boost the performance with less training data or positive data, how to choose the complementary views and how to joint training those views so that remarkable improvement would be obtained.

5. REFERENCES

[1] Loris Bazzani, Marco Cristani, and Vittorio Murino, "Symmetry-driven accumulation of local features for human characterization and re-identification," *Computer Vision and Image Understanding*, vol. 117, no. 2, pp. 130–144, 2013.

[2] Dandan Xu and Huicheng Zheng, "Person re-identification by multi-resolution saliency-weighted color histograms and local structural sparse coding," in *Image and Graphics (ICIG), 2013 Seventh International Conference on*. IEEE, 2013, pp. 477–482.

[3] Bingpeng Ma, Yu Su, and Frederic Jurie, "Covariance descriptor based on bio-inspired features for person re-identification and face verification," *Image and Vision Computing*, vol. 32, no. 6, pp. 379–390, 2014.

[4] Bingpeng Ma, Yu Su, and Frédéric Jurie, "Bicov: a novel image representation for person re-identification and face verification," in *British Machive Vision Conference*, 2012, pp. 11–pages.

[5] Bryan Prosser, Wei-Shi Zheng, Shaogang Gong, Tao Xiang, and Q Mary, "Person re-identification by support vector ranking.," in *BMVC*, 2010, vol. 2, p. 6.

[6] Douglas Gray and Hai Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Computer Vision–ECCV 2008*, pp. 262–275. Springer, 2008.

[7] Slawomir Bak, Etienne Corvee, Francois Brémond, and Monique Thonnat, "Person re-identification using haar-based and dcd-based signature," in *Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on*. IEEE, 2010, pp. 1–8.

[8] Dario Figueira, Loris Bazzani, Ha Quang Minh, Marco Cristani, Alexandre Bernardino, and Vittorio Murino, "Semi-supervised multi-feature learning for person re-identification," in *Advanced Video and Signal Based Surveillance (AVSS), 2013 10th IEEE International Conference on*. IEEE, 2013, pp. 111–116.

[9] Rui Zhao, Wanli Ouyang, and Xiaogang Wang, "Unsupervised salience learning for person re-identification," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 3586–3593.

[10] Rui Zhao, Wanli Ouyang, and Xiaogang Wang, "Learning mid-level filters for person re-identification," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 144–151.

[11] Rui Zhao, Wanli Ouyang, and Xiaogang Wang, "Person re-identification by salience matching," in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 2528–2535.

[12] Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid, "Is that you? metric learning approaches for face identification," in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 498–505.

[13] Alexis Mignon and Frédéric Jurie, "Pcca: A new approach for distance learning from sparse pairwise constraints," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2666–2672.

[14] Kilian Q Weinberger and Lawrence K Saul, "Distance metric learning for large margin nearest neighbor classification," *The Journal of Machine Learning Research*, vol. 10, pp. 207–244, 2009.

[15] Jason V Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S Dhillon, "Information-theoretic metric learning," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 209–216.

[16] Martin Kostinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof, "Large scale metric learning from equivalence constraints," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2288–2295.

[17] Omar Javed, Khurram Shafique, and Mubarak Shah, "Appearance modeling for tracking in multiple non-overlapping cameras," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. IEEE, 2005, vol. 2, pp. 26–33.

[18] Tamar Avraham, Ilya Gurvich, Michael Lindenbaum, and Shaul Markovitch, "Learning implicit transfer for person re-identification," in *Computer Vision–ECCV 2012. Workshops and Demonstrations*. Springer, 2012, pp. 381–390.

[19] Douglas Gray, Shane Brennan, and Hai Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*. Citeseer, 2007, vol. 3.

[20] Dong Seon Cheng, Marco Cristani, Michele Stoppa, Loris Bazzani, and Vittorio Murino, "Custom pictorial structures for re-identification.," in *BMVC*. Citeseer, 2011, vol. 2, p. 6.