Uncooperative Gait Recognition by Learning to Rank

Raúl Martín-Félez [a,*] Tao Xiang [b]

[a] *Institute of New Imaging Technologies, Universitat Jaume I, 12071 Castellón de la Plana (Spain) Tel.: +34 964 728 358 Fax: +34 964 728 435*

[b] *School of Electronic Engineering and Computer Science, Queen Mary University of London, U.K.*

Abstract

Gait is a useful biometric because it can operate from a distance and without subject cooperation. However, it is affected by changes in covariate conditions (carrying, clothing, view angle, etc.). Existing methods suffer from lack of training samples, can only cope with changes in a subset of conditions with limited success, and implicitly assume subject cooperation. We propose a novel approach which casts gait recognition as a bipartite ranking problem and leverages training samples from different people and even from different datasets. By exploiting learning to rank, the problem of model over-fitting caused by under-sampled training data is effectively addressed. This makes our approach suitable under a genuine uncooperative setting and robust against changes in any covariate conditions. Extensive experiments demonstrate that our approach drastically outperforms existing methods, achieving up to 14-fold increase in recognition rate under the most difficult uncooperative settings.

Key words: Gait recognition, covariate conditions, learning to rank, transfer learning, distance learning

* Corresponding author.
Email addresses: martinr@uji.es (Raúl Martín-Félez), t.xiang@qmul.ac.uk (Tao Xiang).

1 Introduction

Gait can be used as a behavioral biometric. Compared to physiological biometrics such as fingerprint, iris and face, it has a number of distinctive pros and cons. The key advantage of gait for person identification is that it can operate from a distance and without subject cooperation. This makes gait ideal for situations where direct contact with or cooperation from a subject is not possible, e.g. surveillance in a public space. However, having uncooperative subjects also means that gait is susceptible to changes in various covariate conditions, which are circumstantial and physical conditions that can affect either gait itself or its perception. Examples of these conditions include clothing, surface, load carrying (e.g. carrying a bag), camera view angle, walking speed, and footwear type. This problem is illustrated in Fig. 1, which shows that due to significant changes in covariate conditions, especially view angle and clothing, features of different people (Fig. 1 (a),(d)) can be more alike than those of a same subject (Fig. 1 (a),(b),(c)).

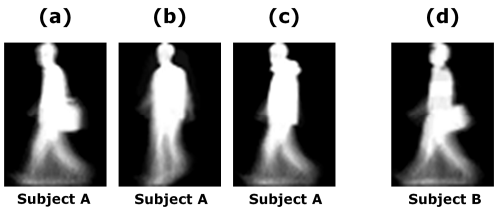

Fig. 1. Comparison of gait representations of Subject A ((a): with a bag, (b): a different viewpoint, and (c): wearing a bulking coat) and Subject B ((d): with a bag). Among (b), (c), and (d), (d) appears to be the best match to (a), because they share the same covariate conditions (view, carrying and clothing), which can easily lead to a wrong match.

As a classification problem (i.e. each person being a different class), gait recognition is challenging. This is not only due to the variable covariate conditions mentioned above, but also because of the lack of training data to cope with the large overlap between classes in the feature space. Specifically, each subject may be captured only in one sequence with a handful of gait cycles for feature extraction, resulting in an extremely under-sampled class distribution. Most existing approaches focus on extracting and selecting the best gait features that are invariable to different conditions [1,2,3]. However, they are based on human *a priori* knowledge (e.g. the most reliable features are in the most dynamic part of human body, i.e. legs) and select features in the highly

overlapped original feature space, which only lead to very limited success. In addition, these methods are designed for addressing specific types of covariate conditions but none of them can cope with large view angle changes. On the other hand, since gait features are particularly sensitive to view angle changes, completely different approaches based on feature transformation [4,5] are developed to deal with the view problem, which in turn do not work on other covariate conditions. Affine moment based features that are invariant to unknown covariant condition changes is proposed in [6]. However, it requires a cooperative setting, relies on clean silhouettes to be extracted from images, and is unable to cope with drastic appearance changes. So far, none of the existing approaches can address all covariate conditions which typically co-exist under an uncooperative setting.

Different from those feature selection and transformation-based methods, some learning-based approaches are also proposed [7,8,9]. These methods attempt to maximize the inter-class distance whilst minimizing intra-class variations, and can be applied after feature selection/transformation. However, they assume that the same classes/people must be present in both the training and test sets and represented with sufficient samples. Both assumptions are often not valid in practice. More importantly, most existing works use a gallery set composed of gait sequences of people under similar covariate conditions and evaluate their performance on a probe set of possibly different but fixed covariate conditions. They therefore make the implicit assumption that the data are collected in a cooperative manner so that the covariate conditions are known *a priori*. This essentially deprives gait of its most useful characteristic as an uncooperative and non-intrusive biometric.

In this paper, a novel approach is proposed which casts gait recognition as a learning to rank problem - a completely different perspective from previous approaches. More specifically, given a training and a test datasets consisting of gait features of different people who may even be captured from a completely different scene, we learn a bipartite ranking model. The model aims to learn a ranking function in a higher dimensional space where true matches and wrong matches become more separable than in the original space. The output of the model is a ranking function which gives a higher score if a pair of gait feature vectors belong to the same person than to different people. This new formulation has three distinctive advantages over the previous ones: (1)

This model is data-driven and can address all covariate conditions including view, i.e. one model for all. (2) Critically, unlike most previous approaches, it does not make any assumption about the gallery and probe sets having the same covariate conditions, either within each set or across the two sets. This makes it particularly suitable for uncooperative person identification, where gait should be used. (3) It does not suffer from the class under-sampling problem. Specifically, since it is based on bipartite ranking, there are only two classes during training: true matches and wrong matches; this also means that gait features from different people captured in different scenes/datasets can be used for training. In essence, it performs cross-class and cross-dataset transfer learning and is able to learn from an auxiliary dataset where plenty of data are available. We assume those data contain the covariate conditions we are modeling, but we do not assume that we know which particular gait sequence contains which covariate (uncooperative setting).

Extensive experiments have been conducted on three benchmark large gait datasets, covering both indoor and outdoor environments. They assess effects of changes in a number of covariate conditions (view angle, surface, carrying conditions and clothing changes) either alone or in combination under both uncooperative and cooperative settings. Results probe that our approach significantly outperforms other contemporary methods, especially under the most demanding uncooperative gait recognition tasks, where an up to 14-fold increase in recognition rate is observed. In addition, our framework is shown to be effective regardless which gait representation is chosen.

2 Related Work

Gait Representations – Most existing gait recognition techniques extract gait information from silhouettes obtained from video sequences. One of the simplest yet effective representations is Gait Energy Image (GEI) [7] (see Fig. 1), which is obtained by averaging silhouettes across a gait cycle. However, it has been shown to be sensitive to various covariate conditions [7,10]. To overcome this problem, a number of variants of GEI have been proposed. Yang *et al.* [3] propose to enhance the dynamic regions of GEI, which are located by a variance analysis. Bashir *et al.* [1] present a method to distinguish

the dynamic and static areas of GEI by using Shannon entropy at each GEI pixel, resulting in a new gait representation called GEnI. Shing and Biswas [2] improve the construction of GEI by using sway alignment instead of upper body alignment, which favors the perception of dynamic information. The basic idea of these methods is to select GEI features from the most dynamic areas of human body, i.e. legs and arms, which are less affected by changes in carrying conditions, clothing, and surface. Various other silhouette-based gait representations have been also developed, including Average Energy Image (AEI) [11], Gait History Image (GHI) [12], and Frame Difference History Image (FDHI) [13]. In addition, an optical flow based representation has been also adopted [14] for a more descriptive representation of gait dynamics. Most of the recently proposed gait representations are designed to be insensitive against certain covariate condition changes. However, none of them is capable of coping with all covariate conditions since there are so many of them and each one has effects on different aspects of gait [14]. The framework proposed in this paper can improve the recognition performance of any gait representation regardless whether they are designed to be invariant to different covariate condition changes or not, as demonstrated by our experiments (Section 4.3.1).

Gait Feature Selection and Transformation – Given a gait representation, recognition can be performed by template matching, i.e. using the one-nearest-neighbor (1NN) classifier based on a certain distance metric. However, to alleviate the effects of various covariate conditions, existing approaches have exploited feature selection and transformation. Feature selection methods such as [1,2,3] select those features from a gait representation that are more invariant to a given covariate condition. Nevertheless, selecting features in the highly overlapped original feature space typically relies on human a priori knowledge (e.g. the most reliable features are in the most dynamic parts of the human body) which only leads to limited success. Others propose to transform the features. On the one hand, some methods perform transformation to represent unknown gait conditions to recreate known covariate conditions. This is usually the preferred method to deal with camera view angle changes [4,5]. Gait features from one view are mapped to another by a learned View Transformation Model (VTM). Recognition is then performed after different views are transformed to the same. A different method is proposed by Bashir *et*

al. [15] which does not reconstruct gait features in different views, but models their correlation using Canonical Correlation Analysis (CCA) and uses the correlation strength as similarity measure. The main limitation of these transformation-based approaches is that the covariate condition(s) must be first recognized to know how the features have to be transformed. Attempts have been made to recognize clothing [16], load carrying [17] or camera viewpoint [15,18]. However, recognizing these covariate conditions under unconstrained conditions is challenging and far from being solved. Furthermore, all the previous feature selection and transformation methods were designed for addressing specific types of covariate conditions and none of them can cope with those combinations of conditions that typically occur in uncooperative scenarios. Recently Iwashita et al. [6] propose a transformation-based method designed to deal with any unknown covariate condition changes. It divides a human body region in multiple areas from which affine moment invariants are extracted as gait features and weighted according to its invariance to covariate condition changes. To compute these weights it requires a gallery set with image of the target subjects under their neutral appearance (e.g. wearing normal cloth and carrying no bag). This is essentially to assumpe a cooperative setting. As demonstrated in our experiments (see Section 4), its performance under uncooperative setting is much poored compared to ours.

Discriminative learning –Apart from feature selection/transformation and covariate condition estimation, there exist other methods which are based on discriminative learning and can be applied after feature selection/transformation. They attempt to maximize the inter-class distance whilst minimizing the intra-class variation. They range from Principal Component Analysis (PCA), combination of PCA with Linear Discriminant Analysis (LDA) [7], Marginal Fisher Analysis (MFA) [8], to general tensor discriminant analysis (GTDA) [9,19,20]. However, in order to learn these discriminant models, one has to make two assumptions: (1) sufficient training samples are available for each class/person; and (2) the training set and the probe set must consist of the same set of people, i.e. the training set needs to be the gallery set. However, these assumptions are often invalid under a real-world setting. For example, there could be only a single gait cycle captured for each person in the gallery set; in that case, LDA, MFA, and GTDA cannot be used. On the other hand, there may be

plenty of data from different people as an auxiliary training set but none of the existing methods could benefit from them. Our ranking-based transfer learning method does not make any of the two assumptions. Importantly, it can leverage those auxiliary data to compensate for the lack of samples in the gallery set.

Transfer Learning and Learning to Rank – Recently, cross-domain [21,22,23,24] and cross-dataset [25] transfer learning have received an increasing interest in computer vision. In these works, the auxiliary dataset and the target dataset are assumed to have the same classes (such as news videos from different countries or the same action classes captured in different scenes). The proposed work differs fundamentally from these works in that the gait classes in the auxiliary and target datasets are different (gait sequences belonging to different people). In this sense, our work is related to existing works on transferring knowledge between different but related classes (e.g. giraffes and horses [26]). They attempt to transfer the knowledge about the shared aspects between classes (e.g. both giraffes and horses have four legs). On the contrary, in this work we wish to transfer the features that are invariant to covariate condition changes across different classes/people. This is achieved by using a bipartite ranking framework, not exploited by the works mentioned above. Our method is inspired by the success of using learning to rank in document retrieval [27] and computer vision [28]. There exist other ranking models such as RankBoost [29] among others [30,31], but RankSVM is chosen because it is more suitable for a large scale learning problem with a severely overlapped feature space, as demonstrated in our experiments (see Section 4). To the best of our knowledge, this is the first work on formulating a learning to rank model for gait recognition. Note that transfer learning for gait recognition has been attempted before by Liu and Sarkar [32]. However, they simply apply a multi-class Linear Discriminant Analysis (LDA) model to transfer the learned discriminant space from an auxiliary dataset to a target dataset, assuming that it can be transferred across different classes. This is a very strong assumption which can be invalid in practice, since the number and nature of classes (people) might be completely different between the auxiliary training set and the test set. We demonstrate that our ranking based transfer learning approach outperforms their approach significantly (see Section 4.2.5).

Contributions – The main contributions of this work are three-fold: (1) To the best of our knowledge, gait recognition is for the first time formulated as a bipartite ranking problem in order to leverage data outside the target gallery set; (2) we introduce a novel solution to uncooperative gait recognition able to deal with any changes from covariate conditions or combinations of them without explicitly estimating them or manually designing appropiate gait features; and (3) we provide extensive evaluation of the proposed model against contemporary methods for a variety of public datasets under both uncooperative and cooperative settings.

An earlier and preliminary version of this work was published in [33]. Compared with [33], this paper provides a more systematic analysis of the ranking models, with a discussion on alternative ranking or distanc-learning based approaches. Much more thorough evaluations are also carried out. These include (1) additional experiments on a new dataset, (2) comparisons against a new baseline [6], (3) validation of our approach using different gait representations, (4) more insights into how our model works.

3 Methodology

3.1 Problem Formulation

Given a gallery set of gait sequences of different people with known identities, the problem of gait recognition can be considered as a retrieval problem. That is, given a probe gait sequence q of a walking subject s (which might be affected by some unknown covariate conditions), we wish to find the most relevant samples to q in the gallery set regardless the type of covariate factors that might affect them. The retrieved gallery gait sequences can be ranked according to a similarity/distance score. We propose to formulate this problem as a learning to rank problem by learning a ranking function able to push the correctly matched gait sequence (i.e. belonging to the same subject s) high on the ranking list, and ideally at the top leading to correct recognition.

As in any learning to rank setting, the training data set T consists of lists of

items with some internal order specified. This order is typically induced by a relevance judgment for each pair of items (q, d), in such a way that the higher the relevance score, the more relevant d is to q and it should be ranked at the top of the corresponding list by the learned model. For instance, in document retrieval, each document has a list of related documents that are relevant to it with different degrees of relevance. In our case, we employ a bipartite ranking model that uses a binary relevance judgment $y(q, d)$, where $y(q, d) = 1$ is given to two samples belonging to a same subject (*true match*), and $y(q, d) = 0$ is assigned otherwise (*wrong match*). Our learning to rank problem is thus a simpler case in that there are only two ranks unlike the aforementioned case of document retrieval. It is possible to introduce more ranks but this would mean comparing the degree of similarity between different people which inevitably would introduce subjectiveness and subsequently it would be subject to bias. For instance, given Person A as the query q, in order to determine the ranks among the gallery people consisting of Person A, B and C, one has to assess whether Person B's gait is less similar to that of Person A, compared to that of Person C, in order to decide whether B should be given a rank 3 and C rank 2 or the opposite. In a bipartite ranking formulation, the problem is much simpler: both B and C are given a rank 2.

Now the original multi-class identification problem is reformulated into a verification problem (genuine or impostor). The new verification task allows to learn information about how to match people's gait against various (unknown) covariate conditions, which can then be used to solve the original gait recognition problem. The reformulation into a two-class problem (true and wrong matches) means that each sample in T is used as query q against all the remaining training samples, which are assigned to one of the following two sets depending on its relevance indicator with respect to q:

- $D(q)^+ = \left\{d_1^+, d_2^+, \cdots, d_{|D(q)^+|}\right\}$, with $y(q, d_i^+) = 1$ for all $d_i^+ \in D(q)^+$, and $|D(q)^+|$ representing the number of relevant sequences (true matches) for the query sequence q.
- $D(q)^- = \left\{d_1^-, d_2^-, \cdots, d_{|D(q)^-|}\right\}$, with $y(q, d_i^-) = 0$ for all $d_i^- \in D(q)^-$, $|D(q)^-|$ representing the number of irrelevant sequences (wrong matches) for the query sequence q.

After using every single sequence as query in turn, each pair (q, d_j^+) or (q, d_j^-) is

represented by the entry-wise absolute difference between their feature vectors z_q and z_d, i.e. $\mathbf{x}(\mathbf{q}, \mathbf{d}) = |z_q - z_d|$ and it has a binary relevance judgment $y(q, d)$ as explained before. We thus obtain a set of preference pairs $P = \{(\hat{D}^+, \hat{D}^-)\}$, where $\hat{D}^+ = \mathbf{x}(\mathbf{q_i}, \mathbf{d_j^+})$ and $\hat{D}^- = \mathbf{x}(\mathbf{q_i}, \mathbf{d_j^-})$ going through all queries q_i as well as their corresponding $D(q_i)^+$ and $D(q_i)^-$. It produces a really higher number of new samples coming from the comparison of each possible pair of training samples, with a number of resulting true matches $\hat{D}^+$ much smaller than the number of wrong matches $\hat{D}^-$, since just a few samples per class (person) are available in T and all the samples of other classes lead to wrong matches. In this way, the original data sparsity problem is overcome and plenty of data are provided for the learning phase even when T has a small number of people.

During the training phase, the aim is to learn a ranking score function (ranking model) for each pair of query sample q and other training sample d in P as follows:

$$\delta(q, d) = \mathbf{w}^T \mathbf{x}(q, d) \tag{1}$$

where $\mathbf{x}(q, d)$ denotes the entry-wise absolute difference between the feature vectors z_q and z_d. Note that q and d refer to samples in P, so it represents the subtraction vector of a pair in P. The optimal $\mathbf{w}$ achieving the best agreement between the ranking induced by the ranking score δ and that induced by the relevance indicators y of the samples is sought, which assures that the score for any true match is higher than that for any related wrong match. From a different perspective, $\mathbf{w}$ can be considered as a weight vector that indicates the importance of each feature dimension towards the ranking score returned by δ. In other words, our ranking model performs implicit feature selection to identify features that are robust against those covariate conditions present in the training set.

In the test stage, we have a test set consisting of a probe/query set and a gallery set. The test set has no overlap with the auxiliary training set T, i.e. gallery and probe sets include samples from people different to those appearing in T. Given a query q in the probe set, the learned ranking score function δ is used to compute a score to each item d in the gallery set according to its relevance to q, taking the entry-wise absolute difference between their two feature vectors as input. Then the gallery items are sorted in descending order according to their scores to obtain a ranked list. If the sample at the top of the

ranking belongs to the same person as the probe one, it is considered a hit, otherwise, a mistake. In essence, we are performing template matching using the ranking score as a similarity measure.

An overview of the whole proposed framework is depicted in Fig. 2 as a summary of the previous explanation.

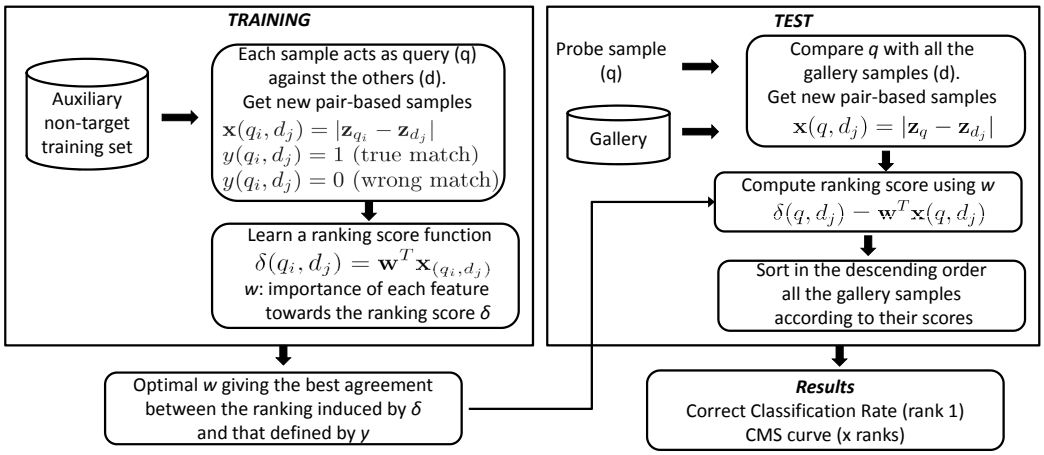

Fig. 2. Overview of our bipartite ranking based approach to gait recognition.

It is worth pointing out that: (1) After reformulating the gait recognition problem into a learning to rank problem, the learned knowledge is independent of the identity of people and only depends on the combinations of covariate factors existing in the auxiliary training set. (2) Only a single model is needed to cope with any covariate condition and combinations of them from those represented in the auxiliary training set. During testing, no assumption is made about the covariate conditions in gallery and probe sets. Both can contain variable unknown covariate factors. The only assumption is that those covariate conditions have appeared in the auxiliary training set and thus the learned model is robust against them. (2) Since the auxiliary training set contains different people/classes from the test set, cross-class and cross-dataset transfer learning can be easily performed.

3.2 Ranking Support Vector Machines

Although any ranking model [31,29,30] could be used with this framework, the primal-based Ranking Support Vector Machine (PrRankSVM) proposed by Chapelle and Keerthi [34] is chosen because it is able to deal with a large scale

and imbalanced learning problem with a severely overlapped feature space, exactly the problems that we are trying to address. In particular, PrRankSVM learns a ranking model $\mathbf{w}$ in a higher dimensional space where true matches and wrong matches become more separable than in the original feature space. Specifically, it aims to solve the following optimization problem:

$$\mathbf{w} = argmin_{\mathbf{w}} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{t=1}^{|P|} \ell \left(\mathbf{w}^T \left(\hat{D}^+ - \hat{D}^- \right) \right) \quad (2)$$

where t is the index of the preference pairs, $|P|$ is the total number of preference pairs used for training, C is a positive importance weight on the ranking performance and ℓ is the hinge loss function such as $\ell(t) = max(0, 1-t)^2$. The algorithm is computationally efficient, because it uses a Newton optimization to solve the non-constraint model of Eq. (2). This optimization allows to remove the explicit computation of the $(\hat{D}^+ - \hat{D}^-)$ pairs by using a sparse matrix. In this work, the C parameter is automatically selected by cross validation on the training set.

Notice that there are other algorithms for learning a RankSVM. In particular, Tsochantaridis et al. [35] introduced a structured output learning framework which achieves a similar level of computational efficiency as the primal RankSVM used in this work.

3.3 Discussion on Alternative Models

Relation to Other Ranking Models – There are many alternative ranking models; among them, the most noticeable one is perhaps RankBoost [29]. As indicated by Eq. (1), our RankSVM model essentially learns a weighted L1 distance between two feature vectors representing a probe and a gallery gait sequences respectively. It can thus be seen as a feature selection method by assigning a continuously valued weight (element of $\mathbf{w}$) to each feature dimension. All weights are learned simultaneously in our framework. RankBoost can also be considered as a feature selection method, by which a feature is removed (i.e. assigned a weight of zero) if the weak classifier learned based on that feature alone can only achieve a classification accuracy (between true and wrong matches) below a threshold (typically 50%). There is therefore a

vital difference between these two ranking methods: in RankBoost, each feature is quantified independently and sequentially, i.e. selected locally, whilst RankSVM quantifies the weights jointly and globally. The advantage of joint and global feature weight learning is demonstrated in our experiments reported later (Section 4.2.5).

Relation to Relative Distance Learning – As mentioned above, our RankSVM model learns a weighted L1 distance. There is a vast literature on distance/metric learning (see [36] for a review). Among the existing methods, those focusing on relative distance learning [37,38,39,40] are relevant to our problem, particularly the Probabilistic Relative Distance Comparison (PRDC) model proposed in [40]. Given two gait sequences q and d, represented by their entry-wise absolute difference vector $\mathbf{x}(\mathbf{q}, \mathbf{d})$, a distance function is learned as:

$$f(q, d) = \mathbf{x}(\mathbf{q}, \mathbf{d})^T \mathbf{M} \mathbf{x}(\mathbf{q}, \mathbf{d}) \tag{3}$$

where the model parameter $\mathbf{M}$ is a semi-definite matrix. The task of distance learning thus becomes estimating the optimal $\mathbf{M}$ such that the best agreement can be achieved between the ranking induced by the distance function and that induced by the relevance indicators of the training data, i.e. making sure that the distance between a true match pair is smaller than that of a relevant wrong match pair [1] . Comparing Eq. (3) with Eq. (1), a ranking model and a distance learning model appear to have a similar goal – to maintain the correct ranking order of the training data pairs as much as possible. The main difference is that a relative distance comparison model such as PRDC is a second-order feature quantification model able to take into account the joined effect between different features, whilst a RankSVM model is a first order one. This is reflected by the fact that the distance function f has a matrix $\mathbf{M}$ as parameters whilst the ranking score function δ has a vector $\mathbf{w}$ as parameters. Therefore, the latter has far fewer parameters (l^2 vs. l with l being the feature dimensionality). A second-order feature quantification model captures the correlations between different feature dimensions explicitly, being thus theoretically superior to a first-order one such as RankSVM. However, in practice, learning $\mathbf{M}$ with a

[1] 'relevant' in this context means that the person in the wrong match pair is the same person in the true match pair.

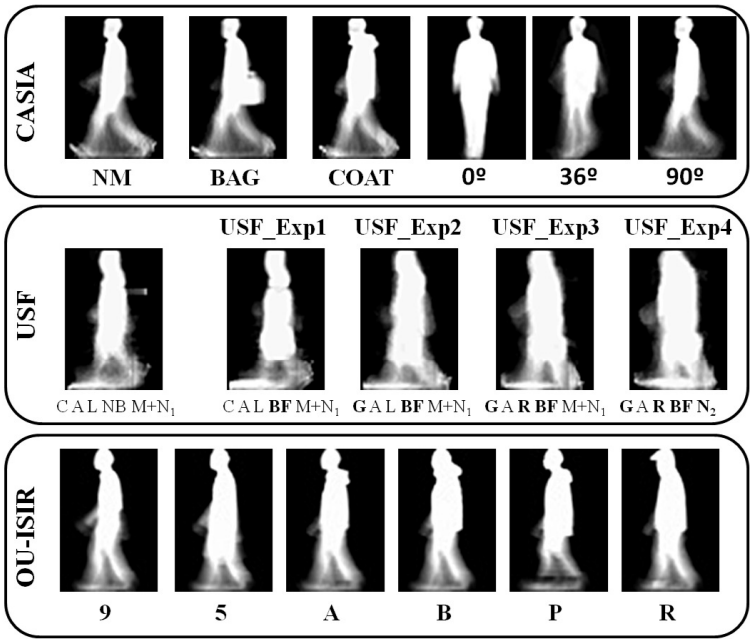

Fig. 3. Examples of GEIs with different covariate conditions.

high-dimensional input (typical in vision problems) is intractable and iterative optimization algorithms with various approximations have to be deployed [40], which makes the model vulnerable to local maximal and sensitive to feature noise. Our experiments in Section 4.2.5 show that a simpler model such as RankSVM is more preferable in practice for gait recognition.

4 Experiments

4.1 Experimental Settings

Datasets – Extensive experiments have been conducted on three of the largest benchmark gait datasets: CASIA [41], USF HumanID [42], and OU-ISIR [43] which cover both indoor (CASIA, OU-ISIR) and outdoor environments (USF). As Table 1 details and Fig. 3 illustrates, changes on different covariate conditions (camera viewpoint, load carrying, clothing, surface, footwear and time) either alone or in combination are the main obstacles to overcome.

Gait features – Unless otherwise stated, Gait Energy Image (GEI) [7] has

been used. Each GEI is normalized to a fixed size of 64×44 pixels using the silhouettes provided by each dataset. Example GEIs are shown in Fig. 3, which clearly show the more challenging nature of the outdoor environment in USF indicated by much noisier silhouettes. In Section 4.3.1, Active Energy Image (AEI) [11] has been used as an alternative gait representation to evaluate our model given different gait representations. Figure 11 shows examples of both gait signatures.

Settings – Firstly, the whole set of subjects considered in a particular experiment was randomly and equally split into two subsets, one for training (auxiliary set) and the other for testing (target set), in such a way that all samples of a same subject were assigned to the same subset. Secondly, the test set was further divided into a gallery set and a probe set. For an *uncooperative setting*, this partition was done in such a way that (1) each subject had at least a different covariate condition across the two subsets, and (2) both the gallery and probe sets had a mix of different covariate conditions. It is a challenging setting because for each probe sequence q of a subject s with a covariate type k, the gallery only contains sequences of the same subject s with a different covariate condition type, and a number of other subjects with the same covariate type k. On the contrary, when the test set is configured as a *cooperative setting*, all the gallery data share fixed covariate condition(s), while the probe set contains samples of different but also fixed covariate condition(s). All experiments have been repeated five times with different training/testing splits to mitigate the effects of subset singularities. We have made public details of our partitions for all the experiments in `http://www3.uji.es/~martinr/PR2013`.

Competitors – Three baseline gait recognition methods have been compared in all experiments. Note that all of them learn from the gallery set unlike our approach that uses a non-target auxiliary training set. They are:

- *1NN*. The k-Nearest Neighbor (1NN) classifier with $k = 1$ is used in the original high dimensional feature space.
- *1NN PCA*. The well-known Principal Component Analysis (PCA) technique is used to only keep those principal components accounting for a 99% of the variance.

- *1NN PCA+LDA*. As in [7,1], PCA is applied along with the Linear Discriminant Analysis (LDA) technique to obtain both the best data representation and the best class separability respectively. After LDA, the number of features become $n = c - 1$, with c being the number of classes (people identities).
- *Moments*. The method proposed in [6] is designed to cope with unknown covariate changes. It extracts affine moment invariants from GEI areas which are weighted according to its invariance to covariate condition changes to give a final similarity measure. We use the parameters suggested in their paper, i.e. we divide the human body in $K = 17$ horizontal areas, and we extract $M = 45$ affine moment invariants from each area. Note that it is not a transfer learning approach; thus no auxiliary dataset is required.

Other published methods have been also compared in individual experiments whenever possible although a direct comparison with the published results is always difficult. This is because as far as we know, only the work of Bashir et al. [1] follows an uncooperative setting. Most previous works were evaluated under a cooperative setting where all sequences in the gallery had the same covariate conditions, which were a priori fixed so were those in the probe. However, we have also conducted some cooperative-based experiments (Section 4.3.2) to directly compare our method with them. In addition, a number of transfer learning methods are introduced and compared in Section 4.2.5.

Performance Measures – Gait recognition performance is evaluated using *Cumulative Match Score (CMS)* curves [44]. A CMS curve shows the percentage of probe sequences the identity of which has been correctly recognized in the gallery among the top x matches. The averaged results from different trials are depicted.

4.2 Experiments under Uncooperative Setting

4.2.1 Results on USF Dataset

We first report results on the most challenging dataset of the three, the USF dataset. The *USF HumanID Gait Dataset (USF)* [42] is composed of videos

Table 1
Description of experiments carried out on CASIA, USF and OU-ISIR gait datasets under uncooperative settings. Covariate conditions: B-Carrying a briefcase or a bag, C-Clothing changes, S-Surface, V-View, T-Time.

Experiment	Covariate conditions	Subsets	#People	#Sequences
USF_Exp1	B	{C A L **NB** $M+N_1$, C A L **BF** $M+N_1$}	121	242
USF_Exp2	B S	{**C** A L **NB** $M+N_1$, **G** A L **BF** $M+N_1$}	117	234
USF_Exp3	B S V	{**C** A **L** **NB** $M+N_1$, **G** A **R** **BF** $M+N_1$}	117	234
USF_Exp4	B S V T	{**C** A **L** **NB** $M+N_1$, **G** A **R** **BF** N_2}	34	68
CASIA_Exp1	B	$\{NM_{90^\circ}, BG_{90^\circ}\}$	124	496
CASIA_Exp2	C	$\{NM_{90^\circ}, CL_{90^\circ}\}$	124	496
CASIA_Exp3	B C	$\{NM_{90^\circ}, BG_{90^\circ}, CL_{90^\circ}\}$	124	744
CASIA_Exp4	V	$\{NM_{90^\circ}, NM_{\theta^\circ}\}$	124	1488
		$\theta^\circ = 18^\circ \cdot X$ with $0 \leq X \leq 5 \in \mathbb{Z}^+$		
OU-ISIR_Exp1	C	{5, 6, 9, A, B, C, J, K, L, M, P, R}	55	660

of 122 subjects captured in an outdoor uncontrolled environment, which comprises up to five covariate conditions: (1) *surface*: subjects walk in two different surfaces, concrete (C) and grass (G); (2) *footwear*: two different shoe types (A) and (B); (3) *view angle*: subjects were captured by two cameras located in the left (L) and right (R) sides of the walking path yielding two view angles both close to side view, i.e. view change between L and R is small; (4) *carrying condition*: carrying a briefcase (BF) or not (NB); and 5) *time*: some subjects were only recorded in November (N_2), while others in both November (N_1) and May (M) which implies clothing changes among others. A total of 32 possible subsets can be obtained based on the different combinations of these covariate conditions in the gallery and probe sets.

We only report results on four representative configurations due to space limitation, resulting in four experiments as shown in Table 1. Starting from the easiest one (USF_Exp1), which copes with only one covariate condition (load carrying), the experiments get more challenging, and the hardest one (USF_Exp4) deals with four covariate condition changes between the gallery and probe at the same time (load carrying, surface, view angle, and time).

The results are shown in Fig. 4. It can be observed that: (1) The existing template matching (1NN) and learning based (1NN PCA+LDA) approaches yield very weak performances under an uncooperative setting. In addition, as expected, their performances become worse as the experiment gets harder.

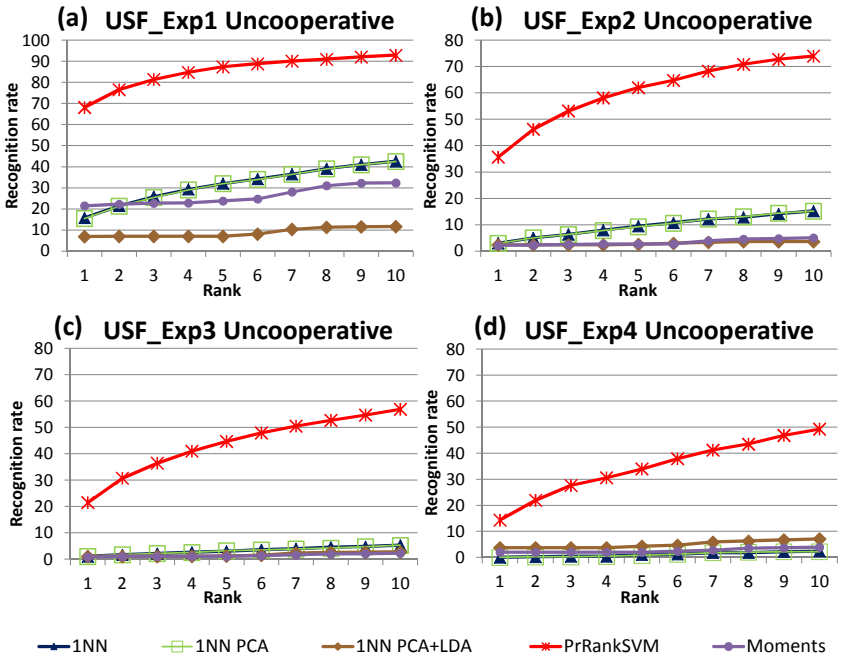

Fig. 4. CMS curves for the four experiments in USF under an uncooperative setting.

For instance, the best rank 1 matching rate (the correct classification rate) among the three drops from 15.8% in USF_Exp1 to 3.8% in USF_Exp4. (2) The learning based method (1NN PCA+LDA) does not fare better. In fact, its performance is even worse than the non-learning based methods in all but one experiments (USF_Exp4). This is because it suffers from the over-fitting problem due to the lack of training data when it learns from the gallery set. In addition, in these experiments, the intra-class variation for LDA is larger than the inter-class variation due to changes of covariate conditions. Under these conditions, LDA does not work as proven in [45]. (3) The affine moment-based method (Moments) shows a really weak performance in all the experiments, sometimes even worse than that of 1NN. This is caused by two reasons. First, the moment-based gait representation is sensitive to silhouette noise, which is a much severe problem for USF than the other two indoor datasets primarily due to unstable lighting condition. Second, this method is designed for cooperative setting, requiring that each target person must have an image of neutral appearance in the gallery set. This condition is obviously not met under our uncooperative setting. (4) Our approach (PrRankSVM) significantly outperforms the compared ones (up to 14-fold in USF_Exp4); and even though the rank 1 result of our approach for USF_Exp4 is poor, the rank 10 result is almost 50%, which makes it of practical use for assisting a

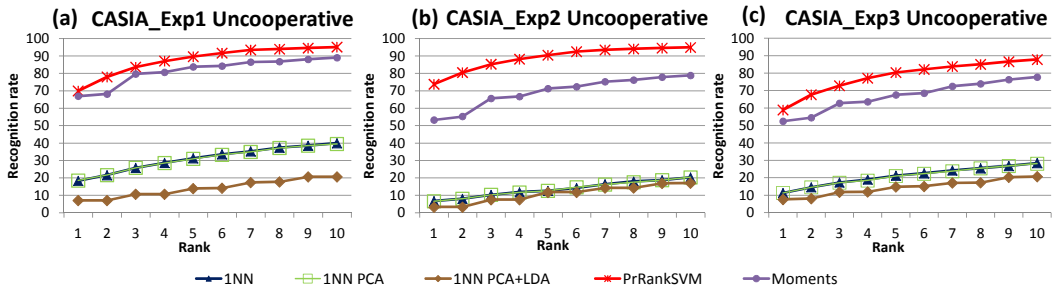

Fig. 5. CMS curves for the first three experiments with CASIA.

human operator in matching people (i.e. the operator does the final matching from a list of candidates selected by the model).

4.2.2 Results on CASIA Dataset

Perhaps the most widely used gait dataset, the *CASIA Gait Database* - Dataset B [41] contains 124 subjects captured under three different covariate condition changes: load carrying, clothing, and view angle. Note that the view changes are much bigger in CASIA than in USF - each subject was captured from 11 different view angles from frontal view (0°) to back view (180°) including side-view (90°). For each view, each subject has 10 gait sequences: six normal (NM) where the person does not carry a bag or wears a coat, two carrying-bag (BG) and two wearing-coat (CL). All the videos were recorded indoors with a uniform background and controlled lighting.

Carrying and clothing condition changes – Three experiments were first conducted to evaluate the different approaches under carrying and clothing condition changes. As shown in Table 1, CASIA_Exp1 focuses on carrying conditions alone, CASIA_Exp2 on clothing changes alone, and CASIA_Exp3 explores both covariate conditions together. For all the three experiments, only side view (90°) gait sequences were used; the effect of view will be investigated in a separate experiment later in this section. From the 10 side-view sequences available for each subject in CASIA, two normal sequences (NM) out of six were randomly selected along with the two in which the subject wears a coat (CL), and the other two carrying a bag (BG). It gave a total of six sequences per person, and 744 in total for CASIA_Exp3. A lower number of sequences were thereby chosen when only one covariate condition change is considered in both CASIA_Exp1 and CASIA_Exp2 (see Table 1).

From Fig. 5, similar observations can be made as those in the USF experiments, although in general higher recognition rates were obtained for all methods because of the cleaner silhouettes as compared to the USF ones (see Fig. 3). Specifically, the results show that: (1) Consistent to the results reported in other works [1,10], clothing changes seem to affect gait more than carrying condition changes, either alone (CASIA_Exp2) or combined with other condition (CASIA_Exp3), for the three compared baseline approaches. However it is not the case for our ranking approach, with which very similar results were obtained for all three experiments. This is the strength of a data-driven learning based approach – it quantifies the features and learn the optimal ranking/distance function given any combination of covariate condition changes. (2) Similarly to the results in USF experiments, 1NN PCA+LDA suffers from over-fitting and its performance is the poorest among all compared methods. (3) Comparing with the results in Fig. 4, The affine moment-based method (Moments) has a much improved performance, significantly outperforming the 1NN based baselines. However, our method (PrRankSVM) still has a clear margin over Moments. This result suggests that being able to obtain clean silhouettes is critical for Moments. Nevertheless, the intrinsic cooperative setting assumption still leads to its inferior performance.

As mentioned before, due to the uncooperative setting we use, our results are not directly comparable with most results published in the literature, which were obtained under a cooperative setting. The only exception is [1], which used a similar setting to our CASIA_Exp3 with a gallery set also containing a mix of NM, BG and CL sequences. Their rank 1 result of 53% is comparable with our 58.9% in CASIA_Exp3. However, there is still a number of vital differences: (1) We used half of the 124 subjects for training whilst they used all for gallery and probe. Importantly their model was learned using the gallery set, thus using the same people as in the probe set; (2) they considered all the NM sequences instead of only two per person in the gallery set to make sure there were enough data in the gallery set to learn their model; and (3) they need to re-learn the LDA model for each pair of gallery and probe sequences, whilst our approach only learns the ranking model once and is able to very efficiently compute the matching score during testing by using Eq. (1). Overall, our method is more generally applicable (i.e. it can deal with any covariate condition changes including view angle, and can work even with just a single

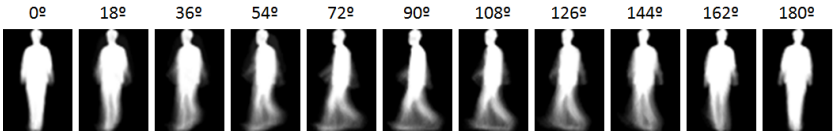

Fig. 6. A subject from CASIA seen from different view angles.

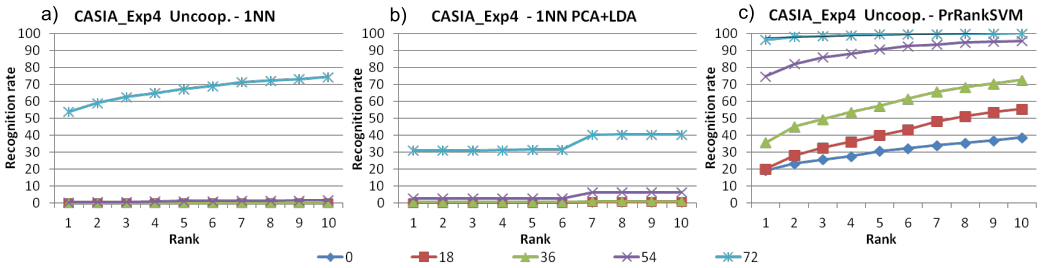

Fig. 7. CMS curves for the cross view experiment (CASIA_Exp4) in CASIA.

sequence per subject in the gallery set) and efficient for real-time applications (see Section 4.2.5 for computation time).

View changes – The experiment CASIA_Exp4 is designed to evaluate our ranking approach under large view angle changes. It aims to match sequences of people seen in their side view (90°), which is considered the best angle for gait to be effective, with respect to sequences in some of the other view angles available in CASIA: $\theta = \{0°, 18°, 36°, 54°, 72°\}$. View angles greater than 90° are not reported because they tend to achieve performances similar to those of their corresponding symmetrical angles [4,1], i.e., 108° is similar to 72°, 126° to 54°, and so on. For each possible pair $(90°, \theta_i)$, an uncooperative setting was adopted as follows. Only the six NM sequences of each subject were considered, and all of them were assigned to either the training or test set. Thus, in the training set, each selected person was represented by six NM sequences from 90° and other six from the other view angle θ_i. The test sequences were split into gallery and test following the procedure explained in Section 4.1. Detailed information of this experiment can be found in Table 1.

Figure 7 shows a comparison of the results of two baseline methods (1NN and 1NN PCA+LDA) and our approach. Each plot depicts the CMS curves for all possible pairs $(90°, \theta_i)$. It is clear that, under an uncooperative setting, both non-learning based methods fail miserably when the view angle difference is beyond 18°. This is unsurprising because, as can be seen in Fig. 6, the GEIs

Table 2
Description of clothing combinations used for OU-ISIR_Exp1. Legend: RP-Regular pants, BP-Baggy pants, SP-Short pants, RC-Rain coat, LC-Long coat, FS-Full shirt, Pk-Parker, DJ-Down jacket, Mf-Muffler.

Clothing combinations	5	6	9	A	B	C	J	K	L	M	P	R
Upper-body	RP	RP	RP	RP	RP	RP	BP	BP	BP	BP	SP	RC
Lower-body	LC	LC	FS	Pk	DJ	DJ	LC	FS	Pk	DJ	Pk	RC
Complements	-	Mf	-	-	-	Mf	-	-	-	-	-	-

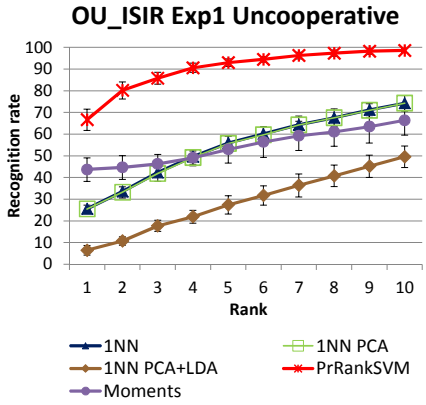

Fig. 8. CMS curves for OU-ISIR_Exp1

of a same subject under large view angle changes look completely different. In addition, given a probe GEI q_t of a subject s in a specific view angle θ_i, the gallery contains samples of s but from a view angle different from θ_i, while it also comprises plenty of other samples from other subjects in the same view angle θ_i. Under this setting, the recognition rate could be worse than random guess because it is almost certain that the probe sample q_t will be matched with a wrong subject with the same view angle θ_i. In comparison, our approach gives much better results especially when the view angle difference becomes larger owing to its ability to transfer useful information about the invariant features under those view change from the auxiliary dataset.

4.2.3 Results on OU-ISIR Dataset

The *OU-ISIR Gait Database* [43] - Dataset *B* includes videos of 68 subjects walking on a treadmill and captured from their side view in an indoor environment. Figure 3 shows that clean silhouettes can be obtained, similar to those in CASIA. This dataset is ideal for studying the effects of clothing changes on gait recognition; in particular, subjects were recorded under up to

32 possible clothing combinations with variations in pants, shirts, skirts, hats, among others. Note that not all combinations were recorded for all subjects. In our experiment (OU-ISIR_Exp1 in Table 1), only the clothing combinations with most of the subjects represented were selected. As a result, 55 subjects were chosen under the 12 clothing combinations summarized in Table 1. More details about the clothing conditions are given in Table 2. We randomly selected 24 subjects for the auxiliary training set. The remaining 31 subjects were used for gallery/probe in the target set. The results in Fig. 8 show that under drastic clothing changes, such as those shown in Fig. 3, our method is able to correctly identify almost 70% of the subjects, with this recognition rate increasing to more than 90% at rank 5. This results show the generalisation capability of our method – it learns a single model to deal with up to 12 clothing combinations in the probe images. Again, our method beats the other compared methods by a large margin. In particular, it is noticed that even with clean silhouettes, the performance of Moments is only at par with the 1NN PCA+LDA method, as opposed the results on CASIA in Fig. 5. This result suggests that the cooperative setting assumption is more problematic given a larger variety of covariate condition changes.

4.2.4 Analysis on What Has Been Learned

The RankSVM model essentially learns a weighted L1 distance/similarity function as the ranking function, with the weight $\mathbf{w}$ as its model parameter. Since each feature correspond to one pixel on a GEI, we can visualize the learned $\mathbf{w}$ as an image which tells us which part of the GEI is more invariant than others under the covariate condition changes in an auxiliary dataset. Figure 9 shows what has been learned by the model in four experiments. It can be seen from Fig. 9a that in the easiest experiment on CASIA (clean indoor background, one covariate condition change only), the high weight values distribute evenly across the GEI with the exception of the center of the body where no useful information exist either for gait itself or the body shape appearance. The areas that are likely to be affected by carrying (see Fig. 1a) are also largely given low weights. When both carrying and clothing condition changes are introduced in CASIA_Exp3, Fig. 9b shows that the important features are now focused on a more narrow band, particularly surrounding the leg and head area, where the effects of clothing and carrying conditions are minimal.

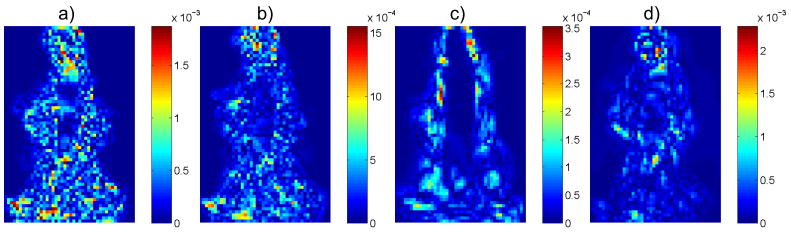

Fig. 9. Visualization of the learned feature weight by RankSVM. a) CASIA_Exp1, b) CASIA_Exp3, c) USF_Exp4, and d) CASIA_Exp4, $\theta = 36°$. The absolute weight values are shown with higher values indicating high importance/more robust against covariate changes.

As more covariate conditions (particularly surface and small view changes) are added in USF_Exp4, the high weight regions become even smaller and concentrated more on the outer boundary of human body (Fig. 9c). Therefore from Fig. 9a to c, one can see clearly how less and less features are favored which correspond to areas that are least affected by a combination of covariate conditions. However, the large view change experiment CASIA_Exp4 shows very different characteristics in the selected features (Fig. 9d). Comparing a GEI of $\theta = 36°$ with that of $\theta = 90°$ in Fig. 6, one can see that a large proportion of the leg areas will not be useful to match subjects directly. Instead, the model discovered that the head movements are more robust against view change as reflected by the high weight values in the head area. Overall, the results in Fig. 9 show that an intuitive and meaningful feature weighting has been learned by the RankSVM model.

4.2.5 Comparison with Alternative Transfer Learning Methods

As discussed in Section 3.3, other transfer learning methods can be used instead of RankSVM. These include alternative ranking methods, distance learning methods and other learning methods that can be trained on an auxiliary dataset. In this experiment, the following approaches are compared:

- *1NN PCA+LDA Transfer Learning (1NN PCA+LDA TL)* is the method introduced in [32]. In this method, PCA and LDA transformations are learned from the auxiliary set and then applied to the gallery and probe samples before matching with 1NN.
- *non-rankSVM Transfer Learning (non-rankSVM TL)* is essentially a binary linear SVM where the two relevance judgment values (true match and wrong

match) are used as the class labels. Compared to the RankSVM, both models use the entry-wise absolute difference features as input and differ only in the formulation of their cost functions.

- *RankBoost* [29] is a boosting-based learning to rank algorithm that selects a subset of optimal features sequentially and independently from the original feature space.
- *Probabilistic Relative Distance Comparison (PRDC)* [40] is a relative distance learning method which was originally formulated for solving the person re-identification problem.

Among the four transfer learning methods, both RankBoost and PRDC are similar to our RankSVM model in that all three can be considered as both ranking and relative distance learning methods (see discussions in Section 3.3). The differences lie on how the features are selected (globally in RankSVM and PRDC vs. locally in RankBoost) and how the feature correlations are modeled (explicitly in PRDC, implicitly in RankSVM, and none in RankBoost). The other two compared methods, 1NN PCA+LDA TL and non-rankSVM TL are both non-ranking based. However they differ significantly from each other – non-rankSVM TL still aims to learn a distance/score function that best separates two classes, the true matches and the wrong matches, whereas 1NN PCA+LDA TL learns an optimal subspace where the multiple classes/people in the auxiliary set are most separable. In addition, non-rankSVM TL uses absolute differences between feature pairs as input, while 1NN PCA+LDA TL uses the original gait feature vectors as input. Therefore non-rankSVM TL is much more similar to our RankSVM with the same input and output. The only difference to RankSVM is that non-RankSVM TL employs a harder constraint on maximizing the distance between the true and wrong match class samples whilst RankSVM imposes a 'soft' constraint on maintaining the ranking order – satisfying the former means automatically satisfying the latter, but not vice versa.

A comparison of results from CASIA_Exp3 and USF_Exp4 is shown in Fig. 10 where the results of two non-transfer learning methods (1NN and 1NN PCA+LDA) are also included. The key findings are:

- All the transfer learning approaches significantly outperform the compared baseline approaches, which proves the benefits of this strategy.

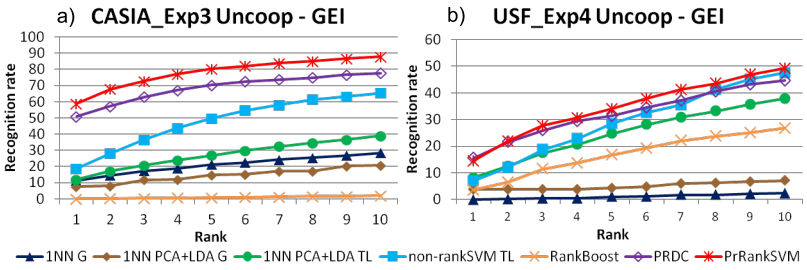

Fig. 10. Comparing different transfer learning approaches.

- 1NN PCA+LDA TL achieves an inferior performance to that of RankSVM and PRDC. The main reason is that despite the learned discriminant subspace contains transferable information for achieving robustness against covariate condition changes (as stated in [32]), it also contains information about how the gait of people in the auxiliary training set differs from each other, and the latter information is non-transferable because the target test set contains a completely different set of people.
- Converting the multiple class classification problem into a verification problem has some benefits as demonstrated by the performance of non-rankSVM TL. However, its stronger constraint seems to have a negative effect leading to clearly lower recognition rates than those of RankSVM and PRDC, particularly at low ranks.
- Among the compared transfer learning methods, RankBoost achieves the lowest performance that demonstrates the importance of selecting features globally. In particular, since the feature dimensionality is fairly high in our case (2816 features), a weak ranker learned using a single feature as in RankBoost would be too weak to be reliable.
- On the contrary, PRDC achieves the closest performance to RankSVM due to the similar nature of the two models. The noticeable improvement of RankSVM over PRDC can be attributed to the simpler formulation of the cost function and the more numerically reliable solver available of the optimization problem. The results suggest that this can more than compensate for the lack of explicit modeling of the correlation between features.

Table 3 shows the training and testing time for different methods on CASIA_Exp3. It can be seen that in terms of testing time, RankSVM is identical to non-rankSVM because both are doing weighted L1 distance during testing. The testing time is fairly similar to 1NN which does a (unweighted) L2 distance. 1NN PCA+LDA and 1NN PCA+LDA TL have very similar test-

Table 3
Comparison of training and testing time for the different methods on CASIA_Exp3.

Methods	Training time	Test time (per image)
1NN	0 s	5.41 ms
1NN PCA+LDA	0.22 s	0.53 ms
1NN PCA+LDA TL	1.34 s	0.49 ms
non-rankSVM TL	123.71 s	7.05 ms
RankBoost	2735.20 s	7.50 ms
PRDC	1395.78 s	6.97 ms
PrRankSVM	1070.33 s	7.05 ms

ing time and are faster than 1NN. All three compute L2 distance, but the two learning based methods operate in a much reduced PCA+LDA space. In terms of training time, the ranking/distance learning based transfer learning models are more expensive than the others, with the RankBoost being the most costly one.

4.3 Further Evaluations

4.3.1 Effects of Different Gait Representations

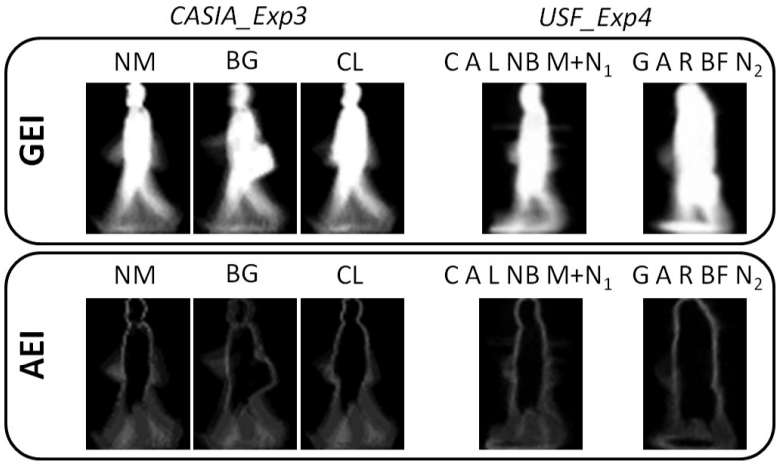

Fig. 11. Examples of different gait representations (GEI and AEI) for different experiments and datasets.

In this experiment, a different gait representation, Active Energy Image (AEI) was used in CASIA_Exp3 and USF_Exp4. AEI was proposed in [11] for enhancing the dynamic characteristics of gait rather than the body shape. It was

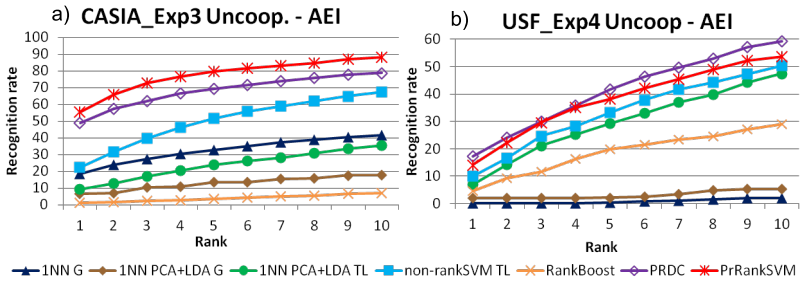

Fig. 12. Results on using AEI as gait representation.

designed to be more robust to carrying condition and clothing changes than GEI, since dynamic characteristics of gait are less affected by those changes. The results in Fig. 12 show that similar observations on the effectiveness of ranking-based transfer learning can be made when a different gait representation is used. Comparing Fig. 12 with Fig. 10, it can be seen that the non-learning based methods (1NN and 1NN PCA+LDA) benefit greatly in CASIA_Exp3 but not in USF_Exp4, whilst the transfer learning methods are less affected by the change of gait representation. This is because CASIA_Exp3 contains clothing and carrying condition changes – what AEI was designed for, whilst USF_Exp4 contains surface and viewpoint changes, which AEI cannot cope with. These results demonstrate the weakness of the existing approaches which address the gait covariate change problem by hand crafting representations, that is, one can never design a representation that works well for any covariate condition changes. In contrast, when using a ranking/distance learning based transfer learning method, one does not need to worry about whether the representation is suitable for the (unknown) covariate conditions one may encounter – just leave the model to do the job.

4.3.2 Experiments under Cooperative Setting

As in the uncooperative experiments we used 50% of people with all their sequences for training, and all the remaining ones for test in the experiments under cooperative setting. The difference is that now the type of sequences (covariate conditions) in gallery and probe are different and a priori known.

Figs. 13a and b show the results for the USF_Exp1 following a cooperative setting. This experiment involves two kinds of sequences (see Table 1): those in which people carry a briefcase (C A L BF $M + N_1$) and those in which they

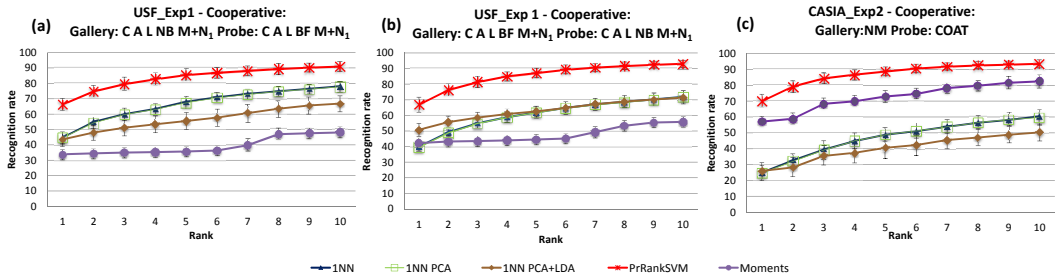

Fig. 13. CMS curves for cooperative experiments: a and b in USF dealing with briefcase covariate, and c in CASIA coping with clothing changes.

do not (C A L NB $M + N_1$). Thus, two different cooperative settings can be considered where both the gallery and probe sets must be composed of only a single type of sequences. The results in Figs. 13a and b show that our rank 1 recognition rates almost double those of the non-ranking methods in both cases. Note that the affine moments-based method does not work even under a cooperative setting here when the silhouettes are noisy.

The results of CASIA_Exp2 using a cooperative setting are depicted in Fig. 13c. Again, our approach gets about 3-fold improvement over the 1NN-based approaches. Note that Moments gets competitive results in this experiment, but still around 10% lower than our result [2]. Under a similar setting, a rank 1 result of 32.7% and 44% are reported by [1,10] respectively, although their experimental setting is still slightly different from ours with larger gallery and probe sets (our learning based method needs to use part of the data for training whilst they do not). Nevertheless, compared with our rank 1 result of 70%, this does give an indication that our model is superior even under cooperative settings.

For cross view recognition, we also reproduced some of the experiments conducted in [4]. In particular, we focused on various combinations of view angles with 90° in the gallery set. Following their experimental settings, only the six NM sequences of each subject were considered and 24 out of 124 subjects were randomly chosen for training leaving all the remaining ones for test. The rank 1 results obtained using our RankSVM model are 5%, 51% and 49% respectively for the three view combinations. These results are comparable with results of the typical SVD-based method [5] (9%,49%, and 52%) but worse

[2] The results of Moments are different from the ones published in [6] because larger gallery and probe sets were used in [6].

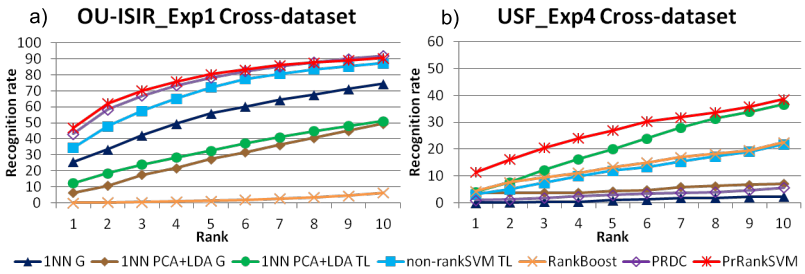

Fig. 14. Cross-dataset experimental results.

than other approaches based on SVR [4] (25%,70%, and 78%). Nevertheless, it is worth pointing out that: those methods are specifically designed for cross-view gait recognition whilst our method can cope with any covariate condition changes and even with a combination of them co-occurring. In addition, we also found from our experiments that similar results can be obtained with the same model but under an uncooperative setting. In contrast, the performance of the methods in [5,4] will decrease under such a setting, because they must first estimate the view angle in the probe, which can only give around 85% accuracy as shown in [15], and then choose one from a set of models learned according to the view estimation.

4.3.3 Cross-dataset Gait Recognition Experiments

So far our ranking model has been learned using different subjects from the same dataset. In the next experiments, our model is learned using one dataset and applied to another one. More specifically, two of the previous experiments already discussed (OU-ISIR_Exp1 and USF_Exp4 from Table 1) were conducted again with identical gallery and probe sets, but this time a different auxiliary dataset, the one used in CASIA_Exp2, was used to learn the transfer learning models. Note that since both the auxiliary dataset and the two target datasets contain clothing changes, there is transferable information to be learned about gait features that are invariant to clothing changes. The objective is thus to compare the ability of different transfer learning models to overcome the dataset bias [46] caused by the differences in the recording environments (e.g. indoor vs. outdoor, natural walking vs. treadmill). The results are shown in Fig. 14. The main findings are: (1) Our RankSVM-based transfer learning model significantly outperforms the non-transfer learning based methods (1NN and 1NN PCA+LDA), indicating that the model can success-

fully learn transferable knowledge even from gait captured from a completely different environment, such as indoors in CASIA and outdoors in USF. This is a significant result as this means that it is not necessary to rely on data collected in the same scene to learn our model. Instead, a large pool of existing labeled gait sequences from other scenes containing a large number of covariate conditions either alone or in combination could be used to learn the model. In a practical sense, our model seems not to need retraining/retuning for a new scene as demonstrates results in Fig. 14, where the same model learned from CASIA (indoors and people walking on a track) works on both OU-ISIR (indoors and people walking on a treadmill) and USF (outdoors and people walking on a track). (2) The RankSVM model also achieves better performance than the alternative transfer learning models in both experiments. In comparison, both RankBoost and PRDC lack consistency when they are applied to different target datasets. Specifically, RankBoost fails completely in OU_ISIR_Exp1 Cross-dataset, whilst it works reasonably well in USF_Exp4 Cross-dataset. PRDC's result is the opposite – fairly close to RankSVM in OU_ISIR_Exp1, but very weak in USF_Exp4. As we discussed before, both models have some unreliabilities – RankBoost relies on the one feature/pixel ranker to select features locally and sequentially, and PRDC employs a numerically unstable iterative learning algorithm that is susceptible to local maximals. These unreliabilities explain their inconsistent performance when applied to different datasets.

5 Conclusions and Future Work

We have proposed a novel gait recognition approach which differs significantly from existing approaches in that the original multi-class classification or identification problem is reformulated into a bipartite ranking problem which learns transferable information independent of the identity of people. In other words, we turn a recognition problem into a verification problem (genuine or imposter) in order to learn features invariant to covariate condition changes that can be generalized to new subjects even in a new dataset. This provides a number of advantages including: (1) unlike most of the existing methods which focus on treating a specific covariate, our approach only needs a single model to cope with any possible covariate condition or even with a combi-

nation of them co-existing; and (2) the model can be learned from different classes/subjects as well as from a different dataset making it more generally applicable with limited data per person in a gallery set (this model can be used even when there is only a single gait sample available for each person in the gallery set). Extensive experiments using three large public datasets have validated the effectiveness of our approach particularly under challenging uncooperative settings.

We have also analyzed the connection between the ranking-based transfer learning methods and relative distance learning-based transfer learning methods. In our context, both models try to achieve the same goal and they differ only in the formulation. In particular, a ranking function is a distance/metric function and a relative learning method also aims to maintain the ranking order in an auxiliary dataset. In the meantime, both models also attempt to quantify gait features to identify the most robust features under different covariate conditions. Our results suggest that a global feature quantification method (e.g. RankSVM, PRDC) is superior to a local one (e.g. RankBoost).

There are a couple of directions for further study: (1) differing from many transfer learning works [21,22,23,24], our current model does not perform model adaption given new data from the target gallery set. This makes the model more vulnerable against dataset bias. Some ideas from the Adaptive SVM [23] can be easily adopted here to make our RankSVM adaptive to new data. In a similar direction, some regularization to the transfer learning process might be included to improve the performance, as some works do in other related areas [47,48]; and (2) we have identified the advantage of a relative distance learning method in terms of modeling the correlation between features explicitly. However, this advantage did not materialize in our experiments due to difficulties in solving the optimization problem. Developing a better optimization solver is part of our ongoing work.

Acknowledgements

This work has partially been supported by projects CICYT TIN2009-14205-C04-04 from the Spanish Ministry of Innovation and Science, PROMETEO

2010-028 from Generalitat Valenciana, and P1-1B2012-22, PREDOC/2008/04 and E-2011-36 from Universitat Jaume I of Castellón.

References

[1] K. Bashir, T. Xiang, S. Gong, Gait recognition without subject cooperation, Pattern Recognition Letters 31 (13) (2010) 2052 – 2060.

[2] S. Singh, K. Biswas, Biometric gait recognition with carrying and clothing variants, LNCS Conf. on Pattern Recognition and Machine Intelligence 5909 (2009) 446–451.

[3] X. Yang, Y. Zhou, T. Zhang, G. Shu, J. Yang, Gait recognition based on dynamic region analysis, Signal Processing 88 (9) (2008) 2350 – 2356.

[4] W. Kusakunniran, Q. Wu, J. Zhang, H. Li, Support vector regression for multi-view gait recognition based on local motion feature selection, in: CVPR, 2010.

[5] Y. Makihara, R. Sagawa, Y. Mukaigawa, T. Echigo, Y. Yagi, Gait recognition using a view transformation model in the frequency domain, in: ECCV, Vol. 3, 2006, pp. 151–163.

[6] Y. Iwashita, K. Uchino, R. Kurazume, Gait-based person identification robust to changes in appearance, Sensors 13 (6) (2013) 7884–7901.

[7] J. Han, B. Bhanu, Individual recognition using gait energy image, PAMI 28 (2) (2006) 316–322.

[8] D. Xu, S. Yan, D. Tao, S. Lin, H. Zhang, Marginal fisher analysis and its variants for human gait recognition and content-based image retrieval, IEEE Trans. on Image Processing 16 (11) (2007) 2811–2821.

[9] D. Tao, X. Li, X. Wu, S. Maybank, General tensor discriminant analysis and gabor features for gait recognition, PAMI 29 (10) (2007) 1700–1715.

[10] S. Yu, D. Tan, T. Tan, A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition, in: ICPR, 2006.

[11] E. Zhang, Y. Zhao, W. Xiong, Active energy image plus 2dlpp for gait recognition, Signal Processing 90 (7) (2010) 2295–2302.

[12] J. Liu, N. Zheng, Gait history image: A novel temporal template for gait recognition, in: IEEE Conf. on Multimedia and Expo, 2007, pp. 663–666.

[13] C.-C. Lee, C.-H. Chuang, J.-W. Hsieh, M.-X. Wu, K.-C. Fan, Frame difference history image for gait recognition, in: ICMLC, Vol. 4, 2011, pp. 1785–1788.

[14] K. Bashir, T. Xiang, S. Gong, Gait representation using flow fields, in: BMVC, 2009.

[15] K. Bashir, T. Xiang, S. Gong, Cross view gait recognition using correlation strength, in: BMVC, 2010.

[16] M. A. Hossain, Y. Makihara, J. Wang, Y. Yagi, Clothing-invariant gait identification using part-based clothing categorization and adaptive weight control, Pattern Recognition 43 (6) (2010) 2281–2291.

[17] D. Damen, D. Hogg, Detecting carried objects in short video sequences, in: ECCV, 2008, pp. 154–167.

[18] K. Sugiura, Y. Makihara, Y. Yagi, Gait identification based on multi-view observations using omnidirectional camera, in: ACCV, 2007, pp. 452–461.

[19] X. Li, S. Lin, S. Yan, D. Xu, Discriminant locally linear embedding with high-order tensor data, IEEE Trans. on Systems, Man, and Cybernetics, Part B 38 (2) (2008) 342–352.

[20] H. Lu, K. Plataniotis, A. Venetsanopoulos, Uncorrelated multilinear discriminant analysis with regularization and aggregation for tensor object recognition, IEEE Trans. on Neural Networks 20 (1) (2009) 103–123.

[21] H. Wang, F. Nie, H. Huang, Transfer dyadic knowledge for cross-domain image classification, in: ICCV, 2011.

[22] S. J. Pan, J. Kwok, Q. Yang, Transfer learning via dimensionality reduction, in: International conference on Artificial intelligence, 2008, pp. 677–682.

[23] J. Yang, R. Yan, A. G. Hauptmann, Cross-domain video concept detection using adaptive svms, in: International conference on Multimedia, ACM, New York, NY, USA, 2007, pp. 188–197.

[24] S. Pan, I. Tsang, J. Kwok, Q. Yang, Domain adaptation via transfer component analysis, IEEE Transactions on Neural Networks 22 (2) (2011) 199 –210.

[25] L. Cao, Z. Liu, T. Huang, Cross-dataset action detection, in: CVPR, 2010.

[26] A. Zweig, D. Weinshall, Exploiting Object Hierarchy: Combining Models from Different Category Levels, in: IEEE International Conference on Computer Vision, 2007.

[27] T. Joachims, Optimizing search engines using clickthrough data, in: Proc. 8th ACM SIGKDD, 2002, pp. 133–142.

[28] B. Prosser, W.-S. Zheng, S. Gong, T. Xiang, Person re-identification by support vector ranking, in: BMVC, 2010, pp. 21.1–21.11.

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

[30] E. Hüllermeier, J. Fürnkranz, W. Cheng, K. Brinker, Label ranking by learning pairwise preferences, Artificial Intelligence 172 (16-17) (2008) 1897–1916.

[31] P. Brazdil, C. Soares, A comparison of ranking methods for classification algorithm selection, in: ECML, 2000, pp. 63–75.

[32] Z. Liu, S. Sarkar, Improved gait recognition by gait dynamics normalization, PAMI 28 (6) (2006) 863–876.

[33] R. Martn-Flez, T. Xiang, Gait Recognition by Ranking, in: Computer Vision - ECCV 2012, Vol. 7572 of Lecture Notes in Computer Science, 2012, pp. 328–341.

[34] O. Chapelle, S. Keerthi, Efficient algorithms for ranking with svms, Information Retrieval 13 (2010) 201–215.

[35] I. Tsochantaridis, T. Joachims, T. Hofmann, A. Y., Large margin methods for structured and interdependent output variables, Journal of Machine Learning Research 6 (2005) 1453–1484.

[36] Y. Liu, An overview of distance metric learning, Tech. rep., Carnegie Mellon University (2007).

[37] M. Schultz, T. Joachims, Learning a distance metric from relative comparisons, in: Advances in Neural Information Processing Systems, 2004.

[38] K. Weinberger, J. Blitzer, L. Saul, Distance metric learning for large margin nearest neighbor classification, in: Advances in Neural Information Processing Systems, 2006.

[39] J. Lee, R. Jin, A. Jain, Rank-based distance metric learning: An application to image retrieval, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008.

[40] W.-S. Zheng, S. Gong, T. Xiang, Person Re-Identification by Probabilistic Relative Distance Comparison, in: 24th Conference on Computer Vision and Pattern Recognition (CVPR), 2011, pp. 649–656.

[41] CASIA, CASIA Gait Database, http://www.sinobiometrics.com (2005).

[42] S. Sarkar, P. Phillips, Z. Liu, I. Vega, P. Grother, K. Bowyer, The HumanId gait challenge problem: data sets, performance and analysis, PAMI 27 (2) (2005) 162–177.

[43] Y. Makihara, H. Mannami, A. Tsuji, M. Hossain, K. Sugiura, A. Mori, Y. Yagi, The ou-isir gait database comprising the treadmill dataset, IPSJ Trans. on Computer Vision and Applications 4 (2012) 53–62.

[44] P. Phillips, H. Moon, S. Rizvi, P. Rauss, The feret evaluation methodology for face-recognition algorithms, PAMI 22 (10) (2000) 1090–1104.

[45] Y. Makihara, Y. Yagi, Cluster-pairwise discriminant analysis, in: ICPR, 2010, pp. 577–580.

[46] A. Torralba, A. Efros, Unbiased look at dataset bias, in: CVPR, 2011, pp. 1521–1528.

[47] S. Si, W. Liu, D. Tao, K.-P. Chan, Distribution calibration in riemannian symmetric space, IEEE Transactions on Systems, Man, and Cybernetics - Part B 41 (4) (2011) 921–930.

[48] X. Tian, D. Tao, Y. Rui, Sparse transfer learning for interactive video search reranking, ACM Trans. Multimedia Comput. Commun. Appl. 8 (3) (2012) 26:1–26:19.