Person Re-Identification by Support Vector Ranking

Bryan Prosser
bryan@eecs.qmul.ac.uk
Wei-Shi Zheng
jason@eecs.qmul.ac.uk
Shaogang Gong
sgg@eecs.qmul.ac.uk
Tao Xiang
txiang@eecs.qmul.ac.uk

Queen Mary Vision Laboratory,
School of Electronic Engineering and
Computer Science,
Queen Mary, University of London,
London, E1 4NS, U.K.

Abstract

Solving the person re-identification problem involves matching observations of individuals across disjoint camera views. The problem becomes particularly hard in a busy public scene as the number of possible matches is very high. This is further compounded by significant appearance changes due to varying lighting conditions, viewing angles and body poses across camera views. To address this problem, existing approaches focus on extracting or learning discriminative features followed by template matching using a distance measure. The novelty of this work is that we reformulate the person re-identification problem as a ranking problem and learn a subspace where the potential true match is given highest ranking rather than any direct distance measure. By doing so, we convert the person re-identification problem from an absolute scoring problem to a relative ranking problem. We further develop an novel Ensemble RankSVM to overcome the scalability limitation problem suffered by existing SVM-based ranking methods. This new model reduces significantly memory usage therefore is much more scalable, whilst maintaining high-level performance. We present extensive experiments to demonstrate the performance gain of the proposed ranking approach over existing template matching and classification models.

1 Introduction

Inter-camera object association, known as *re-identification*, enables tracking of the same objects through different disjoint cameras views either on-the-fly or retrospectively. In small scale CCTV networks, this problem can be approached using temporal information and visual appearance matching. However, in order to match individuals over significantly disjoint views in which the temporal transition time between cameras varies greatly from individual to individual and with a great deal of uncertainty, the problem is made harder as a model must now rely on appearance features alone. Furthermore, inter-camera variations in lighting conditions, changes in object orientation and object pose all make this task substantially harder still. Figure 1 demonstrates the difficulties arising from person re-identification in

BMVC 2010 doi:10.5244/C.24.21

(a) VIPeR Dataset (b) i-LIDS Dataset

Figure 1: (a) Sample image pairs from the VIPeR dataset [6] and (b) the i-LIDS dataset (http://www.ilids.co.uk). Each column represents a matching pair of observations with the top and bottom rows representing different camera views.

public spaces when appearance changes between camera views can render different people almost indistinguishable.

In order to tackle this problem, most existing work has concentrated on compiling feature sets as a template to describe an individual, followed by template matching using a direct distance measure chosen independently from the data. The common feature sets include major colours [9], combinations of colour and texture [10], or complex structural layouts [4]. Typical distance measures include histogram based Bhattacharyya distance [10], K-Nearest Neighbour classifiers [7], L1-Norm [12] or distance measures of relative proportions of colours [9]. Regardless of the choice of features and distance measures, re-identification by this approach is difficult because there is often too much of an overlap between feature distributions of different objects, so much so that given a probe image, an incorrect gallery image can appear to be more similar to the probe than a correct gallery image. Figure 1 shows that incorrect matches can often appear almost identical to the correct match. Based on the assumption that certain features are more suitable for matching than others, Gray and Tao [5] proposed to use Adaboost to search through a large feature set for those features that are more relevant (more discriminative) for more reliable re-identification. However, their feature selection becomes less effective if object feature distributions overlap severely in a multi-dimensional feature space as each of their weak learners only aims to seek the most relevant features in each feature dimension independently, not across the entire multi-dimensional feature space collaboratively.

In this work, we present a novel reformulation of the person re-identification problem. While previous approaches have looked at this problem as a classification of correct vs incorrect match, we propose an approach based on the information retrieval concept of document ranking [1]. Text document ranking aims to produce a ranked list of relevant documents based on a user query for document search. We consider that person re-identification given weakly distinctive (heavily overlapped) visual appearance has similar parallels. Given a query image, we wish to find those observed people who are most relevant, with a focus on the highest ranked person. The main difference between this approach and previous person re-identification techniques is that we are not concerned with comparing direct distance scores between correct and incorrect matches. Instead, we are only interested in the relative ranking of these scores that reflects the relevance of each likely match to the query image. By doing so, we convert the person re-identification problem from an absolute scoring problem

to a relative ranking problem. We thus avoid the need for seeking a maximum distance score and the assumption on the existence of large disparities between the distance of a true match and those of mismatches, nor the need for thresholding.

Ranking can be based on either Boosting or kernel based learning such as Support Vector Machines (SVMs). RankBoost [3] uses a set of weak rankers boosted to form a strong ranker. However, as the re-identification problem intrinsically suffers from a large degree of feature overlapping in a multi-dimensional feature space, picking weak rankers in each individual feature dimension, as considered by [3], is likely to lead to very weak rankers thus reducing matching effectiveness. In contrast, SVM based models such as RankSVM [8] seek to learn a ranking function in a higher dimensional feature space where true matches and wrong matches become more separable than the original feature space via the kernel trick. RankSVM is thus potentially more effective for coping with highly overlapped feature distributions in person re-identification. However, a main issue with running RankSVM on large datasets such as the LETOR dataset [1] is that it is computationally very expensive due to large amount of inequality constraints. As a result, RankSVM based learning to rank is limited as much fewer iterations can be performed, resulting in a sub-optimal ranker. Given the necessarily large number of candidate matches for person re-identification, this poses a severe scalability limitation on RankSVM's applicability to person re-identification.

Chapelle and Keerthi [1] proposed a primal-based RankSVM (PRSVM) to speed up existing RankSVM for document retrieval. We exploit this PRSVM model for addressing the person re-identification problem and show that PRSVM still suffers from another scalability limitation problem. Specifically, as the number of training samples grows, the number of negative samples increases non-linearly, coupled with the high feature dimensionality this means that the memory consumption can become unmanageable. To address this problem, we propose in this work an Ensemble RankSVM that uses a boosting principle on weak PRSVMs to maintain the computational efficiency in a high-dimensional feature space whilst overcoming the scalability limitation problem of PRSVMs in terms of memory usage. An additional benefit of this model is that it integrates the parameter tuning step of PRSVM into a boosting framework removing the need to rebuild training and validation sets.

For validating our model, we test and compare a selection of non-learning, learning and ranking methods on both the VIPeR dataset [6] and the i-LIDS dataset [2]. We show that a ranking based approach to person re-identification gives significant improvement over existing re-identification techniques. We also show that the proposed Ensemble RankSVM is able to achieve comparable results to conventional RankSVM whilst being computationally much more efficient thus having superior scalability.

2 Ranking People for Re-Identification

Person re-identification by ranking can be formulated as follows. Assume there exists a set of relevance ranks $\lambda = \{r_1, r_2, \cdots, r_\rho\}$ such that $r_\rho \succ r_{\rho-1} \succ \cdots \succ r_1$ where ρ is the number of ranks and $\succ$ indicates the order. In our problem there are only two relevance levels/ranks, that of relevant and irrelevant observation feature vectors, i.e. the correct and incorrect matches. Given a dataset $X = \{(x_i, y_i)\}_{i=1}^{m}$ where x_i is a multi-dimensional feature vector representing the appearance of a person captured in one view, y_i is its label and m is the number of training samples (images of people). Each vector $x_i (\in R^d)$ has an associated set of relevant observation feature vectors $\mathbf{d}_i^+ = \{x_{i,1}^+, x_{i,2}^+, \cdots, x_{i,m^+(x_i)}^+\}$ and related irrele-

[1] http://research.microsoft.com/en-us/um/beijing/projects/letor/Baselines/RankSVM.html
[2] http://www.ilids.co.uk

vant observation feature vectors $\mathbf{d}_i^- = \{x_{i,1}^-, x_{i,2}^-, \cdots, x_{i,m^-(x_i)}^-\}$ corresponding to correct and incorrect matches from another camera view. Here $m^+(x_i)$ ($m^-(x_i)$) is the number of relevant (related irrelevant) observations for query x_i and we have $m^-(x_i) = m - m^+(x_i) - 1$. In general, $m^+(x_i) << m^-(x_i)$ because there are likely only a few instances of correct matches and many incorrect matches. The goal of ranking any paired image relevance is to learn a ranking function δ for all pairs of $(x_i, x_{i,j}^+)$ and $(x_i, x_{i,j'}^-)$ such that the relevance ranking score $\delta(x_i, x_{i,j}^+)$ is larger than $\delta(x_i, x_{i,j'}^-)$.

2.1 Ranking by Support Vector Machine

Here we seek to compute the score δ in terms of a pairwise sample $(x_i, x_{i,j})$ by a linear function w as follows:

$$\delta(x_i, x_{i,j}) = w^\top |x_i - x_{i,j}|, \tag{1}$$

where $|x_i - x_{i,j}| = (|x_i(1) - x_{i,j}(1)|, \cdots, |x_i(d) - x_{i,j}(d)|)^\top$ and d is the dimensionality of x_i. We call $|x_i - x_{i,j}|$ the absolute difference vector.

Note that for a query feature vector x_i, we wish to have the following rank relationship for a relevant feature vector $x_{i,j}^+$ and a related irrelevant feature vector $x_{i,j'}^-$:

$$w^\top (|x_i - x_{i,j}^+| - |x_i - x_{i,j'}^-|) > 0, \tag{2}$$

Let $\hat{x}_s^+ = |x_i - x_{i,j}^+|$ and $\hat{x}_s^- = |x_i - x_{i,j'}^-|$. Then, by going through all samples x_i as well as the $x_{i,j}^+$ and $x_{i,j}^-$ in the dataset X, we obtain a corresponding set of all pairwise relevant difference vectors in which $w^\top(\hat{x}_s^+ - \hat{x}_s^-) > 0$ is expected. This vector set is denoted by $P = \{(\hat{x}_s^+, \hat{x}_s^-)\}$. A RankSVM model is then defined as the minimization of the following objective function:

$$\begin{aligned} &\frac{1}{2}\|w\|^2 + C\sum_{s=1}^{|P|}\xi_s \\ &s.t.\ w^\top(\hat{x}_s^+ - \hat{x}_s^-) \geq 1 - \xi_s,\ s = 1, \cdots, |P|,\ \xi_s \geq 0,\ \ s = 1, \cdots, |P|, \end{aligned} \tag{3}$$

where C is a positive parameter that trades margin size against training error.

One of the main problems with using an SVM to solve the ranking problem is the potentially large size of P. In problems with lots of queries and/or queries with lots of associated observation feature vectors, the size of P means that forming the $\hat{x}_s^+ - \hat{x}_s^-$ vectors becomes computationally challenging. Particularly, in the case of person re-identification, assuming there is a training set consisting of m person images in two camera views. The size of P is proportional to m^2, it thus increases rapidly as m increases. SVM-based methods also rely on parameter C, which must be known before training. In order to yield a reasonable model one must use cross validation to tune model parameters. This step requires the rebuilding of the training/validation set at each iteration, thus further increasing the computational cost and memory usage. Hence, the RankSVM in Eqn (3) is not computationally tractable for large-scale constraint problems due to both computational cost and memory use.

Chapelle and Keerthi [1] proposed a method based on primal RankSVM (PRSVM) that relaxes the constrained RankSVM and formulated a non-constraint model as follows:

$$w = \arg\min_w \frac{1}{2}\|w\|^2 + C\sum_{s=1}^{|P|}\ell\left(0, 1 - w^\top\left(\hat{x}_s^+ - \hat{x}_s^-\right)\right)^2, \tag{4}$$

where C is a positive importance weight on the ranking performance and ℓ is the hinge loss function. Moreover, a Newton optimisation method is introduced to reduce the training time of the SVM. Additionally, it removes the need for an explicit computation of the $\hat{x}_s^+ - \hat{x}_s^-$ pairs through the use of a sparse matrix. However, whilst the computational cost of RankSVM has been reduced significantly, the memory usage issue remains. Specifically, in the case of person re-identification, the spacial complexity (memory cost) of creating all the training samples is

$$O\left(\sum_{i=1}^{m} d \cdot m^+(x_i) \cdot m^-(x_i)\right), \tag{5}$$

where d is the feature dimensionality. Assuming there are L people in the training set, and $\frac{m}{L}$ images for each person, we then have $m^+(x_i) = \frac{m}{L} - 1$ and the spacial complexity can be re-written as:

$$O(d \cdot ((\frac{1}{L} - \frac{1}{L^2}) \cdot m^3 + (\frac{1}{L} - 1) \cdot m^2)). \tag{6}$$

This complexity is very high given large number of training samples m and high dimensional feature space d, and it cannot be reduced using PRSVM. In order to make RankSVM tractable for the large scale person re-identification problem we wish to resolve, we propose an Ensemble RankSVM to both significantly reduce the spacial complexity and solve the problem of tuning C in RankSVM.

2.2 Ensemble RankSVM

Rather than learning a batch mode RankSVM, we aim to learn a set of weak RankSVMs each computed on a small set of data and then combine them to build a stronger ranker using ensemble learning. More precisely, a strong ranker w_{opt} is constructed by a set of weak rankers w_i as follows:

$$w_{opt} = \sum_{i}^{N} \alpha_i \cdot w_i. \tag{7}$$

Learning the weak rankers. We divide a data set into groups and each weak ranker is learned based on that group of data. Specifically, assume there are in total L people $\mathscr{C} = \{\mathscr{C}_1, \cdots, \mathscr{C}_L\}$ and we equally divide them into n groups $G_1, \cdots, G_n$ without overlap, i.e. $\mathscr{C} = \bigcup_{i=1}^{n} G_i$ and $\forall\, i \neq j$, $G_i \bigcap G_j = \emptyset$. Then the training data set Z is divided into n groups $Z_1, \cdots, Z_n$ as follows:

$$Z_i = \{(x_i, y_i) | y_i \in G_i\}. \tag{8}$$

The simplest way to learn a weak ranker is to perform RankSVM on each subset Z_i. In order to avoid learning a rather weak ranker, we learn all weak rankers from a subset $\tilde{Z}_i$ and $\tilde{Z}_i = Z_i \bigcup O_i$ so that all weak rankers are not completely learned on separate datasets, where O_i is a subset of data of the same amount $|Z_i|$ randomly selected from the remaining data set $Z - Z_i$. This allows us to learn weak rankers on overlapping subsets. In our experiment (Section 3), for each Z_i and for each importance weight C, a weak ranker is learned; that is if there are s candidate values of parameter C, then $N = s \cdot n$ weak rankers are computed. This makes selection of the parameter C in the primal-based RankSVM unified into the ensemble learning framework, without using any additional cross-validation that requires reforming training samples.

For each $\tilde{Z}_i$, we compute a weak ranker w_i by using a primal-based RankSVM of Chapelle and Keerthi [1], which is tractable given a moderate size dataset. To compute RankSVM, we

first calculate a set of relevant and the related irrelevant absolute difference vectors in $\tilde{Z}_i$, denoted by $P_i = \{(\hat{x}_{i,s}^+, \hat{x}_{i,s}^-)\}$. Then, for some positive parameter C, the primal-based RankSVM solves the squared hinge loss function based on criterion of Eqn. (4).

Learning α_i. Suppose N weak rankers $\{w_i\}_{i=1}^N$ have been learned from the previous step. We now explore boosting to learn the weight α_i on the whole dataset X iteratively (see Algorithm 1). Specifically, at the t step, we first select the best weak ranker w_{k_t} that minimises the following cost function:

$$k_t = \arg\min_i \sum_{s=1}^{|P|} D_t^s \cdot \mathbf{I}_{w_i^\top(\hat{x}_s^- - \hat{x}_s^+) \geq 0} \tag{9}$$

where D_t^s is the weight of pairwise difference vectors at t step, $\sum_{s=1}^{|P|} D_t^s = 1$ and $\mathbf{I}$ is a boolean function. Then, D_t^s is updated as follows:

$$D_{t+1}^s = F^{-1} D_t^s \cdot \exp\left\{\alpha_t \cdot \left(w_{k_t}^\top(\hat{x}_s^- - \hat{x}_s^+)\right)\right\}, \tag{10}$$

where F is the normaliser such that $\sum_{s=1}^{|P|} D_{t+1}^s = 1$ and we initialise $D_1^s = \frac{1}{|P|}$. The weight α_t is then determined by:

$$\alpha_t = 0.5 \cdot \log\frac{1+r}{1-r}, \quad r = \sum_{s=1}^{|P|} D_t^s (w_{k_t}^\top (\hat{x}_s^+ - \hat{x}_s^-)). \tag{11}$$

Note that in order to ensure that the boosting algorithm both converges and updates the above weight, the input weak rankers w_i are normalised by $2 \cdot \max_{i,s} \left|w_i^\top (\hat{x}_s^- - \hat{x}_s^+)\right|$, so that $w_i^\top (\hat{x}_s^+ - \hat{x}_s^-) \in [-1, +1]$, as suggested in [3].

Compared to the batch mode RankSVM, the advantages of Ensemble RankSVM are two-fold. Firstly, it is not required to select the best parameter C for each weak ranker using cross-validation, as the ensemble learning algorithm automatically selects the optimal value of C by assigning different weights to weak rankers of different parameter values of C. Secondly and more importantly, each weak ranker is learned on a small set of data and the boosting process is based on the data projection values of each weak ranker. To learn each weak ranker, the spacial complexity is $O(d \cdot (\frac{1}{n^2}(\frac{1}{L} - \frac{n}{L^2}) \cdot m^3 + \frac{1}{n}(\frac{1}{L} - \frac{1}{n}) \cdot m^2))$, where d is the dimension of each image feature vector and n is the number of subsets. After learning each weak learner, for the ensemble learning process, the space complexity is $O(N \cdot ((\frac{1}{L} - \frac{1}{L^2}) \cdot m^3 + (\frac{1}{L} - 1) \cdot m^2))$ where N is the total number of weak rankers, and as the number of features $d > 2000$ in re-identification we have $N << d$. Overall the space complexity of our Ensemble RankSVM is around $1/n^2$ of that of the original RankSVM. Our experiments show the ensemble RankSVM can obtain comparable performance as the batch mode RankSVM but with significant reduction in memory usage.

3 Experiments

Datasets: Two challenging datasets were used in this work, the VIPeR dataset presented by Gray et al. [6] and a set of images extracted from the i-LIDS dataset [3]. Example images from both datasets can be seen in Figure 1. The VIPeR dataset consists of 632 pedestrian image pairs taken from two camera views. Each of the images has been scaled to a standard

[3] http://scienceandresearch.homeoffice.gov.uk/hosdb/cctv-imaging-technology/i-lids/scenarios

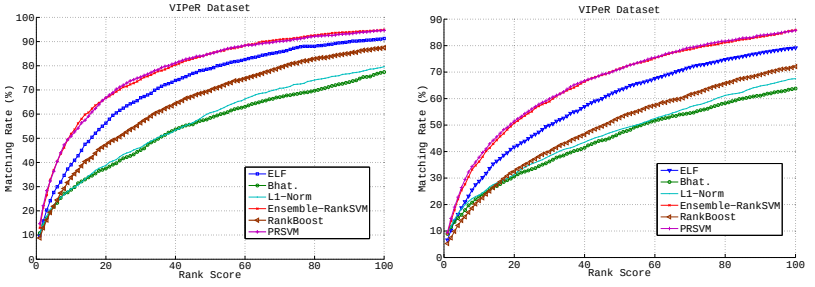
Algorithm 1: Algorithm of Ensemble RankSVM

Data: Pairwise relevant difference vector set P, Initial distribution $D_1 = \{D_1^s\}$
begin
  for $t = 1, \cdots, T$ **do**
    Select the best ranker w_{k_t} by Eqn. (9);
    Compute the weight α_t by Eqn. (11);
    Update the distribution D_{t+1} by Eqn. (10).
  end
end
Output: Output $w_{opt} = \sum_{t=1}^{T} \alpha_t \cdot w_{k_t}$

size and contains stark differences in pose, orientation and illumination making this dataset a good representation of challenging real world data. The i-LIDS dataset used in this work contained 208 image pairs that we have extracted from the HOSDB's i-LIDS multi-camera tracking dataset. Each person has two manually extracted images from two different camera views (one from each). The dataset contains a selection of camera view combinations from different videos in the i-LIDS multi-camera selection. As with the VIPeR dataset these images were scaled to a standard size and were not segmented from the background. As such the i-LIDS dataset in this paper has individuals captured under a diverse set of camera conditions. While the images from both datasets fit to each subject closely, some background noise is present in every image (see Fig. 1).

Feature Extraction: The features used were 8 colour channels (RGB, HS and YCbCr) and 21 texture filters (Gabor [2] and Schmid [11]) applied to the luminance channel. The Gabor filter used had parameters γ, λ, θ and σ^2 that were set to (0.3,0,4,2), (0.3,0,8,2), (0.4,0,4,1), (0.4,0,4,1), (0.3,$\frac{\pi}{2}$,8,2), (0.4,$\frac{\pi}{2}$,4,1) and (0.4,$\frac{\pi}{2}$,8,2) respectively. The Schmid filters used parameters τ and σ set to (2,1), (4,1), (4,2), (6,1), (6,2), (6,3), (8,1), (8,2), (8,3), (10,1), (10,2), (10,3) and (10,4) respectively, similar to [5]. A common bin size was selected for each feature channel of 16 bins. As different regions of the image are likely to contain visually distinct areas of interest some form of spatial representation is clearly needed. Some approaches use a single rectangle to capture the whole appearance [10], and others opt for a more complicated structural representation [4]. These approaches are either too simple or too constrained. Instead we choose a representation using six equal sized horizontal strips in order to roughly capture the head, upper and lower torso and upper and lower legs.

Methods for Comparison: We implemented the PRSVM by selecting parameter C in the set $\{0.0001, 0.005, 0.001, 0.05, 0.1, 0.5, 1, 10, 100, 1000\}$ using cross validation. For Ensemble RankSVM, the number of groups of data n was set to 5. We noted that the performance of method is insensitive to the value of n. For comparison, another four different existing person re-identification models were tested, including two non-learning distance based measures Bhattacharyya and L1-norm, a state-of-the-art Adaboost-based person re-identification system (ELF) [5], and a ranking based model using RankBoost [3]. All six methods were tested using exactly the same image feature set and image representation. We conducted five random trials and report the results averaged over the trials. Presented are the results of using 75% of the total samples for testing with the rest 25% for training, and 50% for testing with the rest 50% training. To evaluate comparatively all six re-identification methods, we display the cumulative matching characteristic (CMC) curve [13], which is based on the ranking of each of the gallery image with respect to the probe, thus resulting in the expectation of the



(a) 316 training, 316 testing (b) 158 training, 474 testing

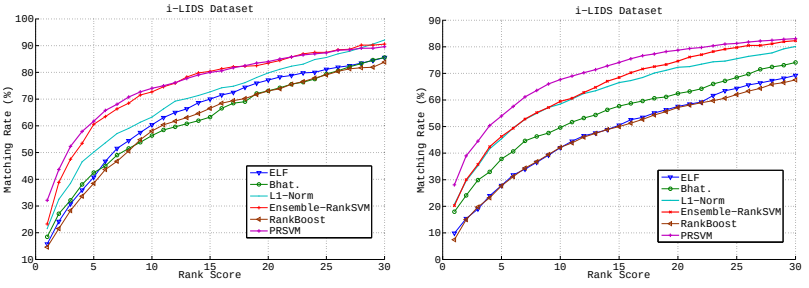
Figure 2: Cumulative Matching Characteristic (CMC) curves for the VIPeR dataset.

correct match being at rank r.

Ranking vs. Non-Ranking Approaches: Figure 2(a) shows the CMC curves for the VIPeR dataset with 50% (316) of the data used for training and 50% for testing while Figure 2(b) uses less samples for training (158) and more for testing (474). Due to the high number of possible matches coupled with the intrinsic difficulty of the data in which objects appear in different viewing conditions, the non-learning based distance measures (Bhattacharyya and L1-Norm) perform fairly poorly overall. In contrast the ELF method shows that by learning from training samples a more accurate distance measure can be obtained. It is clear that by employing a ranking framework we can gain a significant boost in performance with the PRSVM and Ensemble-RankSVM being the best overall. Similarly, the results on the i-LIDS dataset (Figures 3(a) and 3(b)) show that with the exception of RankBoost, explained below, the non-ranking methods still show lower overall performance. Some example query and ranked observation results can be seen in Figure 4.

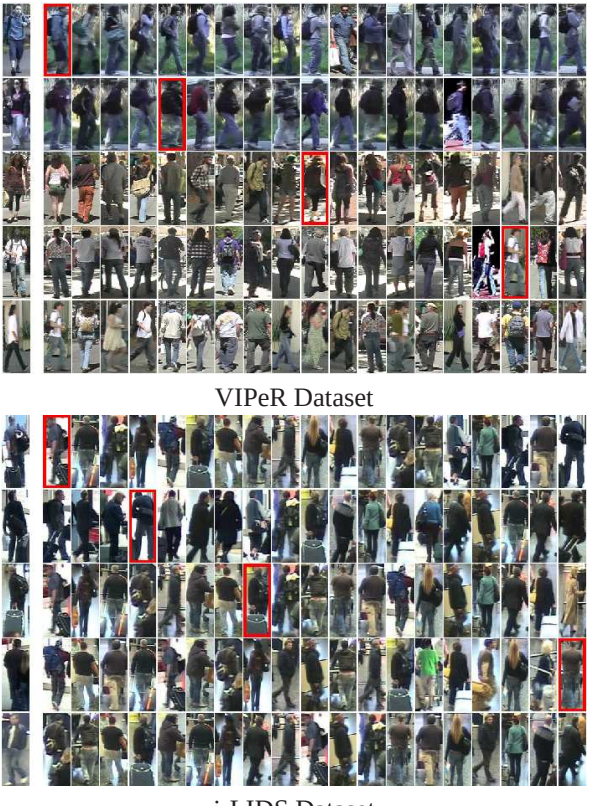
Ensemble RankSVM vs. PRSVM: On the VIPeR dataset (Figures 2(a) and 2(b)) the difference in performance between the Ensemble RankSVM and the PRSVM is negligible. This demonstrates that given a large dataset like VIPeR the Ensemble-RankSVM is an equal in terms of performance, while allowing a better scaling of memory usage (5.6GB needed for the PRSVM with 50% training on the VIPeR dataset, while the Ensemble-RankSVM needed only 740MB and this gap will widen on larger datasets). On the i-LIDS dataset (Figures 3(a) and 3(b)) the gap between them is slightly increased, with the Ensemble RankSVM having a lower overall score when the number of training samples is decreased. This is because that given a small training set, there are too few samples in each subset for learning a weak ranker which affects the performance of our Ensemble RankSVM. Nevertheless, since the primary goal of the introduction of the ensemble framework was to increase scalability, it is natural that on smaller experiments the PRSVM is more suitable.

SVM-based vs. Boosting: From both datasets it is clear that the RankSVM based methods are more suited to this task than the Boosting methods (ELF and RankBoost). The performance on the VIPeR dataset (Figures 2(a) and 2(b)), shows that the ELF method outperforms the RankBoost method with the setting used, both being significantly lower than the two SVM based ranking methods. On the i-LIDS dataset (Figures 3(a) and 3(b)) we can see that the RankBoost method shows similar results to the ELF, both of which are lower even than the baseline non-learning methods, indicating that the weak rankers/classifiers based on single feature channels are not effective. On this dataset the rank 1 matching rate of PRSVM is more than double those of ELF and RankBoost.



(a) 108 training, 108 testing (b) 52 training, 156 testing

Figure 3: Cumulative Matching Characteristic (CMC) curves for the i-LIDS dataset.



VIPeR Dataset

i-LIDS Dataset

Figure 4: Examples of re-identification on the VIPeR and i-LIDS datasets respectively. The first column indicates the query image, the middle column shows the PRSVM ranked results with the correct match in red.

Computation Time: All the experiments were run on a server machine with 8 CPU cores and 24GB of RAM in order to accommodate any required RAM consumption. The implementation was in Matlab, no special effort was made in terms of multi-threading so the

experiments generally took up 1 CPU core and at most 3 for some Matlab functions. The computation times of the SVM-based ranking methods were much lower than that of the ELF and RankBoost methods. For instance, for one-fold training and testing for the VIPeR dataset with a training size of 316, the PRSVM took about 11 minutes and the Ensemble-RankSVM 13 minutes while the ELF took over 5 hours and the RankBoost method 10 days.

4 Conclusion

In this work, we proposed a reformulation of the person re-identification problem as a learning to rank problem. We have shown that a ranking relevance based model can improve the reliability and accuracy in person re-identification under challenging viewing conditions. In addition, we formulated an Ensemble RankSVMs in order to overcome the computational scalability limitation of existing RankSVM models, which is especially severe when there is a large number of people to match and the feature space has a high dimensionality. We further incorporated the tuning of parameters for training a PRSVM in our ensemble learning method to eliminate the need for iterative cross-validation in model training. The proposed approach shows a significant improvement over other boosting based ranking models where the weak rankers were constructed from individual features.

References

[1] O. Chapelle and S. Keerthi. Efficient algorithms for ranking with svms. *Information Retrieval*, 18, 2010. to appear.

[2] I. Fogel and D. Sagi. Gabor filters as texture discriminator. *Biological Cybernetics*, 61(3):102–113, 1989.

[3] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer. An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research*, 4:933–969, 2003.

[4] N. Gheissari, T. Sebastian, P. Tu, and J. Rittscher. Person reidentification using spatiotemporal appearance. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 1528–1535, 2006.

[5] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *European Conference on Computer Vision*, Marseille, France, 2008.

[6] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *IEEE International Workshop on Performance Evaluation for Tracking and Surveillance*, Rio de Janeiro, Brazil, October 2007.

[7] M. Hahnel, D. Klunder, and K.-F. Kraiss. Color and texture features for person recognition. In *IEEE International Joint Conference on Neural Networks*, volume 1, pages 647–652, July 2004.

[8] T. Joachims. Optimizing search engines using clickthrough data. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142, New York, NY, USA, 2002. ACM. ISBN 1-58113-567-X. doi: http://doi.acm.org/10.1145/775047.775067.

[9] C. Madden, E. Dahai Cheng, and M. Piccardi. Tracking people across disjoint camera views by an illumination-tolerant appearance representation. *Machine Vision and Applications*, 18(3):233–247, 2007.

[10] B. Prosser, S. Gong, and T. Xiang. Multi-camera matching under illumination change over time. In *ECCV Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications*, Marseille, France, 2008.

[11] C. Schmid. Constructing models for content-based image retrieval. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 39–45, Hawaii, USA, December 2001.

[12] H. Wang, D. Suter, and K. Schindler. Effective appearance model and similarity measure for particle filtering and visual tracking. In *European Conference on Computer Vision*, pages 606–618, Graz, Austria, 2006.

[13] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu. Shape and appearance context modeling. In *IEEE International Conference on Computer Vision*, pages 1–8, Rio de Janeiro, Brazil, October 2007.