Transitive Re-identification

Yulia Brand[1]
yulias@techunix.technion.ac.il
Tamar Avraham[2]
http://www.cs.technion.ac.il/~tammya
Michael Lindenbaum[2]
http://www.cs.technion.ac.il/~mic

[1] Electrical Engineering Department
[2] Computer Science Department

Technion - I.I.T.
Haifa 32000, Israel

Abstract

Person re-identification accuracy can be significantly improved given a training set that demonstrates changes in appearances associated with the two non-overlapping cameras involved. Here we test whether this advantage can be maintained when directly annotated training sets are not available for all camera-pairs at the site. Given the training sets capturing correspondences between cameras A and B and a different training set capturing correspondences between cameras B and C, the Transitive Re-IDentification algorithm (TRID) suggested here provides a classifier for (A,C) appearance pairs. The proposed method is based on statistical modeling and uses a marginalization process for the inference. This approach significantly reduces the annotation effort inherent in a learning system, which goes down from $O(N^2)$ to $O(N)$, for a site containing N cameras. Moreover, when adding camera $(N+1)$, only one inter-camera training set is required for establishing all correspondences. In our experiments we found that the method is effective and more accurate than the competing camera invariant approach.

1 Introduction

Person re-identification (ReID) consists of recognizing individuals over different camera views. The ReID problem has lately received increasing attention especially due to its important role in surveillance systems, which should 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 any overlapping or non-overlapping camera.

ReID can use spatio-temporal cues (e.g., [7, 10, 14, 15, 20, 26]) and appearance based cues, on which we focus here. There are three major approaches to appearance based ReID: (1) searching for features invariant to changes in illumination, resolution, pose, and background, while using some fixed distance measure for making a *same* or *not-same* decision (e.g., [3, 4, 8, 9, 13, 18, 21]); (2) designing metrics that aim to bring feature vectors associated with shared identities close to one another and those associated with different identities far from one another (e.g., [11, 16, 22, 23, 25, 28]); and (3) trying to learn the transformation that the appearance of a person in one domain undergoes when passing to another domain (e.g. [2, 19]).

As opposed to the first two approaches, which are *camera-invariant* and deal with re-identifying a person at any new location, the third approach is *camera-specific*. It focuses on

the natural setup of surveillance systems in which the cameras are stationary, and exploits the fact that for each pair of cameras, the transfer domain is limited. It was shown in [2] that this approach can yield better ReID performance.

The main downside inherent in the camera-specific ReID approach is that distinct inter-camera training sets must be collected for each camera-pair. That is, for each camera-pair, videos consisting of the same people passing in front of both cameras must be collected and annotated. Thus, given a site with N cameras $C_1, C_2, ..., C_N$, $N(N-1)/2$ inter-camera transformations must be learned using $N(N-1)/2 = O(N^2)$ distinct inter-camera training sets. This requirement may be impractical.

In this paper, we aim at reducing the number of required direct inter-camera training sets from $O(N^2)$ to $O(N)$. We present a transitive algorithm which uses inter-camera training sets only for $N-1$ camera-pairs (C_i, C_{i+1}), $i = 1, ..., N-1$, and refer to these pairs as *directly trainable pairs*; see Fig. 1(a). The transitive method presented here suggests a way to infer a ReID classifier for any camera-pair in the system, given only these limited training sets. The proposed algorithm enables the inference of a classifier for a *non-directly trainable pair* (C_i, C_{i+2}) given only the available training sets associated with the camera pairs (C_i, C_{i+1}) and (C_{i+1}, C_{i+2}); see Fig. 1(b). The inter-camera classifiers for any other camera pair can be deduced by recursively applying the transitive algorithm; see Fig. 1(c). Here we focus on a triplet case and notate: $[C_i\ C_{i+1}\ C_{i+2}] = [A\ B\ C]$; see Fig. 2.

A naive solution would be to use the union of the available (A,B) and (B,C) training sets for learning the (A,C) ReID classifier, assuming that the inter-camera variability information embedded in them reflects some of the true (A,C) variability. In a sense, this approach is camera-invariant. It works, but its performance is still inferior to that of a directly trained classifier; see Fig. 3. Our goal is to find a more effective use of the available training sets that will decrease this performance gap.

The proposed Transitive Re-IDentification algorithm (TRID) establishes a path between the non-directly trainable camera pair (A,C) by marginalization over the domain of possible appearances in camera B. Camera B plays the role of the 'connecting element' between cameras A and C. This approach indeed minimizes the performance gap while effectively narrowing the number of required training sets.

2 The Transitive Re-identification Algorithm (TRID)

Given two training sets S_{AB}, S_{BC} associated with camera-pairs (A,B) and (B,C), respectively, TRID infers a classifier for the (A,C) camera-pair, for which S_{AC} is missing (Fig. 2).

After some notations, Sec. 2.2 derives the transitive inference, Sec. 2.3 presents the TRID algorithm, and Sec. 2.4 provides further analysis of its discriminative ability.

2.1 Notations

We consider the description of a person's appearance as a d dimensional random variable. It clearly depends both on the person's identity and the camera. The notation x_A refers to a feature vector describing the appearance observed by camera A. $X_A = x_A$ implies that the random variable X_A associated with the appearance in camera A gets the value x_A. For short notation, we usually denote this event simply by "x_A". We also use the binary variable Y_{AB}, which gets the value 1 if and only if the appearances given in A and in B are of the same identity. When the feature vector is known to correspond to a particular individual of identity

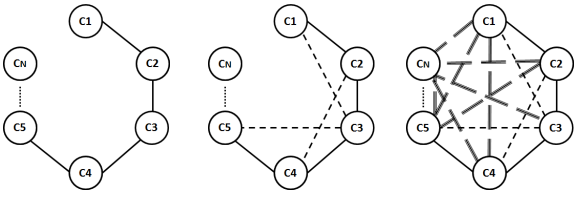


Figure 1: Illustration of a system consisting of N cameras $C_1, C_2, ..., C_N$: (a) Inter-camera training sets are available only for $N-1$ *directly trainable pairs* (C_i, C_{i+1}), $i = 1, ..., N-1$. (b) The transitive algorithm infers classifiers for the *non-directly trainable pairs* (C_i, C_{i+2}). (c) Inter-camera classifiers for any other camera pair in the system can be deduced by recursively applying the transitive algorithm.

Figure 2: The transitive ReID setup. Given the training sets S_{AB} ((a)left) capturing correspondences between people $1...n$ and a training set S_{BC} ((a)right) capturing correspondences between different people $n+1, ..., m+n$, we would like to be able to classify (A, C) pairs (b) without an S_{AC} training set.

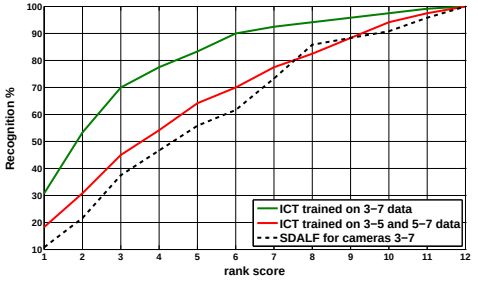


Figure 3: Motivation: A typical case of ReID with 3 cameras: $[A\ B\ C]$=[3 5 7] from the SAIVT-SoftBio [5] database. We present CMC curves (for more details about CMC, see Sec. 3.2.2) comparing the performance of 3 approaches for ReID for cameras (A, C). The *camera specific* ICT algorithm [2] trained with annotated data from cameras A and C outperforms the *camera invariant* SDALF [4] method. A noticeable performance gap exists between ICT trained with (A, C) data and ICT trained using the union of (A, B) and (B, C) training sets. The TRID algorithm aims to minimize this gap by exploiting the (A, B) and (B, C) training sets more wisely, for cases where a direct training set (A, C) is not available.

i, we denote it by x_A^i. Thus, the pair $\{(x_A^i, x_B^i)\}$ is a pair of feature vectors corresponding to the same person but to different cameras.

2.2 The Transitive Inference

Our goal is to estimate the conditional probability $P(Y_{AC}|x_A, x_C)$. Although a training set consisting of annotated pairs $\{(x_A^i, x_C^i)\}$ is not available, we can exploit the annotated sets $S_{AB} = \{(x_A^i, x_B^i)\}$, $i = 1,...,n$ and $S_{BC} = \{(x_B^j, x_C^j)\}$, $j = n+1,...,n+m$. We use different indexes to emphasize that the two training sets possibly correspond to disjoint sets of people.

To obtain $P(Y_{AC}|x_A, x_C)$ we marginalize over all values of Y_{AB},Y_{BC} and X_B:

$$P(Y_{AC}|x_A, x_C) = \sum_{y_{AB}\in\{0,1\}} \sum_{y_{BC}\in\{0,1\}} \left[\int_{x_B \in \mathbb{R}^d} P(Y_{AC}, Y_{AB} = y_{AB}, Y_{BC} = y_{BC}, x_B | x_A, x_C) dx_B \right] , \tag{1}$$

establishing a transitive path between the non-directly trainable camera-pair (A, C)[1]. Applying the chain rule,

$$P(Y_{AC}|x_A, x_C) = \sum_{y_{AB}} \sum_{y_{BC}} \left[\int_{x_B} P(Y_{AC}, Y_{AB} = y_{AB}, Y_{BC} = y_{BC} | x_A, x_B, x_C) f_{X_B}(x_B) dx_B \right] , \tag{2}$$

where $f_{X_B}(x_B)$ is a multi-dimensional probability distribution function that represents all possible appearance vectors in camera B, and which is independent of the particular appearance vectors observed in cameras A and C. Explicitly writing the four additive terms,

$$\begin{aligned} &P(Y_{AC}|x_A, x_C) = \\ &\int_{x_B} P(Y_{AC}, \overline{Y_{AB}}, \overline{Y_{BC}} | x_A, x_B, x_C) f_{X_B}(x_B) dx_B + \int_{x_B} P(Y_{AC}, \overline{Y_{AB}}, Y_{BC} | x_A, x_B, x_C) f_{X_B}(x_B) dx_B \\ &+ \int_{x_B} P(Y_{AC}, Y_{AB}, \overline{Y_{BC}} | x_A, x_B, x_C) f_{X_B}(x_B) dx_B + \int_{x_B} P(Y_{AC}, Y_{AB}, Y_{BC} | x_A, x_B, x_C) f_{X_B}(x_B) dx_B , \end{aligned} \tag{3}$$

we can see that two of them can be eliminated: for a (X_A, X_B, X_C) triplet for which $Y_{AB} = 0$ and $Y_{BC} = 1$, Y_{AC} cannot be 1. The same applies for the case where $Y_{AB} = 1$ and $Y_{BC} = 0$. We apply the chain rule again for the first and fourth terms left in eq. (3) and get

$$\begin{aligned} &P(Y_{AC}|x_A, x_C) = \\ &\int_{x_B} P(\overline{Y_{AB}}|x_A, x_B, x_C) P(\overline{Y_{BC}}|x_A, x_B, x_C, \overline{Y_{AB}}) P(Y_{AC}|x_A, x_B, x_C, \overline{Y_{AB}}, \overline{Y_{BC}}) f_{X_B}(x_B) dx_B \\ &+ \int_{x_B} P(Y_{AB}|x_A, x_B, x_C) P(Y_{BC}|x_A, x_B, x_C, Y_{AB}) P(Y_{AC}|x_A, x_B, x_C, Y_{AB}, Y_{BC}) f_{X_B}(x_B) dx_B . \end{aligned} \tag{4}$$

Y_{AB} is independent of X_C. Y_{BC} is independent of both X_A and of Y_{AB} (given X_B). Y_{AC} is independent of X_B when $Y_{AB} = Y_{BC} = 0$ and is 1 when $Y_{AB} = Y_{BC} = 1$. Therefore,

$$\begin{aligned} P(Y_{AC}|x_A, x_C) &= P(Y_{AC}|x_A, x_C) \int_{x_B} P(\overline{Y_{AB}}|x_A, x_B) P(\overline{Y_{BC}}|x_B, x_C) f_{X_B}(x_B) dx_B \\ &+ \int_{x_B} P(Y_{AB}|x_A, x_B) P(Y_{BC}|x_B, x_C) f_{X_B}(x_B) dx_B . \end{aligned} \tag{5}$$

By a simple reorganization, the desired conditional probability is expressed by

$$P(Y_{AC}|x_A, x_C) = \frac{\int_{x_B} P(Y_{AB}|x_A, x_B) P(Y_{BC}|x_B, x_C) f_{X_B}(x_B) dx_B}{1 - \int_{x_B} P(\overline{Y_{AB}}|x_A, x_B) P(\overline{Y_{BC}}|x_B, x_C) f_{X_B}(x_B) dx_B} . \tag{6}$$

[1] Readers interested only in the final result may go directly to eq. (6).

2.3 The Algorithm

2.3.1 Obtaining the Probability for Directly-Trainable Camera-Pairs

According to eq. (6), in order to estimate the probability $P(Y_{AC}|x_A,x_C)$ for a match associated with cameras A and C, one must provide the probabilities for matches $P(Y_{AB}|x_A,x_B)$ and $P(Y_{BC}|x_B,x_C)$ associated with camera-pairs (A,B) and (B,C), respectively. A few algorithms were suggested for obtaining *same vs. not-same* classifiers for ReID (e.g., [2, 16, 25, 28]). Such classifiers, when provided with an inter-camera training-set, output either a binary decision or a continuous score. The scores, or decision values, are usually used for the ranking of different candidate appearance pairs. Any of these methods can be exploited for our need, provided that it can be modified to output a probability.

We chose the ICT algorithm [2], which has shown state-of-the-art results for modeling the transfer of appearances associated with two specific-cameras. Given a training set S_{AB} associated with a specific camera pair (A,B), the ICT algorithm trains a classifier on the concatenation of appearances from the two cameras: Let $[a|b]$ denote the concatenation of two vectors a and b. The ICT classifier is trained using *positive examples* $[x_A^i|x_B^i]$ for which the identity associated with x_A^i and x_B^i is the same, and *negative examples* $[x_A^i|x_B^j]$ for which the identities associated with x_A^i and x_B^j are different. In our implementation we used color based descriptors (though, in principle, any other descriptors could be used). ICT trains an SVM using an RBF kernel. Given a candidate pair (x_A,x_B), the concatenation $[x_A|x_B]$ is fed to the classifier and the decision value is obtained. In TRID, the decision values are converted to probability estimates using a sigmoid according to Platt's widely used method [24].

2.3.2 Integrating Over the 'Connecting Elements'

The multi-dimensional probability distribution function, $f_{X_B}(x_B)$, can be estimated by different methods for density estimation using the x_B vectors available in the S_{AB} and S_{BC} training sets, and also other, non-annotated, instances associated with people passing in front of camera B. However, estimating high-dimensional density is hard and the integration is computationally costly. Thus, the proposed TRID algorithm approximates the integral by a sum over all available x_B vectors. Let S_B be the set of all such x_B vectors,

$$P(Y_{AC}|x_A,x_C) \approx \frac{\frac{1}{|S_B|}\sum_{x_B\in S_B} P(Y_{AB}|x_A,x_B)P(Y_{BC}|x_B,x_C)}{1-\frac{1}{|S_B|}\sum_{x_B\in S_B} P(\overline{Y_{AB}}|x_A,x_B)P(\overline{Y_{BC}}|x_B,x_C)} \quad . \tag{7}$$

In the process of optimizing the transitive ReID, the SVM parameters were determined such that the probability estimates and the implied integrand are wide and smooth functions of the integration variable x_B. This smoothness and the large number of available (non-annotated) examples makes the sum a good approximation of the integral.

2.4 Asymptotic Analysis

To evaluate the discriminative ability of TRID, we propose an asymptotic simplified analysis of eq. (7). Assume that the x_B's are $K = |S_B|$ randomly drawn examples and the response of the probabilistic classifier $P(Y_{AB}|x_A,x_B)$ is a binary random variable with mean p. When the appearances observed in cameras A and C belong to different people, the variables Y_{AB} and Y_{BC} are independent and $P(Y_{AB}|x_A,x_B)P(Y_{BC}|x_B,x_C) = p^2$. For large K, the numerator of eq. (7) converges to its expected value p^2, while its variance converges to $\frac{1}{K}p^2(1-p^2)$. When the observed appearances are of the same identity, the responses are positively correlated and

$P(Y_{AB}|x_A,x_B)P(Y_{BC}|x_B,x_C) = pp'$, where $p' \in (p,1]$ reflects the amount of dependency. In this case, for large K, the numerator converges to pp' and the variance to $\frac{1}{K}pp'(1-pp')$. The s.d. decrease with K. Therefore, for increasing K, the value of the numerator for the same identities case is larger than the numerator for the different identities case with arbitrary large probability. As for the denominator, a similar analysis implies that it converges to $2p-p^2$ for the different identities case and to $2p-pp'$ (which is smaller) for the same identities case. Thus, dividing by the denominator further emphasizes the distinction between the values corresponding to same and different identities.

3 Experiments

First, in Sec. 3.1 we use simple low-dimensional synthetic data to demonstrate how TRID achieves the transitive inference and to enhance the reader's intuitive understanding of our method. Then, in Sec. 3.2 we describe experiments on a multi-camera setup. We describe the dataset used, provide some implementation details, and compare the performance of TRID to that of other non-transitive algorithms [2].

3.1 Synthetic Demonstration

This section describes a simple experiment using synthetic data, where x_A, x_B, x_C are all one dimensional. This enables us to illustrate the situation upon the joint features spaces associated with two cameras.

In the first experiment we assume that the data is generated as follows: we randomly sample from a uniform distribution, each sample thus representing an "individual". These values are the 'clean' x_A values. The 'clean' x_B and x_C values are created by simple linear functions: $x_B = a_1 x_A + b_1, x_C = a_2 x_B + b_2$. The x_A, x_B, x_C values available to the learning procedure are a noisy version of these clean values. Clearly the relation between feature points x_C and their corresponding points x_A are also given by a linear function which is a composition of the two functions above. See the locations of the feature pairs in Fig. 4left(a,b,c). Disjoint sets of samples are used for training and test. The function $P(Y_{AC}|x_A,x_C)$ learned by the transitive TRID algorithm is plotted in Fig. 4left(g). Note that this plot is similar both to the plot of the true feature pairs (Fig. 4left(c)), and to the results of direct learning from (A,C) examples (Fig. 4left(f)). Also note that the naive algorithm fails to learn the true transformation (Fig. 4left(e)). A CMC curve in Fig. 4left(d) shows the ReID results. We see that TRID, which uses indirect inference, performs similarly to ICT, which uses a direct training set.

The second experiment is similar except that the transformations from one camera to the other are not linear and not even single valued. Rather they are implicitly defined by a random selection of one of two linear functions; see Fig. 4right(a,b,c). Though a general deterioration of performance is observed for this more challenging case, TRID still closely follows the performance of the directly trained ICT (Fig. 4right(d)).

[2] The Matlab source code of TRID as well as of the ICT are available at: http://www.cs.technion.ac.il/~tammya/Reidentification.html .

3.2 Experimenting with Multi-Camera Setup

3.2.1 SAIVT-SoftBio Database

To test the TRID algorithm, we needed an annotated dataset associated with a site with at least 3 stationary cameras. Unfortunately most common ReID benchmark datasets are unsuitable. They are either limited to images taken by only two cameras, (VIPeR [17], CAVIAR4REID [8]), annotated without including camera identities (iLIDs MCTS [1, 27]), or contain images taken from a moving platform (ETHZ [12]).

We used the recently presented multi-camera surveillance database SAIVT-SoftBio [5]. It includes annotated sequences (704×576 pixels, 25 frames per second) of 150 people, each captured by a subset of eight different indoor cameras, providing various viewing angles and varying illumination conditions; see Fig. 5. A coarse bounding box indicating the location of the annotated person in each frame is provided.

3.2.2 Implementation and Experimentation Details

To implement the TRID algorithm, we had to train two ICT ReID classifiers. For each ICT, we used a training set corresponding to two cameras and a set of M_{Train} people. 10 frames per person and camera were sampled. Following the descriptors recommended in [2], each bounding box was divided into five horizontal stripes and each stripe was described by a histogram with 10 bins for each of the color components H, S, and V. x_A, x_B and x_C are therefore feature vectors with $d = 150$ dimensions. For describing a pair, the corresponding feature vectors were concatenated. We built all $100M_{Train}$ positive examples (same person, different cameras), and randomly selected $30M_{Train}(M_{Train}-1)$ negative examples (different people and cameras). The C and γ parameters for the RBF SVM were taken from the code supplied by the authors of the ICT algorithm. We found that overall optimization of the transitive ReID performance yielded similar values. All experiments were carried out with the same parameters. We modified ICT to output the posterior conditional probabilities using LibSVM [6] implementation and deduced $P(Y_{AB}|x_A,x_B)$ and $P(Y_{BC}|x_B,x_C)$.

We used eq. (7) to evaluate the matching probability, where the summation is performed over all (M_B) x_B appearance descriptors available from camera B except those corresponding to the test set. Here we performed a multi-shot version, where 10 images of the same person were taken from the camera, and all 100 pairs were tested using eq. (7), followed by averaging their estimated matching probabilities.

The main tool for evaluating the results is the Cumulative Match Characteristic (CMC) curve, widely accepted for evaluation of ReID algorithms. For each instance in the test set, each algorithm ranked the matching of the appearance in camera A with the appearances of all instances in the test set in camera C. The CMC curve summarizes the statistics of the ranks of the true matches.

We performed four transitive ReID experiments. In each experiment a different camera-triplet played the role of cameras A, B and C. The four selected triplets were those that included the largest number of commonly annotated subjects. We repeated the following procedure 10 times: The annotated set was randomly divided into three subsets corresponding to disjoint identities. One, containing M_{Test} subjects, was used for testing, and the two other equal subsets, S_{AB} and S_{BC} of size M_{Train}, were used for training the two ICT direct ReID classifiers. The particular camera triplets and the corresponding M_{Train}, M_{Test}, M_B values are: ([1 5 7], 40, 21, 102), ([3 5 7], 30, 12, 111), ([1 5 3], 30, 14, 109), ([1 3 8], 30, 15, 99).

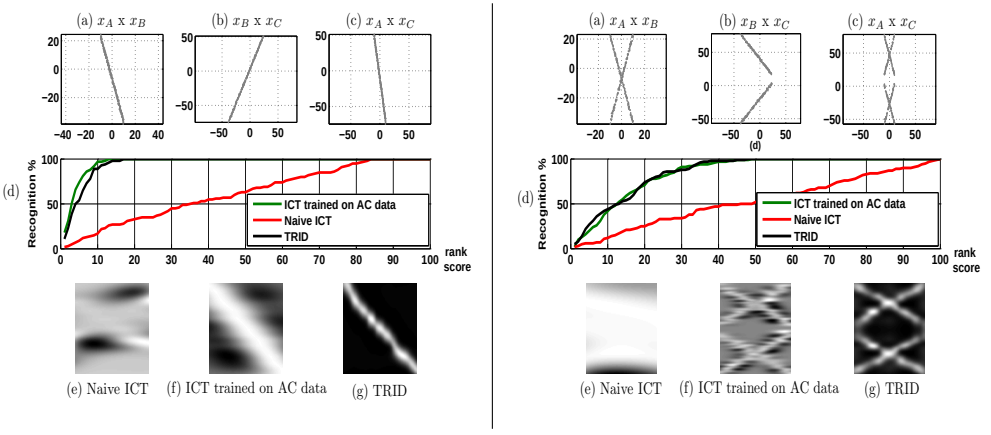


Figure 4: Synthetic experiments for feature space visualization. See text (Sec. 3.1).

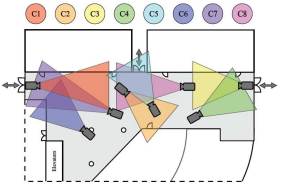
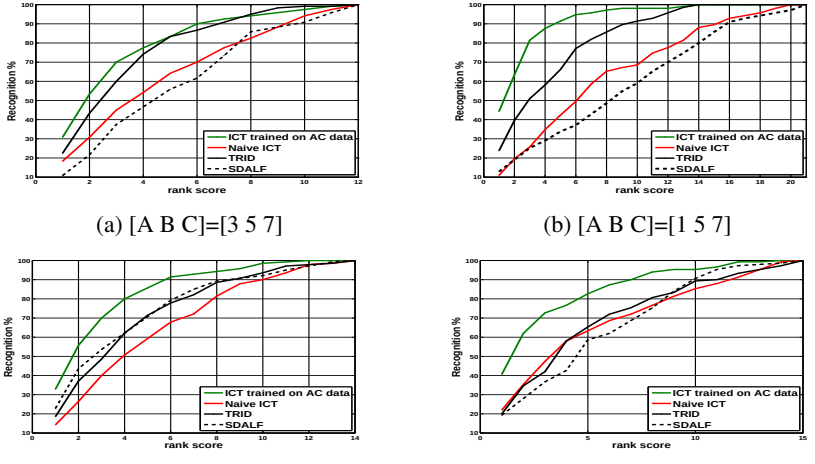


Figure 5: Approximate camera placement and orientation in the Multi-Camera SAIVT-SoftBio Surveillance Database. This image was taken from [5].



(a) [A B C]=[3 5 7]

(b) [A B C]=[1 5 7]

(c) [A B C]=[1 5 3]

(d) [A B C]=[1 3 8]

Figure 6: CMC curves comparing the performance of TRID to that of other, non-transitive algorithms, over 4 combinations of camera-triplets from the SAIVT-SoftBio dataset.

3.2.3 Results and Comparisons

As mentioned above, and as demonstrated in Fig. 3, the best classifier can be obtained when annotated data, S_{AC}, specific for the (A,C) camera-pair, is available for training. The motivation behind the development of the TRID algorithm was to get close as possible to that performance when this specific data is not available, by transitively using the non-direct training sets S_{AB} and S_{BC}. Therefore, we compared TRID results to results obtained by the following algorithms:

- **ICT trained on AC data:** The ICT algorithm was trained using a direct annotated training set S_{AC} of size N_{train}.
- **Naive-ICT:** The ICT algorithm was trained by the union of S_{AB} and S_{BC}.
- **SDALF:** The SDALF camera-invariant approach (that does not have a training phase). We used the original SDALF multiple-shot code[3]. As advised by the SDALF authors, we used a background subtraction pre-processing step to obtain the people's silhouettes, which included the subtraction of a background image (supplied with the dataset for each camera), followed by low pass filtering and thresholding[4].

Fig. 6 presents the results for the 4 combinations of camera-triplets. The four camera triplets represent a variety of possible inter-camera relations in a multi-camera system (Fig. 5). The most significant performance gap reduction occurred for $[A\ B\ C] = [3\ 5\ 7]$ (Fig. 6(a)), where TRID performed almost as well as the reference ICT trained on AC data. Here, cameras A and C were located far away from each other and associated with significantly different appearances due to different illumination conditions and background. Camera B was located just in the middle, and seems to provide a good transitive path. It seems that when either the A,B or the B,C views are similar (Fig. 6(d)), the advantage of the TRID over the naive approach is smaller. Overall, the results show that TRID performed better than the camera invariant method (SDALF), with notably improved performance over the naive ICT method.

4 Discussion

ReID benefits from training with corresponding appearance-pairs captured by specific cameras pairs, as the background, illumination, resolution and pose are camera dependent (Fig. 2). Nevertheless, the practical difficulty in collecting such data sets is a limitation. In this paper we proposed a new approach for reducing this limitation: collecting data only from a small subset of the camera pairs, and using transitivity to improve ReID performance for all camera pairs. The transitive inference ReID algorithm, denoted TRID, is based on statistical modeling. We demonstrate its principles on simple low dimensional data and then show that it indeed improves ReID performance on camera pairs for which annotated pairs are not available. Moreover, the accuracy of the TRID is superior to that of a state-of-the-art camera invariant, which used a fixed similarity measure. It is also more accurate than a method that makes more simplistic use of training sets associated with other camera-pairs. The TRID algorithm is in fact a general framework that may be combined with different probabilistic classifiers (not necessarily ICT) and of course different features.

[3] The original SDALF code is available at: http://www.lorisbazzani.info/code-datasets/sdalf-descriptor .

[4] We have also tried using binary masks supplied with the dataset, and got similar results.

A common requirement in camera systems is to add a new camera to a set of N previously installed cameras. Camera specific ReID requires in this case the data be collected from N camera pairs. With the proposed transitive approach, collecting data for only one pair is required, which is clearly more practical.

To the best of our knowledge, this work is not only the first approach to transitivity in ReID but also, more generally, transitivity in domain adaptation. In our future work we intend to test the TRID algorithm on other domain adaptation tasks, and to examine alternative transitive approaches. One direction for the latter would be to represent the direct ReID classifiers using two explicit transformations and to combine them by function composition. While the transitive composition is straightforward in this case, representing the transformation by single valued function seems inadequate due to the non-unique appearance of the same object in the same camera.

Transitive use of indirect training sets, as proposed here, can be applied to strengthen direct learning, when the set of direct annotated pairs set is small (but not empty). Many interesting topics arise when trying to extend the transitive approach from camera-triplets to larger camera sets. We intend to study the recursive TRID application for obtaining ReID estimates for all camera pairs, to study this estimation with multiple paths, to study the optimal topology, and to analyze the cumulative error.

5 Acknowledgments

This research was supported by the MAGNET program in the Israeli ministry of industry and commerce, by the Israeli ministry of science and by the E. and J. Bishop research fund. We would like to thank the SAIVT Research Labs at Queensland University of Technology (QUT) for freely supplying us with the SAIVT-SoftBio database.

References

[1] i-lids multiple camera tracking scenario definition. *UK Home Office*, 2008.

[2] T. Avraham, I. Gurvich, M. Lindenbaum, and S. Markovitch. Learning implicit transfer for person re-identification. In *The 1st International Workshop on Re-Identification (Re-Id 2012), in conjunction with ECCV, LNCS*, pages 381–390, 2012.

[3] L. Bazzani, M. Cristani, A. Perina, M. Farenzena, and V. Murino. Multiple-shot person re-identification by HPE signature. In *International Conference on Pattern Recognition (ICPR)*, pages 1413–1416, 2010.

[4] L. Bazzani, M. Cristani, and V. Murino. Symmetry-driven accumulation of local features for human characterization and re-identification. *Computer Vision and Image Understanding*, 117:130–144, 2013.

[5] A. Bialkowski, S. Denman, P. Lucey, S. Sridharan, and C. B. Fookes. A database for person re-identification in multi-camera surveillance networks. In *Digital Image Computing: Techniques and Applications (DICTA 2012)*, pages 1–8, 2012.

[6] C. Chang and C. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/$\sim$cjlin/libsvm.

[7] K. W. Chen, C. C. Lai, P.J. Lee, C. S. Chen, and Y. P. Hung. Adaptive learning for target tracking and true linking discovering across multiple non-overlapping cameras. *IEEE Transactions on Multimedia*, 13:625–638, 2011.

[8] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino. Custom pictorial structures for re-identification. In *British Machine Vision Conference (BMVC)*, pages 68.1–68.11, 2011.

[9] E. D. Cheng and M. Piccardi. Matching of objects moving across disjoint cameras. In *International Conference on Image Processing (ICIP)*, pages 1769–1772, 2006.

[10] A. Dick and M. Brooks. A stochastic approach to tracking objects across multiple cameras. In *Australian Conference on Artificial Intelligence*, pages 160–170, 2004.

[11] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja. Pedestrian recognition with a learned metric. In *Asian Conference on Computer Vision (ACCV)*, pages 501–512, 2010.

[12] A. Ess, B. Leibe, and L. Van Gool. Depth and appearance for mobile scene analysis. In *International Conference on Computer Vision (ICCV)*, 2007.

[13] N. Gheissari, T.B. Sebastian, and R. Hartley. Person reidentification using spatiotemporal appearance. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1528–1535, 2006.

[14] A. Gilbert and R. Bowden. Tracking objects across cameras by incrementally learning inter-camera colour calibration and patterns of activity. In *European Conference on Computer Vision (ECCV)*, pages 125–136, 2006.

[15] A. Gilbert and R. Bowden. Incremental, scalable tracking of objects inter camera. *Computer Vision and Image Understanding*, 111:43–58, 2008.

[16] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *European Conference on Computer Vision (ECCV)*, pages 262–275, 2008.

[17] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *PETS Workshop in conjunction with ICCV*, 2007.

[18] W. Hu, M. Hu, X. Zhou, T. Tan, and J. Lou. Principal axis-based correspondence between multiple cameras for people tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:663–671, 2006.

[19] O. Javed, K. Shafique, and M. Shah. Appearance modeling for tracking in multiple non-overlapping cameras. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26–33, 2005.

[20] P. KaewTrakulPong and R. Bowden. A real-time adaptive visual surveillance system for tracking low resolution colour targets in dynamically changing scenes. *Journal of Image and Vision Computing*, 21:913–929, 2003.

[21] I. Kviatkovsky, A. Adam, and E. Rivlin. Color invariants for person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 99, 2012.

[22] C. Liu, S. Gong, C. C. Loy, and X. Lin. Person re-identification: What features are important. In *The 1st International Workshop on Re-Identification (Re-Id 2012), in conjunction with ECCV, LNCS*, volume 7583, pages 391–401, 2012.

[23] B. Ma, Y. Su, and F. Jurie. Local descriptors encoded by Fisher vectors for person re-identification. In *The 1st International Workshop on Re-Identification (Re-Id 2012), in conjunction with ECCV, LNCS*, volume 7583, pages 413–422, 2012.

[24] John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74, 1999.

[25] B. Prosser, W. Zheng, S. Gong, and T. Xiang. Person re-identification by support vector ranking. In *British Machine Vision Conference (BMVC)*, pages 21.1–21.11, 2010.

[26] C. Stauffer. Learning to track objects through unobserved regions. *IEEE Workshop Motion and Video Computing*, pages 96–102, 2005.

[27] W. Zheng, S. Gong, and T. Xiang. Associating groups of people. pages 23.1–23.11, 2009.

[28] W. Zheng, S. Gong, and T. Xiang. Person re-identification by probabilistic relative distance comparison. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 649–656, 2011.