Domain Transfer for Person Re-identification

Ryan Layne, Timothy M. Hospedales, Shaogang Gong
Queen Mary University of London
London, England
{rlayne, tmh, sgg}@eecs.qmul.ac.uk

ABSTRACT

Automatic person re-identification in is a crucial capability underpinning many applications in public space video surveillance. It is challenging due to intra-class variation in person appearance when observed in different views, together with limited inter-class variability. Various recent approaches have made great progress in re-identification performance using discriminative learning techniques. However, these approaches are fundamentally limited by the requirement of extensive annotated training data for every pair of views. For practical re-identification, this is an unreasonable assumption, as annotating extensive volumes of data for every pair of cameras to be re-identified may be impossible or prohibitively expensive.

In this paper we move toward relaxing this strong assumption by investigating flexible multi-source transfer of re-identification models across camera pairs. Specifically, we show how to leverage prior re-identification models learned for a set of source view pairs (domains), and flexibly combine these to obtain good re-identification performance in a target view pair (domain) with greatly reduced training data requirements in the target domain.

Categories and Subject Descriptors

I.5.4 [**Pattern Recognition**]: Applications; I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis

General Terms

Surveillance, Transfer Learning

Keywords

Person Re-identification, Support Vector Machines

1. INTRODUCTION

Person re-identification, or inter-camera entity association, is the task of recognizing an individual across heterogeneous non-overlapping camera views against a background of similar persons. When an individual disappears from one view they need be differentiated from numerous possible alternative people and re-identified in another view, potentially under a different and unknown view angle, pose, lighting conditions and clutter or occlusion (see Figure 1 for examples). This is critical to a variety of safety, security and efficiency tasks which require long-term maintenance of consistent identity across space and time. In particular, it is a fundamental capability for long-term tracking across multiple disjoint camera views [11].

Relying on manual re-identification in large camera networks is prohibitively costly and error prone. For these reasons, there has recently been extensive work in the computer vision community on automated re-identification [7, 24, 10] . This is very challenging because of extreme intra-class (person identity) variability in appearance across views with different lighting, pose and occlusion; and limited inter-class variability in appearance among many similarly clothed pedestrians. Existing approaches can be broadly broken down into two complementary categories: those which focus on developing effective feature representations [7, 5], and those which focus on developing learning methods to better discriminate identity using a given representation [21, 24, 10].

Feature design approaches [7, 5] suffer from the problem that it is extremely challenging if not impossible to design features that are discriminative enough to distinguish people reliably; while simultaneously being invariant to all the covariates which occur in practice such as, motion blur, view angle and pose change, lighting and occlusion. In contrast, learning approaches [21, 24, 10] try to improve on a given set of features, and focus on discriminative training of models to maximize re-identification performance, for example distance metric learning [24, 10, 4] and support vector machines (SVM) [21, 1]. Recently, discriminative approaches have significantly improved state of the art benchmark performance [24, 10, 1] treating re-identification as a binary (same versus different person) rather than multi-class (person identity) problem.

A central limitation of existing discriminative learning approaches, is that they are more suited to closed-world benchmark problems than realistic open-world scenarios. In particular they require many pairs of person images annotated by same/different, *for each camera pair* between which the system is required to operate. This is reasonable for training/testing splits on benchmark datasets that are already exhaustively annotated by person identity. However it is highly impractical for real-world use, where there may be very many pairs of cameras in a given network, *each* requiring exhaustive annotation – making this "calibration" requirement of such a system impossible or prohibitively expensive. Ideally, we would like to deploy a re-identification system between a pair of cameras with minimal calibration/training annotation. What a system learns from annotations of one camera pair should be exploited by another pair without requiring exhaustive annotation in the new pair.

ARTEMIS'13, October 21, 2013, Barcelona, Spain.

ACM 978-1-4503-2393-2/13/10 ...$15.00.
http://dx.doi.org/10.1145/2510650.2510658.

Figure 1: Examples from all of the datasets we use in our experiments; from the top: VIPeR, PRID, GRID, and CUHK. Note the dramatic appearance variations in both the people and backgrounds; as well as how image quality varies.

This is an issue in *transfer learning* [19, 6, 12]. Transfer learning is already important for many classical vision problems such as object recognition [22] with multiple classes or domains. However it is critically important for re-identification because the number of domains (camera pairs) is *quadratic in the number of cameras*. Therefore obtaining exhaustive training data for each domain is even more impractical than for conventional vision applications, and transfer learning becomes critical. Nevertheless, no prior re-identification studies have addressed this issue, relying solely on benchmark datasets with sufficient annotated data in each camera-pair of interest.

In this paper we relax the practically unrealistic assumption of exhaustive training data within each domain by generalizing recent ideas in learning re-identification [1] and SVM transfer learning [12]. Specifically, we consider re-identification based on binary-relation learning [1, 13], and show how to generalize this approach to to achieve effective cross-domain learning by combining non-linear decision boundaries from source domains to create a more accurate target domain re-id classifier. In this way we are able to improve on within-domain learning both for sparse and even non-sparse training data volumes. Moreover we show how to achieve this while systematically avoiding negative transfer, even when there are multiple and irrelevant source domains.

2. RELATED WORK

2.1 Feature design

Contemporary approaches to re-identification (re-id) typically exploit features such as color, texture, spatial structure, or combinations thereof [3, 5, 21, 7]. Once a representation has been designed, nearest-neighbor [7] may be used for re-identification given a suitable distance metric (e.g., Euclidean) to measure the similarity between two samples. Beyond the intrinsic challenge of designing view invariant but identity variant features, a fundamental problem is that features which are most effective in one domain (re-id view pair) may be less effective in another domain (a new re-id view pair) as we will show in Section 4.5. For this reason among others, learning techniques have been studied that are trained to maximize re-identification performance within a given domain.

2.2 Learning re-identification

Learning approaches to re-identification typically learn distance metrics [10, 24, 4], or model-based matching procedures such as boosting [8] and ranking [21] based on annotated training pairs. These have recently improved state of the art re-id performance significantly [10, 1]. Another line of research learns mid-level attributes [14] to replace or augment low level features. In this case inter-camera invariance is obtained via the generalization performance of learned attribute classifiers. However, this only applies within domains where annotated attribute data is available. The recently proposed binary relation learning approach [1, 13] obtains state of the art re-id results by exploiting strong SVM classifiers trained to make same/different judgements on pairs of images. This strategy does not assume that instances of the same person are more similar than instances of different people, and instead implicitly learns the mapping between appearance in pairs of training cameras. A serious issue with all these approaches is that they do not generalize well across domains (different re-id view pairs; see Section 4.5); and hence require extensive volumes of training data for *each pair* of cameras to be re-identified between. This is possible for benchmark scenarios, but unreasonable in practice.

2.3 Cameras and Domains

In this work we consider a camera *pair* to make up a *domain*, and this should not be confused with some other studies which consider a particular *camera* to be a domain [22]. For classification [22] and detection [6], an individual camera encompasses the notion of a domain because a camera's parameters impart a systematic impact on the observations, which the model must learn to interpret. However in re-identification, a model's task is to infer something about pairs of observations, and the systematic impact of the environment is therefore defined by the pair of cameras.

2.4 Transfer Learning

Transfer learning [19] has been used to good effect in numerous classical computer vision problems, for example object categorization [12, 22]. The motivation is typically to scale systems to many classes [12] or domains [22, 6] without requiring prohibitive amounts of training data. While transfer learning is already an important issue in classical vision tasks, it will turn out to be even more central to the re-identification problem. This is because since *pairs* define domains in this context, it is unreasonable to collect exhaustive training data for a quadratic number of domains.

Only very recently has transfer learning for re-identification begun to be considered [15, 25]. However these studies consider only improving within-domain (camera pair) re-identification by transferring knowledge learned from one group of people to help identify another group of people. This is intrinsically a much more restricted scenario than the more general and useful case of transferring across domains to permit re-identification in a new camera pair with sparse annotations.

A central issue in transfer learning [19] is that of *from where to transfer*. When there is only one source of information available,

and that source is known to be highly relevant to the task of interest, then transfer learning is much simpler than in the more general and realistic case where there are multiple sources of information of greatly varying relevance. In this latter case, it is non-trivial to design models which avoid negative transfer [19]. Our problem of transferring mappings across camera pairs falls squarely into the latter more difficult case. Since the relevance of one camera pair to another depends on similarity in their viewing angles and lighting, many pairs will not be similar and working out from where to transfer is of critical importance.

Our Approach.

We address all the mentioned issues by generalizing the state of the art binary relation approach to re-id [1], but tackle the new challenges in addressing the training data requirements via multi-source transfer. There are many potential approaches to transfer learning [19], but in this study we will develop a SVM multi-kernel learning (MKL) [17, 12] transfer strategy. This will allow us to integrate multiple source domains of unknown relevance, while avoiding negative transfer via an inter-kernel sparsity regularizer

We make the following specific contributions: (i) Framing the problem of generalizing re-identification as a domain-transfer problem; (ii) Developing a specific framework for domain-transfer re-id for multiple domains of varying relevance by way of expressing the task as a SVM multi-kernel learning problem; (iii) Revealing the limitations of existing approaches to re-identification by way of a systematic and quantitative cross-domain evaluation; and (iv) Extensive evaluation of our proposed method on four of the largest public re-id datasets available.

3. METHODS

3.1 Concept Illustration

To provide intuition before introducing the details of the proposed method, Figure 2 provides an schematic illustration of our re-id transfer learning framework. In this illustration, the feature space within each camera is one dimensional. A domain, consisting of pairs of observations made by two cameras, can thus be represented as a point on a two dimensional plane. Pairs of cross-view images corresponding to the same person are shown with circles, and pairs corresponding to different people with crosses. Binary-relation [1] based re-id is the strategy of learning a decision boundary in this space (Figure 2, blue lines). In an easy re-identification scenario, the feature-space is the same in each view, so distinguishing true pairs from false pairs requires only a simple decision boundary (Figure 2(a)). In a realistic scenario, there will be a non-trivial and unknown transformation [20] in feature space from one camera view relative to another (Figure 2(b) and (c)). In this case a strong non-linear classifier could learn the decision boundary separating true from false pairs, and hence an implicit inter-camera mapping.

In this illustration, we assume there are three source domains (camera pairs; Figure 2(a)-(c)) for which annotated data (red and green symbols) is plentiful, and good binary relation based re-id models have been learned (blue lines). Now suppose we wish to deploy our re-identification system to a new location where we can only annotate a very limited amount of training data. With limited data, a re-identification classifier learned in the conventional way – solely from local data – will be much less accurate, clearly misclassifying many regions of the input space (Figure 2(e), unlabeled grey symbols on the wrong side of the decision boundary). In contrast, a re-identification classifier taking advantage of our domain transfer framework will realize that the limited data is best explainable by the model learned from the second source domain (Figure 2(b)),

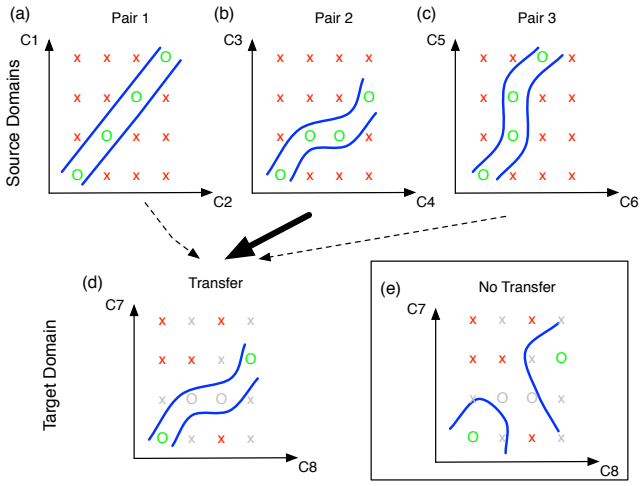

Figure 2: An illustration of how domain transfer can assist re-identification. Symbols indicate same/different pairs, grey symbols are un-annotated and lines indicate decision boundaries.

and borrow that classifier's strength to help learn a much more informative and accurate boundary than is possible using local data alone (Figure 2(d) vs (e)). (The intuition for how this works is finding a source domain classifier or combination thereof which fit the few available data points in the target domain). Finally, note that simple averaging of all the source classifiers is insufficient: in this example the mean of source classifiers (a)-(c) is very similar to classifier (a) which will be wrong for the target domain (d). We shall validate these intuitive observations experimentally in our experiments (Section 4.4).

3.2 Within Domain Re-identification

3.2.1 *Training*

We first consider the case of learning to re-identify people within one particular domain corresponding to a camera pair a and b. Here we largely follow a binary-relation learning approach [1, 13], but review the method for completeness. We assume training data $\{\mathbf{x}_i^a, z_i^a\}_{i=1}^{N_a}$ describing N_A people observed in camera a, and $\{\mathbf{x}_j^a, z_j^b\}_{j=1}^{N_b}$ describing N_B people appearing in camera B, where $\mathbf{x}$ represents a feature vector, and z indicates the identity of each person. From this data we can generate:

- A set of cross-camera positive pairs of the same person: $\{y_k = 1, \mathbf{x}_k = [\mathbf{x}_i^a || \mathbf{x}_j^b]_k\}, \forall (z_i = z_j)$,
- A set of cross-camera negative pairs of different people: $\{y_k = -1, \mathbf{x}_k = [\mathbf{x}_i^a || \mathbf{x}_j^b]_k\}, \forall (z_i \neq z_j)$,

where $[\cdot||\cdot]$ denotes concatenation and $k = 1 \ldots N$ indexes observation pairs $\mathbf{x}_k$. Note that there are a quadratic number of negative pairings, and actually constructing all pairs is typically prohibitive, so using a random subset of negative examples is typically adopted [1, 21].

To learn a re-identification model, we train a classifier on pair data $\{y_k, \mathbf{x}_k\}_{k=1}^{N}$ to distinguish matching pairs from non-matching pairs. This can be formalized as a support vector machine learning

problem as:

$$\min_{\mathbf{w},\xi} \|\mathbf{w}\|^2 + \frac{C}{N}\sum_{k=1}^{N}\xi_k,$$
$$\text{s.t. } y_k \mathbf{w}^T \phi(\mathbf{x}_k) \geq 1 - \xi_k, \ \forall_k, \quad (1)$$

where $\phi(\cdot)$ is a non-linear mapping, and we maximize the margin subject to the soft constraint (slack variable ξ_k) that true pairs should be positive and false pairs should be negative.

Discussion.
Note that this objective (Eq. 1) pursues positive true pairs and negative false pairs, without any assumption of their visual similarity/dissimilarity. With the RBF kernel, binary-relation SVM implicitly learns an arbitrarily complex transformation mapping between cameras, e.g., uncovering their lighting [20] or view transformation, as well as relative relevance for each feature within that domain. In contrast, the common RankSVM [21] approach has two limitations: (i) it only models a first-order weighting of features, without considering their covariance, and (ii) it operates under the explicit assumption that true pairs should be more similar than false pairs (i.e., Figure 2(a)). In practice this means that for camera pairs which deviate sharply from a simple linear transformation model (e.g., Figure 2(a)) to a more complex transformation (e.g., Figure 2(b) or (c)), binary relation SVM outperforms RankSVM, as shown in [1]. Mahalanobis metric learning objectives [10, 4, 24] are more powerful than RankSVM in modelling feature covariance, however they also still assume that true pairs are more similar than false pairs.

3.2.2 Re-identification

For online re-identification of persons across cameras, putative pairs of images are concatenated $\mathbf{x}_* = [\mathbf{x}_*^a, \mathbf{x}_*^b]$ and the score of a test pair $\mathbf{x}_*$ is is evaluated as $f(\mathbf{x}_*) = \mathbf{w}^T \phi(\mathbf{x}_*)$. The pair can be classified as same or different via $\text{sign} f(\mathbf{x}_*)$, or the continuous score itself can be used to relatively rank putative matches. Given this re-identification framework, we next address how to transfer learned models across domains.

3.3 Domain Transfer Re-Identification (DTR)

3.3.1 Training

Assume a set of source domains $s = 1 \ldots S$ are given, for which we have learned re-identification models as per Section 3.2. To leverage the learned experience of these domains in a new target domain t, we take the strategy of multi-kernel learning [2]. Each source domain s can be seen as providing a score $f_s(\mathbf{x})$ indicating its confidence that a given pair $\mathbf{x}$ is a matching pair under the model of that domain. We therefore formalize a domain transfer prediction task, which classifies a pair $\mathbf{x}$ in the target domain, taking into account both target and source domain knowledge, as:

$$\begin{aligned} f_t(\mathbf{x}) &= \overline{\mathbf{w}}^T \overline{\phi}(\mathbf{x}), \\ &= \mathbf{w}_t^T \phi_t(\mathbf{x}) + \sum_{s=1}^{S} \mathbf{w}_s^T \phi_s(f_s(\mathbf{x})), \end{aligned} \quad (2)$$

where parameters $\overline{\mathbf{w}} = [\mathbf{w}_t, \mathbf{w}_s]$ to be determined weight the relative informativeness of the target domain and each source domain knowledge.

Given this task formulation, the within-domain learning objective in Eq (1) can be generalized to the case of domain-transfer

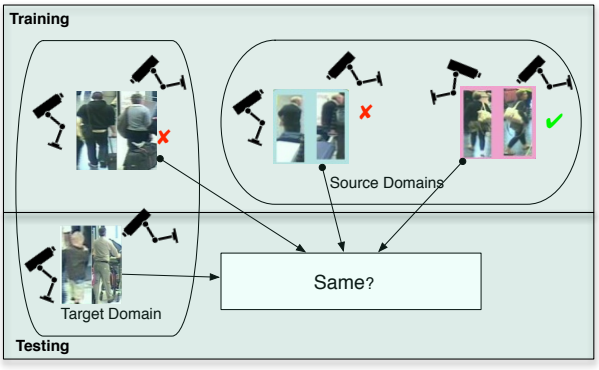

Figure 3: Schematic overview of our framework.

learning to estimate $\overline{\mathbf{w}}$ as:

$$\min_{\overline{\mathbf{w}}} \Omega(\overline{\mathbf{w}}) + \frac{C}{N}\sum_{k=1}^{N} L(\overline{\mathbf{w}}, \mathbf{x}_k, y_k), \quad (3)$$

where L denotes the hinge loss

$$L(\overline{\mathbf{w}}, \mathbf{x}, y) = \left| 1 - y\overline{\mathbf{w}}^T \overline{\phi}(x) \right|_+, \quad (4)$$

and $\Omega(\overline{\mathbf{w}})$ denotes the weight regularizer. Note that [12] use a linear kernel ϕ for computational tractability. In our case, because the problem is binary, we are able to use the RBF kernel instead without great penalty. This is indeed necessary because we need to learn a complex transformation.

Evaluating Domain Relevance.
An important issue for domain transfer in the general unconstrained case is that we do not know in advance which source domain is going to be relevant, and indeed the majority are likely to be irrelevant. For this reason we seek a sparse solution for the optimization problem in Eq (3). Previously l_1 norm regularizers have been proposed to provide sparsity across kernels. However this is hard to optimize effectively [2]. The l_p $(1 < p \leq 2)$ norm regularizer has recently been shown to effectively induce sparsity while providing significantly easier optimization [18]. We therefore take the $(2, p)$ group-norm as the regularizer: providing l_2 regularization within domains, while encouraging l_p sparsity across the set of $S+1$ kernels which reflect the cues from the target domain and the S source domains:

$$\begin{aligned} \Omega(\overline{\mathbf{w}}) &= \frac{1}{2}\|\overline{\mathbf{w}}\|_{2,p}^2, \\ &= \frac{1}{2}\|[\|\mathbf{w}_t\|_2, \|\mathbf{w}_1\|_2, \ldots, \|\mathbf{w}_S\|_2]\|_p^2. \end{aligned} \quad (5)$$

This avoids negative transfer because any source kernels which mismatch the available target domain data will be allocated zero coefficients. Expressed in this form, we can exploit existing efficient stochastic gradient-descent algorithms [17] for solving the cross-domain re-identification learning problem in Eq (3).

4. EXPERIMENTS

4.1 Feature Extraction

The main imagery feature that we will use with our DTR model is the 150 dimensional HSV color descriptor as detailed in [1]. Ad-

ditionally we compared the commonly used ensemble of local features (ELF) which encodes both color and texture in 2784 dimensions as detailed in [8, 21]; as well as symmetry driven accumulation of local features (SDALF) as detailed in [7]. Note that SDALF provides a distance matrix directly, rather than a feature encoding.

4.2 Experimental Settings

We tested the model using the four largest publicly available re-id datasets: VIPER [23], PRID [9], GRID [16] and CUHK [15], which provide 316, 200, 250, and 971 matched pairs respectively. These datasets cover a diverse variety of image sizes (in the region of [128x48] to [128x64].), typical view angles and camera conditions (Figure 1). We evaluated cross-domain re-identification performance on these datasets in four "leave one dataset out" folds. In each case we considered three datasets as source domains and the fourth dataset as the target domain. For the source domains we learned within-domain re-identification models with all available data for each (Section 3.2). For the held out domain, we performed 2-fold cross-validation, training the domain transfer model on half (or less) of the data (Section 3.3), and using the held out half for testing. For testing, we consider the matched pairs between cameras within the domain, taking each person in turn (probe) and matching them against the people in the other camera (gallery). Within the source domains, SVM slack parameter C was cross-validated to optimize expected rank. In the target domain we set $C = 10$ throughout. We fixed the RBF kernel parameter γ to the median of each distance matrix. For the SVM methods we select 10 negative examples per positive pair.

4.3 Evaluation

As baselines we consider where relevant three non-learning methods and three learning methods. For non-learning methods we consider: (i) HSV features [1], (ii) ELF [8] and (iii) SDALF [7]; in each case with nearest neighbor (NN) matching and Euclidean distance where relevant. For learning methods, we consider:

ATTR: Re-identification using Euclidean NN matching on learned mid-level attributes [14] from ELF [8] features.

BR-SVM: Binary-relation based re-identification using SVMs [1, 13]. Note that BR-SVM has already been shown to decisively outperform the commonly applied RankSVM [21, 25] and prior metric learning methods [24].

DTR: Our proposed new Domain-Transfer re-identification model, using multi-kernel learning.

We evaluate re-identification performance using two metrics: For visualization, the normalized Cumulative Matching Characteristic (nCMC) curve, which indicates the probability of the correct match to a probe image appearing in the top n results from the gallery for varying n[1]. For quantitative summarization, we use the expected rank (ER) metric [8, 1], which is the mean rank of the true result[2]. This metric has the advantage that it reflects a physically meaningful quantity, which is how many items an operator has to scan in a ranked list before reaching the true match for the probe, and hence the average time it takes a human operator to find the true match using such a system [8].

4.4 Domain Transfer Experiments

[1]Here, higher curves are better; enclosing an area of 1 is perfect; and an area of 0.5 is random

[2]Lower is better; a mean rank of 1 is perfect; and a mean rank of half the gallery size is random

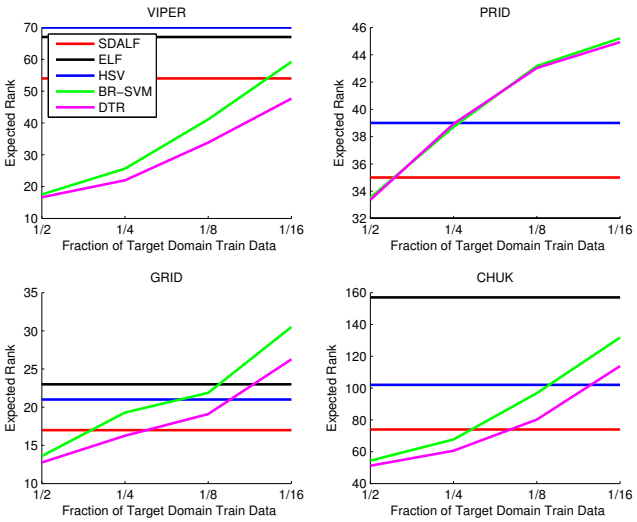

Figure 4: Re-identification performance as a function of volume of training data. Lower expected rank is better. Each dataset is evaluated as a leave-one-dataset out domain transfer problem. Our proposed DTR model systematically outperforms BR-SVM within-domain learning.

4.4.1 *Domain transfer compensation for a lack of target domain data*

We first evaluate re-identification performance as a function of target domain training data volume. Figure 4 summarizes the expected rank (ER) of each model for logarithmically varying volumes of training data. Also shown (flat lines) are the performance of LLF models SDALF (red), HSV (blue) and ELF (black). Clearly performance for the learning models degrades with sparser training data (Figure 4, ER of learned models higher to the right). However, our proposed DTR model (magenta) systematically outperforms the within-domain BR-SVM model [1] (green), especially with increasingly sparse data. We obtain between a margin of improvement over BR-SVM of 5-20%, 6-16% and 6-17% for VIPER, GRID and CUHK respectively. Meanwhile we obtain a margin of improvement over SDALF of up to 70%, 5%, 25% and 31% for VIPER, GRID and CUHK. At some point, for all learning models, the data will be sufficiently sparse that LLF approaches will be best. However DTR's margin over BR-SVM, means that standard LLFs can be outperformed with less training data than before. DTR model outperforms the best LLF with down to 1/16th data for VIPER, 1/4 data for GRID and 1/8th data for CUHK. Importantly, performance of DTR is usually dramatically better than simple nearest-neighbor on HSV (blue), which is the feature on which DTR was trained. Note that our weaker result on the PRID dataset can be understood by the generally poor performance of the HSV feature used by our DTR in this domain (see Section 4.5). This could in general be ameliorated by including other feature types within our MKL framework.

These results are also visualized in Figure 5, showing the CMC curve for each domain and data sparsity condition (line-style), of BR-SVM-based re-identification versus our domain-transfer model (color). The magenta CMC curves representing the transfer condition enclose the green non-transfer curves in each case. Finally, for GRID and CUHK we observe that even with the maximum volume of training data, transfer learning is still able to improve perfor-

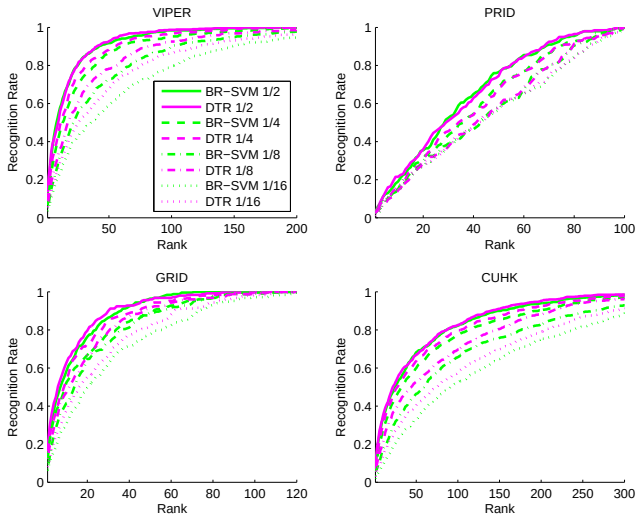

Figure 5: CMC curves for re-identification with and without transfer. Each line-type illustrates a different volume of training data. Ir each case the transfer CMC curve encloses the non-transfer curve.

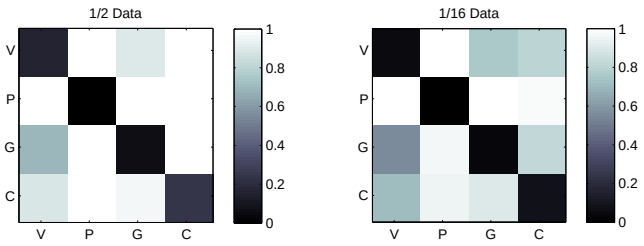

Figure 6: Cross-dataset affinity for re-identification. Darker blocks indicate a stronger cue.

	HSV[1]	SDALF[7]	ELF[8]
VIPeR	70.24	**53.64**	67.73
PRID	38.91	34.85	**32.50**
GRID	20.64	**16.70**	23.18
CUHK1	101.72	**73.70**	156.86

Table 1: Low-Level Features (LLFs) often do not generalize across domains. Columns are LLFs used in NN re-id on four public datasets (rows). We report Expected Rank (ER), lower scores are better. Bold scores are best; red scores are worst.

mance (Figure 5, solid magenta CMC curves enclosing solid green CMC curves; Figure 4, magenta curves under green curves).

Some visual examples of the improvement provided by our DTR approach over BR-SVM in each dataset are shown in Figure 7. In each case, the correct match to the probe is highlighted in green and the upper rows show the ranked matches by DTR versus ranked matches by BR-SVM in lower rows. Finally, Table 3 summarises some accuracies of each method at different ranks under the various conditions. In the majority of cases DTR clearly outperforms BR-SVM.

4.4.2 *Cross-Domain Analysis*

To provide some insight into the cross-domain results above, we present some analysis of the affinity between the major re-identification datasets by way of the learned weights for each kernel. Figure 6 plots the weights for re-identification for each target domain (rows) against the data source (columns). As expected, each dataset is highly relevant to itself (strong diagonal). Cross-dataset transfer is illustrated by the off-diagonal weights. It is evident that the VIPER re-identifier is relevant to assist both GRID and CUHK, but not PRID. Interestingly, there is some degree of transferability between VIPER, GRID and CUHK. However, the PRID dataset is neither useful as a source for any others, nor making use of any others as a source. This reflects the previous (Figure 4) results showing that the transfer performance for PRID was no better than the local only performance. Nevertheless, it is reassuring that in this case of an irrelevant source, the sparsity prior of our transfer framework was able to apply zero weighting (Figure 6) and hence avoided automatically negative transfer (Figure 4).

4.5 Additional Analysis

We next provide some additional analysis about the existing models and datasets to provide some insight into the domain transfer problem, and further validate our contribution as illustrated in Sections 3.1 and 4.4.

4.5.1 *Cross-domain generalization of low-level features*

To investigate the generalisation of low-level features, we perform re-identification using non-learned nearest-neighbor matching on the four datasets. The results are shown in Table 1, expressed as expected rank. The best results are highlighted in bold, and the worst in red. The important point to note here is that the best and worst low-level feature vary significantly by domain. That is, the ranking of different feature types does not generalize across domain. This highlights that selecting a generic good feature for all domains is not plausible, and leveraging learning based methods to adapt to the appearance of a given camera view is critical. We note that while SDALF [7] is the most effective feature overall, it is extremely computationally extensive to extract and thus of limited suitability for practical real-time applications.

4.5.2 *Cross-domain generalization of learning models*

We next perform re-identification using two learning methods: BR-SVM [1] and attribute learning [14], each of which provides at least near state-of-the-art performance when applied within a single domain. To evaluate cross-domain generalization, we train the methods on each domain (VIPER, PRID, GRID, CUHK) and apply them to all domains, thus obtaining 16 conditions[3] per method as shown in Table 2. The important points to note here are that (i) for both learning methods, the within-domain performance (diagonal of the table) is **significantly** better than the across-domain performance, i.e., *the methods do not directly generalize across-domain*; and (ii) the performance of the learning methods when applied across-domains is actually worse than the low-level feature methods (Table 1). This shows that achieving a useful level of performance with learning methods outside of closed-world benchmarks is non-trivial, and hence highlights the value of our contribution in this paper.

The above results together show that neither low-level features nor learning methods generalize directly and reliably across-domains. The only viable route to good performance therefore *is to learn a new model for each pair of cameras*. However the quadratic num-

[3]Minus ATTR for CUHK because we had no attribute annotation for this domain.

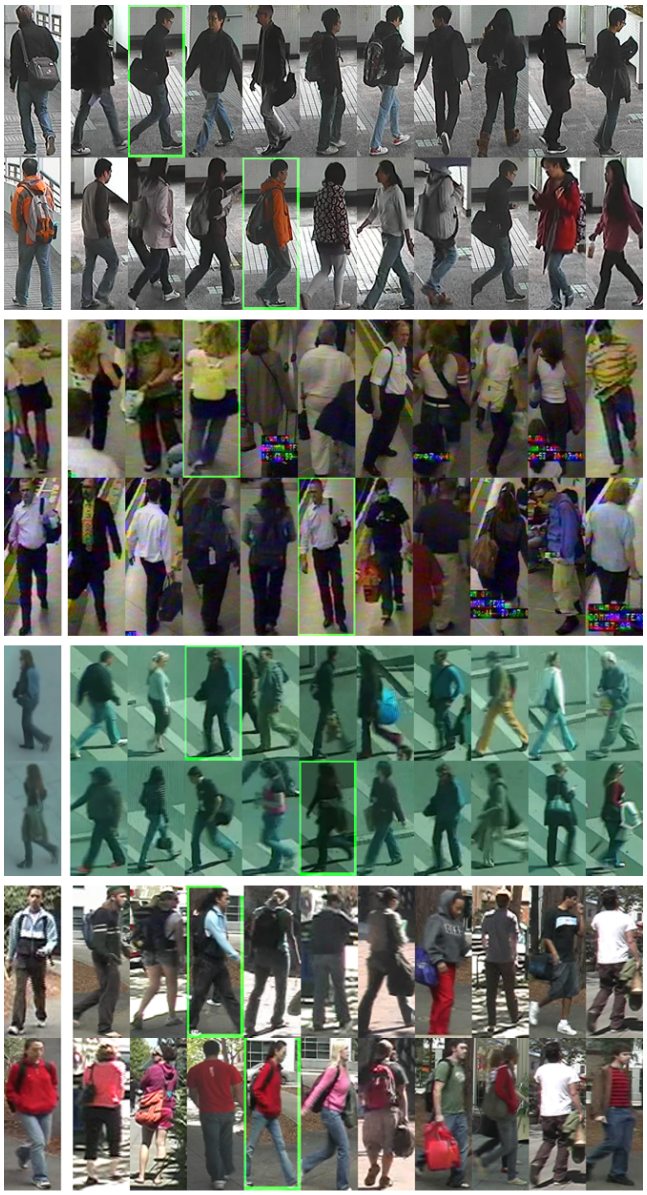

Figure 7: Some examples of early-rank matches from our system. The leftmost image is the probe image, with gallery images ranked by similarity to the right. The correct match to the probe is highlighted in green. From top to bottom, we present two examples from VIPeR, PRID, GRID and CUHK.

BR-SVM[1]	VIPeR	PRID	GRID	CUHK
VIPeR	**16.17**	50.23	39.01	166.11
PRID	155.23	**34.35**	59.70	240.72
GRID	119.38	49.17	**11.60**	202.55
CUHK	96.51	48.93	47.39	**52.24**

ATTR[14]	VIPeR	PRID	GRID	CUHK
VIPeR	**48.19**	43.38	26.22	185.61
PRID	98.82	**26.06**	39.01	201.50
GRID	94.28	46.69	**21.82**	194.29

Table 2: Learning-based re-id methods may transfer "blind" and retain some utility on untrained datasets but performance is penalised. Rows are training sources, columns are testing targets. Scores are the Expected Rank (ER), lower scores are better.

ber of pairs means that in practice exhaustive annotation is unreasonable beyond benchmark dataset testing exercises. This is turn shows the value of our contribution of transferring re-identification models for reducing training data requirements.

4.5.3 Computational Efficiency

The practically relevant aspect of performance is online matching efficiency. As with any SVM approach, our model is linear in the number of support vectors at test time. In particular it requires S times the computation of [1] for S source domains. In practice this means that our multi-kernel matching took about a millisecond per comparison (79ms including ELF feature extraction) with our unoptimized Matlab implementation. We note that despite making use of a strong model, this is still faster than state of the art LLFs such as SDALF [7], which requires approximately 460ms per comparison.

5. CONCLUSION

In this paper we introduced the problem of domain transfer for re-identification. This is a highly relevant challenge for taking re-identification out of closed-world benchmarks and making it useful for real-world applications. By formulating domain-transfer re-id as a SVM multi-kernel learning problem, we were able to achieve good performance on a wide variety of public benchmark datasets with a fraction of the training data required by previous methods. Moreover, our approach is able to evaluate available source domains automatically, weighting the relevant sources appropriately and ignoring irrelevant sources, thus avoiding negative transfer. We achieved these results despite the fact that the datasets used were unrelated and independently collected. With a wider selection of source datasets to choose from, the ability to construct a mapping to the target domain of interest (Figure 2) will be increased [12], and our results are therefore expected to only improve further as additional datasets are released.

There are many remaining issues for future work in order to improve performance and further reduce the amount of training required data for good performance can be achieved. So far we have only used the simplest color feature available, absolute performance should improve using better features as input, and multiple features can readily be included our MKL framework. With regards to negative instance selection, we thus far randomly selected 10 negative pairs per positive pair for training. Re-identification accuracy can be increased at the cost of additional computation by increasing this ratio [1]. However, more interesting is investigating active learning or instance mining approaches to optimally select the right instances from the quadratic number of pairs is there-

VIPeR, Rank:	1	10	20	50
BR-SVM 1/2	12.34	55.22	74.37	91.77
DTR 1/2	13.45	51.58	74.68	92.72
BR-SVM 1/4	6.49	42.72	62.66	86.71
DTR 1/4	9.02	45.09	68.20	88.29
BR-SVM 1/8	5.06	30.38	47.78	72.78
DTR 1/8	5.85	34.49	53.96	78.32
BR-SVM 1/16	3.48	22.15	37.50	58.86
DTR 1/16	5.54	27.06	44.15	67.41
PRID, Rank:	1	10	20	50
BR-SVM 1/2	3.00	19.00	35.50	76.50
DTR 1/2	2.50	21.50	36.00	74.50
BR-SVM 1/4	2.50	15.00	31.00	69.00
DTR 1/4	2.50	14.50	30.50	68.50
BR-SVM 1/8	2.00	15.00	30.00	58.00
DTR 1/8	2.00	15.00	28.50	58.00
BR-SVM 1/16	1.00	15.50	26.50	56.50
DTR 1/16	1.00	17.00	27.00	58.50
GRID, Rank:	1	10	20	50
BR-SVM 1/2	14.40	58.40	76.40	96.40
DTR 1/2	18.00	62.80	79.60	96.00
BR-SVM 1/4	12.00	52.40	66.80	87.60
DTR 1/4	16.40	53.60	72.80	93.20
BR-SVM 1/8	6.80	41.60	64.80	89.20
DTR 1/8	12.40	47.20	66.80	91.60
BR-SVM 1/16	6.00	30.80	49.60	77.60
DTR 1/16	8.00	37.60	55.20	81.60
CUHK, Rank:	1	10	20	50
BR-SVM 1/2	7.84	33.09	45.88	68.25
DTR 1/2	8.04	34.64	47.73	67.01
BR-SVM 1/4	5.46	27.22	40.21	59.90
DTR 1/4	6.80	29.90	42.89	63.61
BR-SVM 1/8	3.30	17.32	26.60	46.49
DTR 1/8	5.46	20.52	33.61	52.78
BR-SVM 1/16	1.96	8.97	16.49	32.99
DTR 1/16	2.89	13.71	23.20	40.41

Table 3: We present rank scores for each of the target datasets and target annotation volumes for both Binary-Rank SVM (BR-SVM) and our Domain Transfer Re-identification (DTR) approach. Higher scores are more desirable, as are earlier ranks which are more useful to human operators. Our approach shows that even with extremely reduced annotations on the target dataset, re-identification knowledge can be transferred in order to improve performance over low-level features alone.

fore an important open question. Finally, we would also like to transductively exploit the unlabeled data distribution in the target domain, and eventually move towards completely annotation free transfer learning for re-id.

6. REFERENCES

[1] T. Avraham, I. Gurvich, M. Lindenbaum, and S. Markovitch. Learning implicit transfer for person re-identification. In *Workshop on Re-Identification, ECCV*, 2012.

[2] F. R. Bach, G. R. G. Lanckreit, and M. I. Jordan. Multiple kernel learning, conic duality, and the smo algorithm. In *ICML*, 2004.

[3] L. Bazzani, M. Cristani, A. Perina, M. Farenzena, and V. Murino. Multiple-shot person re-identification by hpe signature. In *ICPR*, 2010.

[4] H. Bischof, P. M. Roth, M. Hirzer, P. Wohlhart, and M. Kostinger. Large scale metric learning from equivalence constraints. In *CVPR*, pages 2288–2295, 2012.

[5] S. Bąk, G. Charpiat, E. Corvée, F. Brémond, and M. Thonnat. Learning to match appearances by correlations in a covariance metric space. In *ECCV*, 2012.

[6] L. Duan, I. W. Tsang, D. Xu, and S. J. Maybank. Domain transfer svm for video concept detection. In *CVPR*, 2009.

[7] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2010.

[8] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, 2008.

[9] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *SCIA - Proceedings of the 17th Scandinavian conference on Image analysis*, SCIA'11, 2011.

[10] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof. Relaxed pairwise learned metric for person re-identification. In *ECCV*, 2012.

[11] O. Javed, K. Shafique, Z. Rasheed, and M. Shah. Modeling inter-camera space-time and appearance relationships for tracking across non-overlapping views. *Computer Vision and Image Understanding*, 109:146–162, 2008.

[12] L. Jie, T. Tommasi, and B. Caputo. Multiclass transfer learning from unconstrained priors. In *ICCV*, 2011.

[13] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and Simile Classifiers for Face Verification. In *ICCV*, Oct 2009.

[14] R. Layne, T. M. Hospedales, and S. Gong. Towards person identification and re-identification with attributes. In *Workshop on Re-Identification, ECCV*, 2012.

[15] W. Li, R. Zhao, and X. Wang. Human reidentification with transferred metric learning. In *ACCV*, 2012.

[16] C. Loy, T. Xiang, and S. Gong. Time-delayed correlation analysis for multi-camera activity understanding. *IJCV*, 90:106–129, 2010.

[17] F. Orabona and L. Jie. Ultra-fast optimization algorithm for sparse multi kernel learning. In *ICML*, 2011.

[18] F. Orabona, L. Jie, and B. Caputo. Online-batch strongly convex multi kernel learning. In *CVPR*, pages 787 –794, june 2010.

[19] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE TKDE*, 22(10):1345–1359, 2010.

[20] B. Prosser, S. Gong, and T. Xiang. Multi-camera matching using bi-directional cumulative brightness transfer functions. In *BMVC*, 2008.

[21] B. Prosser, W.-S. Zheng, S. Gong, and T. Xiang. Person re-identification by support vector ranking. In *BMVC*, pages 1–11, 2010.

[22] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *ECCV*, 2010.

[23] W. Schwartz and L. Davis. Learning discriminative appearance-based models using partial least squares. In *Computer Graphics and Image Processing (SIBGRAPI), 2009 XXII Brazilian Symposium on*, pages 322 –329, oct. 2009.

[24] W.-S. Zheng, S. Gong, and T. Xiang. Re-identification by relative distance comparison. *PAMI*, 2012.

[25] W.-S. Zheng, S. Gong, and T. Xiang. Transfer re-identification: From person to set-based verification. In *CVPR*, 2012.