Learning Implicit Transfer for Person Re-identification

Tamar Avraham, Ilya Gurvich, Michael Lindenbaum, and Shaul Markovitch

Computer Science Department, Technion - I.I.T., Haifa 32000, Israel

Abstract. This paper proposes a novel approach for pedestrian re-identification. Previous re-identification methods use one of 3 approaches: invariant features; designing metrics that aim to bring instances of shared identities close to one another and instances of different identities far from one another; or learning a transformation from the appearance in one domain to the other. Our *implicit* approach models camera transfer by a binary relation $R = \{(x, y)|x$ and y describe the same person seen from cameras A and B respectively$\}$. This solution implies that the camera transfer function is a multi-valued mapping and not a single-valued transformation, and does not assume the existence of a metric with desirable properties. We present an algorithm that follows this approach and achieves new state-of-the-art performance.

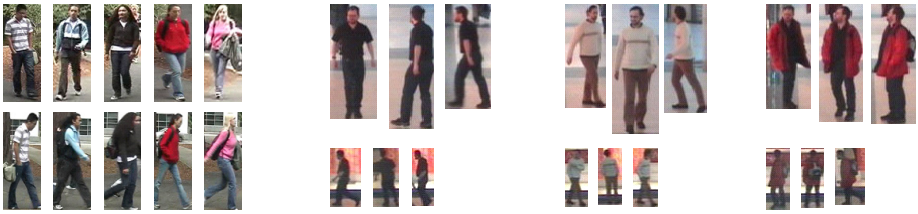
1 Introduction

The re-identification problem has received increasing attention in the last five to six years, especially due to its important role in surveillance systems. It is desirable that computer vision systems will be able to keep track of people after they have left the field of view of one camera and entered the field of view of the next, even when these fields of view do not overlap.

We make the distinction between the *general re-identification* problem, in which the goal is to re-identify a person in any new location, and the *camera-specific re-identification* problem, in which the goal is to provide a solution for a specific site. In this work we tackle the second goal. Given a pair of stationary cameras, A and B, capturing two non-overlapping regions, and a training set of annotated people captured by those two cameras, our objective is to recognize correspondence between the appearance of a never-before-seen person in camera A and his or her appearance in camera B. As can be seen in the examples in Fig. 1, learning the domain of the camera-specific transformations may be very informative. Each camera is associated with a limited variety of backgrounds, illumination conditions, and sometimes human poses. We propose an algorithm that exploits these properties. Our algorithm is based on the observation that the transfer between two cameras is a multi-valued mapping which can be estimated using implicit function learning.

Previous re-identification methods have used solutions that belong to one of three families of methods: those that seek for invariant features; those that seek for a metric in which instances associated with the same person are close and

A. Fusiello et al. (Eds.): ECCV 2012 Ws/Demos, Part I, LNCS 7583, pp. 381–390, 2012.
© Springer-Verlag Berlin Heidelberg 2012



(a) VIPeR examples (b) CAVIAR4REID examples

Fig. 1. (a) Examples from the VIPeR dataset: five people captured by one camera (top row) and another camera (second row). (b) Examples from the CAVIAR4REID dataset: three people captured by one camera in multi-shots (top), and the same three people captured by a second camera (bottom). We see that the background, illumination, resolution and sometimes pose are camera dependent.

instances associated with different people are far; and those that try to learn a transformation, i.e., a function, that transfers the descriptors of people as they 'move' from one camera to the other. Our *implicit* approach models camera transfer by a binary relation $R = \{(x, y)|x$ and y describe the same person seen from cameras A and B respectively$\}$. This solution implies that the camera transfer function is a multi-valued mapping and not a single-valued transformation. Moreover, it does not assume the existence of a metric that can bring all instances of shared identities close to one another and instances of different identities far from one another. Instead, given a person's appearance described by a feature vector of length k, the binary relation models a (not necessarily continuous) sub-space in $\mathbb{R}^{2k}$. That is, we divide the $\mathbb{R}^{2k}$ space to 'positive' regions (belonging to the relation) and 'negative' regions (not belonging to the relation). As a result, this modeling does not build only on a feature-by-feature comparison, but models also dependencies between different features.

Our algorithm, denoted *ICT* (short for *Implicit Camera Transfer*), models the binary relation by training a (non-linear) binary classifier with concatenations of pairs of vectors, the first describing an instance associated with camera A, and the second describing an instance associated with camera B. One class includes the *positive pairs* – pairs of instances capturing the same person with the two different cameras, and the second class includes the *negative pairs* – pairs of instances whose members are associated with two different people and two different cameras. This algorithm, although so simple, provides state-of-the-art results. It can work for single-shots per person as well as for multi-shots (video).

We consider the optimal number of negative examples to use for training and show that utilizing the more abundant negative examples allows us to learn the transfer associated with two cameras from rather small sets of inter-camera example pairs. The ICT algorithm simultaneously learns to distinguish between changes that are camera and location dependent and those that depend on the person's identity. This allows the use of simple features extracted from the bounding boxes surrounding the people, without incorporating high-level, risky, and time consuming, preprocesses.

In Sec. 2 we review related work, in Sec. 3 we describe the ICT algorithm, and in Sec. 4 we describe the experiments on the VIPeR [1] and the CAVIAR4REID [2] datasets. Sec. 5 concludes.

2 Related Work

Object re-identification is a challenge that has been receiving increasing attention (e.g., face re-identification [3, 4], car re-identification [5]). Person, or pedestrian, re-identification is a special focus of recent research, mainly due to its important role in surveillance systems. One common approach proposes *invariant* features that are stable to illumination, resolution, pose, and background changes. A 'same' or 'not-same' decision is then made using some fixed distance measure. In [6], for instance, normalized color and salient edgel histograms are the basis for matching segmented parts. In [7] a similarity measure based on principal axis correspondence is used. In [8] the similarity between two sets of signatures, each describing a person's video track, is measured by the width of the margins of a linear SVM. In [9] features extracted from a person's track are compacted with an epitomic analysis that recognizes the presence of recurrent local patterns. In [10] each semantic body part is described by a signature composed of features that are stable to changes in pose, viewpoint, resolution and illumination.

Some recent methods focus on learning characteristics of the similarity between feature vectors describing two instances of the same person against that of two vectors describing instances of different people. These similarity-based methods usually use the absolute distance as the characteristic to be learned. The ELF method [11] models the feature-wise difference distribution using Adaboost for feature selection and classification. In [12] it is observed that what matters is not the similarity itself, but the relative similarity: positive pairs should be ranked higher than negative pairs. The goal is to weigh the features in a way that maximizes the difference between absolute differences of negative pairs and absolute differences of positive pairs. The method in [13] takes a similar approach using probabilistic modeling. In contrast to these methods, we do not assume that greater similarity implies 'same'. Moreover, as opposed to methods that use the absolute distances as a starting point, or that compare histograms bin-by-bin, we do not perform a feature-by-feature comparison, and allow dependencies between any two features in the two input descriptors.

Most of the aforementioned methods try to solve the *general re-identification* setup. When two specific cameras are considered, the correlation between the cameras' identities and the expected background, illumination, and human pose may be exploited. The following situation may then be considered: An instance associated with person i in camera A undergoes a transformation function T and is then captured as an instance in camera B. In [14] it was shown that the domain of possible transformations between color histograms lies in a low-dimensional subspace. This paper is based on modeling the transformation as well. We take a different approach and argue that the transformation is a multi-valued function (or a binary relation). Moreover, unlike the approach in [14], which uses only

positive examples, our approach allows the utilization of negative examples to better model the transformation domain.

Some methods start with a pre-process for separating the people from the background (e.g., [6, 8, 10]) and some also attempt to divide the person into semantic parts. For instance, [6] begins with a spatio-temporal segmentation process and then searches for correspondence between different segments. In [10], regions are separated into parts corresponding to head, torso and legs by vertical asymmetries. In [2] pictorial structures are extracted in order to fit corresponding body parts. These high-level processes indeed lead to more accurate recognition but may also lead to mistakes that will then be dragged into the training and classification stages. In our work we use bounding boxes surrounding the people and yet achieve very good performance. This is because our algorithm is implicitly trained to filter out the background by recognizing the background associated with each camera as person-independent. This approach is not limited to re-identifying people as it does not rely on a specific model for their appearance. As a result, it also allows items carried by the people (e.g., bags) to be used as cues without additional explicit analysis. Note that high-level semantic analysis requires processing time that is unlikely to allow real-time performance, while the method proposed here can be used for real-time re-identification.

3 Implicitly Learning Inter-camera Transfer

In this section we describe the ICT algorithm. Given that there are two stationary cameras A and B, covering two non-overlapping regions of a site, our algorithm is trained to find correspondence between people captured by the two cameras. Let $V^A_{i,k}$ describe the k'th appearance of a person with identity i captured by camera A, and let $V^B_{j,l}$ describe the l'th appearance of a person with identity j captured by camera B. Given a pair $(V^A_{i,k}, V^B_{j,l})$, the goal is to distinguish between *positive* pairs with the same identity ($i = j$), and *negative* pairs ($i \neq j$). Our algorithm trains a binary classifier using concatenations of such positive and negative pairs of vectors coming from training data. Then it classifies new such pairs by querying the classifier on their concatenations. See Fig. 2. A detailed description of the algorithm follows.

The ICT Algorithm

The Training Stage:

The Input:

- A set $\{V^A_{i,k} | i = 1, ..., n; k = 1, ..., m^A_i\}$ of vectors describing instances of n people captured by camera A.
- A set $\{V^B_{i,k} | i = 1, ...n; k = 1, ..., m^B_i\}$ of vectors describing instances of the same n people captured by camera B.

That is, for each person and each camera we may be provided with a few descriptor vectors, each associated with his or her appearance in a different video frame.

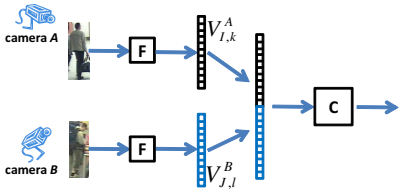


Fig. 2. Illustration of the classification stage of the ICT algorithm. From each of the instances captured by cameras A and B, features are extracted (F). The concatenation of those two feature vectors, $V^A_{I,k}$ and $V^B_{J,l}$, is the input to the classifier C.

Let $[a\|b] = (a_1, ..., a_n, b_1, ..., b_m)$ denote the concatenation of vectors $a = (a_1, ..., a_n)$ and $b = (b_1, ..., b_m)$. The training input for the binary classifier is:

- A set of positive examples $\{[V^A_{i,k}\|V^B_{i,l}]\ |,\ i \in \{1, ..., n\}, k \in \{1, ..., m^A_i\}, l \in \{1, ..., m^B_i\}\}$.
- A set of negative examples $\{[V^A_{i,k}\|V^B_{j,l}]\ |i \neq j, i, j \in \{1, ..., n\}, k \in \{1, ..., m^A_i\}, l \in \{1, ..., m^B_j\}\}$.

For the type of descriptors used and for details about the classifiers used in our experiments, see Sec. 4. Note that there are $\sum_{i=1}^{n} m^A_i m^B_i$ positive examples, while there is a quadratic number $\sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} m^A_i m^B_j$ of negative examples. We do not use all the negative examples but show that even a fraction of them significantly contribute to the success of the algorithm. See Sec. 4.2.

The Classification/Decision Stage:

The Input:

- A set $\{V^A_{I,k} | k = 1, ..., m^A_I\}$ of vectors describing a person's track as captured by camera A.
- A set $\{V^B_{J,l} | l = 1, ..., m^B_J\}$ of vectors describing a person's track as captured by camera B.

The Decision: Apply the trained classifier on each of the concatenations $[V^A_{I,k}\|V^B_{J,l}]$, $k = 1, ..., m^A_I$, $l = 1, ..., m^B_J$. One possibility is to use the binary classifications and to output a binary decision by their majority. However, more informative is to output a continuous score that allows different candidate matches to be ranked. The way to obtain such a score depends on the classifier used. In our experiments we use an SVM as the classifier and output the average of the decision values: let $y_{k,l}$, $k = 1, ..., m^A_I$, $l = 1, ..., m^B_J$ be the decision values obtained from the classifier. The algorithm returns the mean $Y = \sum_{k=1}^{m^A_I} \sum_{l=1}^{m^B_J} y_{k,l}/m^A_I m^B_J$.

4 Experiments

After providing additional implementation details (Sec. 4.1), we test ICT's performance as a function of the number of negative examples utilized for training

(Sec. 4.2). Then we compare its performance to that of the latest state-of-the-art for the single-shots case on the VIPeR dataset (Sec. 4.3). In Sec. 4.4 we compare ICT's performance to that of recent state-of-the-art for multi-shot setups on the CAVIAR4REID dataset[1,2].

4.1 Implementation Details

Features. We use a common and simple description of bounding boxes surrounding the people: each bounding box is divided into five horizontal stripes. Each stripe is described by a histogram with 10 bins for each of the color components H, S, and V. This results in feature vectors with 150 dimensions. We did not focus on finding optimal features. Any alternative descriptors (e.g., textural descriptors [13], temporal features [6], or semantic features [2, 10]) can be easily used as well, and may further improve the algorithm's performance.

Classifiers. We use an RBF kernel binary SVM as the classifier for the concatenated vectors. In one of our experiments below we test the use of a one-class-SVM also with an RBF kernel. We use the implementation provided by LibSVM [16].

Evaluation Methods. In the experiments described below we output and compare average Cumulative Match Characteristic (CMC) curves. This is the most widely accepted way to evaluate re-identification algorithms. For each person in the test set, each algorithm ranks the matching of his or her appearance in camera A with the appearances of all the people in the test set in camera B. The CMC curve summarizes the statistics of the ranks of the true matches. For quantitative comparison we use the measure rank(i), which denotes the percentage of true matches found within the first i ranked instances, the CMC-expectation measure, which is the mean rank of the true match, and the nAUC (normalized Area Under Curve).

4.2 The Role of Negative Examples

As mentioned in Sec. 3, the number of negative examples that can be used for training is quadratic in the number of positive examples. Using all the negative examples can lead to a strong bias and is computationally expensive. Do we need all the negative examples? Do we need negative examples at all? In our first set of experiments we tested the contribution of the negative examples by checking the algorithm's performance as a function of the number of negative examples used for training. These experiments use the VIPeR dataset, the most commonly used dataset for evaluating re-identification methods. It contains 632 pedestrian image pairs. Each pair contains two images of the same individual seen from different

[1] The Matlab source code used in all the experiments is available in `http://www.cs.technion.ac.il/~tammya/Reidentification.html`.

[2] We are aware of the set of data annotated by [15] and corresponding to 119 people appearing in the i-LIDs videos. That set includes a few instances for each person without indication of the camera's identity. It was thus unsuitable for our setup.

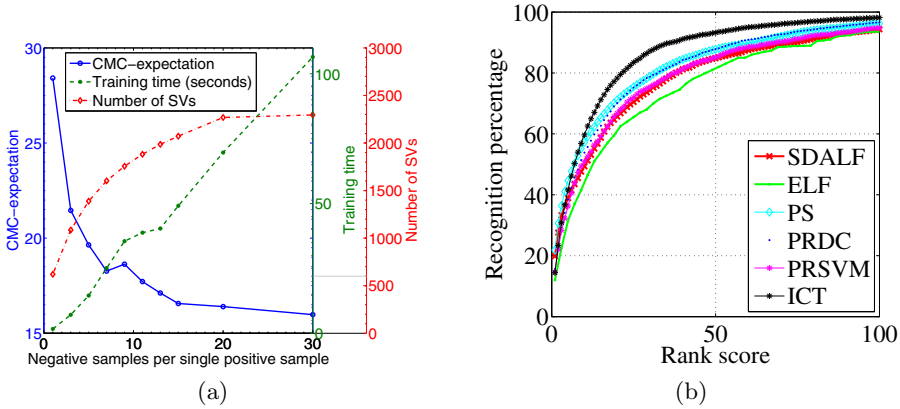


Fig. 3. (a) ICT's performance on the VIPeR dataset as a function of κ, the number of negative examples per single positive example, measured by the CMC expectation, the training time, and the support-vectors used by the SVM. (b) CMC curves comparing ICT's results on VIPeR with recent state-of-the-art reported in [2, 10–13].

viewpoints by two cameras. See examples in Fig. 1(a). We perform a 2-fold cross-validation, dividing the 632 pedestrians into equal-size training and test sets. We repeat this process four times with different random choices for the sets. The number of positive examples available for training is $P = 316$ (one concatenated pair for each person). We test the performance of ICT for different numbers of negative examples $N = \kappa P$, where $\kappa = 1, 3, 5, 7, 9, 11, 13, 15, 20, 30$. That is, for each positive example associated with person i, κ of the $N - 1$ negative examples involving person i's appearance in camera A are randomly selected. Each training involves a parameter learning stage: we learn the optimal c and γ parameters for the RBF SVM by a 4-fold cross-validation inside the training set, searching for the parameters that result in the lowest CMC-expectation.

See Fig. 3(a) for ICT's performance as a function of κ. It reports the CMC expectation, the training time, and the number of support vectors found by the SVM. We see that the expectation drops as κ increases, at a high slope for small κ's and at an almost zero slope for $\kappa > 15$. We also see a similar convergence in the number of support vectors, which means that adding more than a certain number of negative examples does not add information. Note that the computation time for training grows linearly with κ. We also tested a variation of the algorithm that learns only from positive examples using one-class-SVM (i.e, $\kappa = 0$). The one-class SVM test, which followed a similar procedure, yields a CMC expectation value of 45.6, which is worst than the CMC expectation achieved by the binary SVM for $\kappa = 1$.

We learned that (a) not all negative examples are essential and training time can be saved by selecting only some of them; (b) the negative examples play an important role in compensating for the usually small number of positive examples, by helping in defining the borders of the "cloud" formed by the positive transformations.

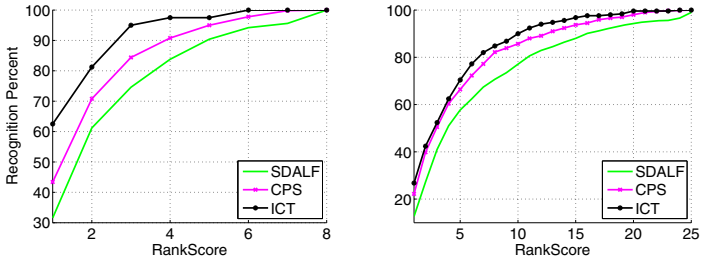
Table 1. Results of ICT on the VIPeR dataset compared to the models in [2, 10–13].

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**

4.3 Comparing to State-of-the-Art on VIPeR

In order to compare ICT's performance on the VIPeR dataset with that of recent work we repeated the above experiment, this time performing cross validations for 10 random splits, using $\kappa = 30$. See Fig. 3(b). The results of the ELF [11] and the SDALF [10] algorithms were kindly provided by the authors of [10]. The results of PRDC were kindly provided by the authors of [13]. The results of the PS based algorithm were kindly provided by the authors of [2]. The results of PRSVM are those presented in [12]. See Table 1 for a comparison of the CMC expectation, rank(1), rank(10), rank(20), and nAUC of the different methods. The CMC-expectation and the nAUC are much better for ICT than for all previous methods. ICT does not achieve the best rank(1) performance, but performs best for all ranks 8 and up.

The different measures show different aspects of the algorithm's performance. We argue that while the lower ranks are desirable, they are not achievable for the majority of the cases, which makes the higher ranks and the CMC expectations at least as important. The few lower ranks only reflect the algorithm's performance on the easy cases, while the CMC-expectation reflects the average human operator effort, and together with higher ranks, measures the algorithm's performance for average and difficult cases. The higher ranks on the VIPeR data may be more relevant for realistic applications in which the set of candidates contain only a few people. Consider, for example, a common surveillance scenario in which a suspect is recognized as he is captured by a certain camera, and we wish to continue tracking him. Yet the tracker has lost him because of a short occlusion, or because he passed through a 'blind' area not covered by any camera. Now, we can define a set of possible candidates for this 'lost' suspect. The number of candidates in such a case will be rather small. Hence, instead of 316 candidates (as tested in the VIPeR experiment setup), we may have, say, 8 candidates on whom we can apply a re-identification algorithm. If we scale the CMC curves accordingly, we may expect, on the average, that rank(1) for 8 candidates is approximately equivalent to rank(40) for 316 candidates. In this case ICT promises 91% success, while the next runner up promises success of only 84%.



(a) Training with 42 people and an 8-person test set

(b) Training with 25 people and a 25-person test set

Fig. 4. CMC curves comparing ICT's results on CAVIAR4RfEID with those of SDALF [10] and CPS [2]

4.4 Comparing to State-of-the-Art on CAVIAR4REID

In this section we compare the performance of ICT with that of state-of-the-art on the newly released CAVIAR4REID dataset. This dataset includes 50 pedestrians captured by two different cameras. For each person in each camera there are 10 available appearances. We report results for two setups in Fig. 4, demonstrating the relative performance as a function of the size of the training data available. In the first setup (Fig. 4(a)), 42 people are included in the inter-camera training set and 8 others in the test set. In the second setup (Fig. 4(b)), the 50 people are equally divided into a training set of 25 and a test set of 25. For each setup, we average results on 10 random divisions. Our results are compared to those of SDALF and CPS reported in [2]. In [2] the test set consists of all 50 inter-camera people. We estimated the performance for test sets of 25 and 8 by normalizing the CMC curves reported in [2][3]. We see that for training sets of 25 people our algorithm meets the state-of-the-art performance of CPS, and outperforms SDALF and CPS for larger training sets. Note that the SDALF and CPS methods include high-level semantic analysis that requires heavy processing during the classification stage, while our classification stage includes very basic feature extraction and classifier calls that can run in real-time.(For instance, the runtime of an SVM RBF classifier with $\sim$ 1000 support vectors on one concatenated vector is 0.8 milliseconds on a standard laptop.)

5 Discussion

This paper considers the re-identification task and contributes the observation that the transfer between two cameras is a multi-valued mapping (a binary relation) which can be estimated using implicit function learning. We show that utilizing the more abundant negative examples allows us to learn the transfer

[3] If a person's true match was rated m among n people, then on the average it will be ranked $(m-1)*(k-1)/(n-1)+1$ among k people.

associated with two cameras from rather small sets of inter-camera example pairs. The algorithm yields an extremely fast classifier. We present new state-of-the-art re-identification performance.

The paper focuses on the camera-specific context, which enables the algorithm to implicitly "filter out" the irrelevant, person-independent, features without high-level semantic analysis. Yet we intend to test the utility of combining analysis of this sort in our algorithm, with the goal of finding the optimal combination that will bring maximum performance with minimum training.

Acknowledgments. This work was supported by the VULCAN consortium, a Magnet project administrated by the Office of the Chief Scientist at the ministry of Industry and Trade, Israel. The authors would like to thank Loris Bazzani, Dong Seon Cheng, and Wei-Shi Zheng for sharing their data and/or results.

References

1. Gray, D., Brennan, S., Tao, H.: Evaluating appearance models for recognition, reacquisition, and tracking. In: PETS Workshop in Conjunction with ICCV (2007)
2. Cheng, D.S., Cristani, M., Stoppa, M., Bazzani, L., Murino, V.: Custom pictorial structures for re-identification. In: BMVC (2011)
3. Pinto, N., DiCarlo, J., Cox, D.: How far can you get with a modern face recognition test set using only simple features? In: CVPR (2009)
4. Wolf, L., Hassner, T., Taigman, Y.: Descriptor based methods in the wild. In: ECCV (2008)
5. Ferencz, A., Learned-miller, E., Malik, J.: Learning to locate informative features for visual identification. IJCV 77, 3–24 (2008)
6. Gheissari, N., Sebastian, T., Hartley, R.: Person reidentification using spatiotemporal appearance. In: CVPR (2006)
7. Hu, W., Hu, M., Zhou, X., Tan, T., Lou, J.: Principal axis-based correspondence between multiple cameras for people tracking. PAMI 28, 663–671 (2006)
8. Cong, D., Khoudour, L., Achard, C., Meurie, C., Lezoray, O.: People re-identification by spectral classification of silhouettes. Signal Processing 90 (2010)
9. Bazzani, L., Cristani, M., Perina, A., Farenzena, M., Murino, V.: Multiple-shot person re-identification by HPE signature. In: ICPR, pp. 1413–1416 (2010)
10. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person re-identification by symmetry-driven accumulation of local features. In: CVPR (2010)
11. Gray, D., Tao, H.: Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 262–275. Springer, Heidelberg (2008)
12. Prosser, B., Zheng, W., Shaogang, G., Xiang, T.: Person re-identification by support vector ranking. In: BMVC (2010)
13. Zheng, W., Gong, S., Xiang, T.: Person re-identification by probabilistic relative distance comparison. In: CVPR (2011)
14. Javed, O., Khurram, S., Mubarak, S.: Appearance modeling for tracking in multiple non-overlapping cameras. In: CVPR (2005)
15. Zheng, W., Gong, S., Xiang, T.: Associating groups of people. In: BMVC (2009)
16. Chang, C., Lin, C.: LIBSVM: a library for support vector machines (2001), Software, `http://www.csie.ntu.edu.tw/~cjlin/libsvm`