Joint Learning for Attribute-Consistent Person Re-Identification

Sameh Khamis[1(✉)], Cheng-Hao Kuo[2], Vivek K. Singh[3], Vinay D. Shet[4], and Larry S. Davis[1]

[1] University of Maryland, College Park, MD, USA
sameh@umiacs.umd.edu
[2] Amazon.com, Seattle, USA
[3] Siemens Corporate Research, Princeton, USA
[4] Google, Mountain View, USA

Abstract. Person re-identification has recently attracted a lot of attention in the computer vision community. This is in part due to the challenging nature of matching people across cameras with different viewpoints and lighting conditions, as well as across human pose variations. The literature has since devised several approaches to tackle these challenges, but the vast majority of the work has been concerned with appearance-based methods. We propose an approach that goes beyond appearance by integrating a semantic aspect into the model. We jointly learn a discriminative projection to a joint appearance-attribute subspace, effectively leveraging the interaction between attributes and appearance for matching. Our experimental results support our model and demonstrate the performance gain yielded by coupling both tasks. Our results outperform several state-of-the-art methods on VIPeR, a standard re-identification dataset. Finally, we report similar results on a new large-scale dataset we collected and labeled for our task.

1 Introduction

Person re-identification is the problem of matching people across multiple, typically non-overlapping, cameras [8]. Matching is complicated by variations in lighting conditions, camera viewpoints, backgrounds, and human poses. Research in person re-identification has been motivated by increasing safety and security concerns in public spaces, where face recognition and other fine biometric cues are not available because of the insufficient image resolution [9].

Approaches addressing this problem usually focus on either representation, where better descriptors or features are used to specifically address this problem [6–8,10], or learning, where a better similarity or distance function is proposed [1,11,12,20,31]. Our work falls into the latter category. While some recent approaches use one or more standard distance metrics (*e.g.* Euclidean distance or

This work was done while the authors were at Siemens Corporate Research, Princeton, NJ.

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part III, LNCS 8927, pp. 134–146, 2015.
DOI: 10.1007/978-3-319-16199-0_10

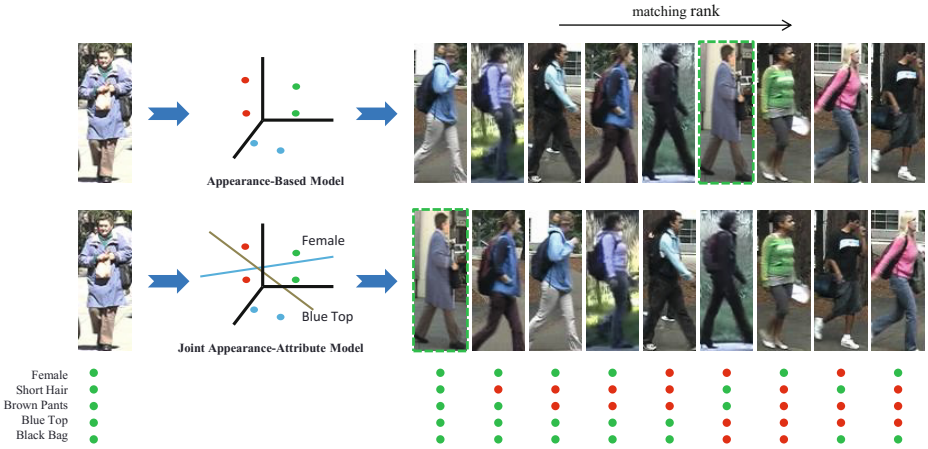

Fig. 1. Overview of our approach. An image of a person of interest on the left (the probe) is used to rank images from a gallery according to how closely they match that person. The correct match, highlighted in a green box, can be difficult even for humans to find given the severe lighting and pose differences between the two images. Similarly, approaches that model only appearance are likely to suffer from these challenges. Our main contribution, on the other hand, is the integration of a semantic aspect, through attributes, into the matching process. We jointly learn a distance metric that optimizes matching and attribute classification by projecting the image descriptors to a coupled appearance-attribute space. Leveraging this representation, our approach gains some invariance to lighting and pose and achieves better performance on the re-identification task.

Bhattacharyya distance) for matching [13,15,21], approaches that instead learn a distance metric for the problem have had better success.

We illustrate our approach to person re-identification in Figure 1. Given an image of a person of interest (the probe), we are interested in a list of subjects images (the gallery), ranked according to how well they match the probe. It can be quite challenging even for a human to find the correct match for the probe image in the figure. Consequently, appearance-based models, represented at the top of the figure, tend to suffer from the severe lighting and pose changes. We instead approach this by augmenting appearance with a semantic attribute-based description of the subject and jointly optimize both ranking and attribute classification. As shown in the figure, the semantic representation in our model imposes an attribute-consistent matching, introducing invariance to the extreme lighting and pose changes, and at the same time the resulting attribute classifiers are better tuned because of the regularization imposed by the matching constraints. We validate both claims empirically. It is noteworthy that our approach is not limited to person re-identification and applies to any matching problem.

We first demonstrate how learning a distance metric optimized over ranking loss, which is a natural aspect of the re-id problem, outperforms one subject to

binary classification constraints as in [19,31]. We learn a distance metric that projects images of the same person closer together than to images of a different person. We then augment the projection subspace using semantic attribute information. Our semantic representation is based on the types of attributes that humans might use in describing appearance (short sleeves, plain shirt, blue jeans, carrying bag, etc.). We jointly optimize for the ranking loss and the attribute classification loss and validate how attribute-consistent matching in this coupled space achieves performance better than several state-of-the-art approaches on VIPeR [9], a standard person re-identification dataset. We also report our results on a new dataset (Indoor-ReID) we collected and labeled for this task.

The rest of this paper is organized as follows. The current literature is surveyed in Section 2. We introduce our ranking formulation and extend it with attribute-consistency in Section 3 and discuss how to efficiently learn a metric over the coupled appearance-attribute space. We then explain our experimental setup and evaluate our approach in Section 4. Finally, we conclude and summarize our work in Section 5.

2 Related Work

Approaches for person re-identification are generally composed of an appearance descriptor to represent the person and a matching function to compare those appearance descriptors. Over the years, several contributions have been made to improve both the representation as well as the matching algorithm in order to increase robustness to the variations in pose, lighting, and background inherent to the problem. Many researchers addressed the problem by proposing better feature representations for the images. Ma *et al.* [18] use local descriptors based on color and gradient information and encode them using high dimensional Fisher vectors. Liu *et al.* [17] use different feature weights for each probe image based on how unique these features are to the subject. Zhao *et al.* [30] use unsupervised saliency estimation and dense patch matching to match subjects, which can even be augmented with a supervised approach.

Several approaches also exploit the prior knowledge of the person geometry or body structure to obtain a pose invariant representation. Farenzena *et al.* [7] accumulate features based on the symmetry of the human body. Gheissari *et al.* [8] match fitted triangulated meshes over segmented images of subjects. Bak *et al.* [2] use body parts detectors with spatial pyramid matching. Similarly, Cheng *et al.* [6] utilize Pictorial Structures (PS) to localize the body parts and match their descriptors. However, these approaches tend to suffer if the pose variations are too extreme, which can invalidate symmetry or break part-based detectors.

Given feature based representations of a pair of images, an intuitive approach is to compute the geodesic distance between the descriptors, for instance, using the Bhattacharyya distance between the histogram-based descriptors or L2-norm between descriptors in a Euclidean space. However some features may be more

relevant for appearance matching than others. To this end, several approaches have been proposed to learn a matching function in a supervised manner from a dataset of image pairs. For instance, Gray *et al.* [10] use boosting to find the best ensemble of localized features for matching. Prosser *et al.* [22] propose ensemble RankSVM to solve person re-identification as a ranking problem, while Wu *et al.* [29] solved a ranking problem subject to a listwise loss instead of a pairwise loss in an effort to realize results closer to what a human would generate.

On the other hand, approaches that learn a distance metric on the feature descriptors have had better success on standard benchmark datasets. Zheng *et al.* [31] learn a metric by maximizing the probability of a true match to have a smaller distance as compared to a wrong match. Köstinger *et al.* [12] learn a Mahalanobis metric that also reflects the properties of log-likelihood ratio test and reports better performance over traditional metric learning. Hirzer *et al.* [11] propose a distance learning approach which is not guaranteed to learn a pseudo-metric, but nonetheless achieves expected performance with a reduced computational cost. Pedagadi *et al.* [20] extract very high dimensional features from the subject images, which then go through multiple stages of unsupervised and supervised dimensionality reduction to estimate a final distance metric. An *et al.* [1] learn a distance metric using Regularized Canonical Correlation Analysis (RCCA) after projecting the features to a kernelized space. Finally, Mignon *et al.* [19] learns a PCA-like projection to a low-dimensional space while preserving pairwise constraints imposed by positive and negative image pairs. We extend the latter's work to a joint optimization framework and learn a projection to a coupled appearance-attribute subspace, and we successfully report a significant performance improvement. The integration of attribute-consistency into the matching process, which is the main thesis of our work, is also applicable to other metric learning approaches.

With the recent success of attribute-based modeling approaches [4,5,14], earlier attempts have also been made to overcome the lighting and pose variations by integrating attributes into the matching process. However, the existing approaches simply augment the extracted feature descriptors with attribute labels obtained independently and thus fail to capture the interactions between the attributes and the identities [15,16,27]. Our work attempts to simultaneously learn matching and attribute classification, and through the coupled process improve the performance of both tasks.

Our approach also does not require access to the attribute labels at test time. In that aspect our work is also related to matching with privileged information. Learning Using Privileged Information (LUPI) is a learning framework where additional information, in our case the attribute labels, is available at training time but is not provided at the test stage [26]. Recent work investigated using attributes in a two stage approach, where the result of the attribute classifiers is used to scale SVM margins in the second stage [24]. We instead integrate both attributes and matching in a single objective function and optimize them jointly.

3 Approach

3.1 Attribute-Consistent Matching

Most work on person re-identification focuses on appearance-based methods, which intuitively suffer from the lighting and pose changes inherent to any matching problem. We instead propose to complement appearance models, which are nonetheless crucial to matching, with a semantic aspect. The semantic representation in our approach is introduced through the integration of attributes into the matching process. We jointly learn a discriminative projection to a joint appearance-attribute subspace. Unlike LMNN [28], this subspace is of lower dimensionality and is discriminatively learned for the purpose of matching [19]. By performing matching in this space, our model exhibits some invariance to the lighting and pose conditions that impede models which rely only on appearance.

We start by introducing our notation. We initially extract a set of feature vectors $\mathbf{x}$ for the subjects in the dataset. We index the features as triplets, where first two vectors $\mathbf{x}_{(i,1)}$ and $\mathbf{x}_{(i,2)}$ corresponds to images of the same subject, while the third vector $\mathbf{x}_{(i,3)}$ corresponds an image of a different subject. We are also given attribute labels $\mathbf{y}$ for the same subjects, where y_{jk} denotes the attribute value of image j for attribute k. To this end, we optimize the joint regularized risk function:

$$\begin{aligned}
&\min_{\mathbf{A},\mathbf{B}} \quad F(\mathbf{A},\mathbf{B}) = \\
&\min_{\mathbf{A},\mathbf{B}} \quad \frac{\lambda_A}{2}\|\mathbf{A}\|_F^2 + \frac{\lambda_B}{2}\|\mathbf{B}\|_F^2 + \\
&\qquad \sum_i \ell(1 + D_{AB}^2(\mathbf{x}_{(i,1)}, \mathbf{x}_{(i,2)}) - D_{AB}^2(\mathbf{x}_{(i,1)}, \mathbf{x}_{(i,3)})) + \\
&\qquad C\sum_j \sum_k \ell(1 - y_{jk}\mathbf{b}_k\mathbf{x}_j), \qquad (1)
\end{aligned}$$

where the two linear mappings $\mathbf{A}$ and $\mathbf{B}$ map input vectors $\mathbf{x}$ into the joint subspace defined by the two projections, and ℓ is a loss function. The projection is learnt so as to satisfy the joint constraint set. The subspace defined by $\mathbf{A}$ only imposes ranking constraints on feature vector triplets; distances between images of the same subject are smaller than those between images of different subjects. The subspace defined by $\mathbf{B}$ includes additional classification constraints that encode the semantic aspect of our model, where each dimension in this subspace represents an attribute, and each row $\mathbf{b}_k$ of the matrix $\mathbf{B}$ is basically a linear classifier for that attribute.

The objective function in Equation 1 is jointly minimizing two loss functions: the ranking loss for the matching and the classification loss for the attribute subspace. Optimizing a ranking loss is arguably more appropriate for person re-identification where ranking is performed at test time. One advantage of this formulation is that both the distance constraints and the attribute classification constraints can be sparse. We explicitly included a regularization term for both matrices to avoid trivial solutions and to achieve a faster convergence rate.

The squared distance between two images in this coupled space can be defined as:

$$D^2_{AB}(x_i, x_j) = \left\| \begin{bmatrix} \mathbf{A} \\ \mathbf{B} \end{bmatrix} (x_i - x_j) \right\|_2^2 = \|\mathbf{A}(x_i - x_j)\|_2^2 + \|\mathbf{B}(x_i - x_j)\|_2^2, \quad (2)$$

which is also the sum of the squared distances in the subspaces defined by the two linear mappings $\mathbf{A}$ and $\mathbf{B}$.

To empirically validate our claim that our attribute-consistent model is more discriminative, we strip our model of the attribute classification constraints. The resulting model is parameterized only by the linear operator $\mathbf{A}$ which projects the input vectors $\mathbf{x}$ to the same dimensionality as our original model in Equation 1. This allows us to isolate the effect of the attribute classification constraints on the matching process. This stripped baseline model is then defined as follows:

$$\min_{\mathbf{A}} G(\mathbf{A}) = \min_{\mathbf{A}} \frac{\lambda}{2}\|\mathbf{A}\|_F^2 + \sum_i \ell(1 + D^2_A(\mathbf{x}_{(i,1)}, \mathbf{x}_{(i,2)}) - D^2_A(\mathbf{x}_{(i,1)}, \mathbf{x}_{(i,3)})) \quad (3)$$

where the distance in the low dimensional space is defined as

$$D^2_A(x_i, x_j) = \|\mathbf{A}(x_i - x_j)\|_2^2. \quad (4)$$

3.2 Optimization

To this point we set the loss function in our experiments to the hinge loss:

$$\ell(x) = \max(0, x), \quad (5)$$

which is convex but not differentiable. There are many smooth approximations of the hinge loss (*e.g.*, the generalized logistic function [19]), but given that we regularize our objective explicitly, convergence rate was not an issue. Under the hinge loss our distance constraints are similar to those in LMNN [28], while our classification constraints correspond to the constraints of a linear SVM. This means that the distance constraints in the objective function are not convex with respect to $\mathbf{A}$ or $\mathbf{B}$. However, an iterative subgradient descent approach on the parameters of both matrices has been shown to converge to good local optima [25,28].

We can now compute a subgradient of the objective with respect to the variables $\mathbf{A}$ and $\mathbf{B}$. A subgradient with respect to $\mathbf{A}$ is:

$$\frac{\partial H(\mathbf{A}, \mathbf{B})}{\partial \mathbf{A}} = \lambda_A \mathbf{A} + 2\mathbf{A} \sum_{i \in I(\mathbf{A}, \mathbf{B})} (\mathbf{C}_{(i,1),(i,2)} - \mathbf{C}_{(i,1),(i,3)}) \quad (6)$$

where the set I is the subset of triplets $\mathcal{T}$ that violate the distance constraints and is defined formally as:

$$I(\mathbf{A},\mathbf{B}) = \big\{i \in \mathcal{T} : D^2_{AB}(\mathbf{x}_{(i,1)},\mathbf{x}_{(i,3)}) - D^2_{AB}(\mathbf{x}_{(i,1)},\mathbf{x}_{(i,2)}) < 1\big\}, \tag{7}$$

and $\mathbf{C}_{i,j}$ is the outer product matrix for the difference between two feature vectors $\mathbf{x}_i$ and $\mathbf{x}_j$:

$$\mathbf{C}_{i,j} = (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T. \tag{8}$$

Similarly, a subgradient with respect to row $\mathbf{b}_k$ in $\mathbf{B}$ is:

$$\frac{\partial H(\mathbf{A},\mathbf{B})}{\partial \mathbf{b}_k} = \lambda_B \mathbf{b}_k + 2C \sum_{j \in J_k(\mathbf{B})} y_{jk}\mathbf{x}_j^T + 2\mathbf{b}_k \sum_{i \in I} (\mathbf{C}_{(i,1),(i,2)} - \mathbf{C}_{(i,1),(i,3)}) \tag{9}$$

where the set J_k is the subset of feature vector indices that are misclassified by attribute classifier k, which is represented by row $\mathbf{b}_k$:

$$J_k(\mathbf{B}) = \big\{j \in \mathcal{L} : y_{jk}\mathbf{b}_k^T\mathbf{x}_j < 1\big\} \tag{10}$$

To learn the linear mappings we then use a projected subgradient descent algorithm [23]. The iterative projections to the constrained Frobenius norms dramatically sped up the convergence rate for the learning procedure. We also employ restarts to avoid local minima. The details of the approach are provided in Algorithm 1.

4 Experiments

4.1 Setup

We evaluated our model on VIPeR [9], a standard person re-identification dataset, as well as the new dataset, Indoor-ReID, which we collected and labeled. VIPeR contains 632 images of 316 subjects captured from 2 cameras. The images are captured outdoors and have significant lighting and viewpoint variations. We use the 15 binary attributes annotated by [15]. The dataset is split randomly into a training set and a testing set. In one set of experiments the splits are of equal sizes (316 subjects each) and in another set the training set has only 100 subjects and the test set has 532 subjects. Testing is done by considering images from Camera A as probe and evaluating their matches from images in Camera B. All the benchmarking results are averages of 10 runs.

Indoor-ReID was collected in an indoor office environment, using 4 different cameras at varying angle and under different lighting conditions. It contains over 28,000 images from 290 different subjects. The images were generated by sampling images from several trajectories obtained over a few hours of surveillance

Algorithm 1. Learning Attribute-Consistent Matching

INPUT: $\mathbf{x}, \mathbf{y}, \mathcal{T}, \mathcal{L}, \lambda_A, \lambda_B, C, T$
Initialize $\mathbf{A}_1$ and $\mathbf{B}_1$ randomly
for $t = 1 \dots T$ **do**
  Find the violating triplets $I(\mathbf{A}_t, \mathbf{B}_t)$ (Equation 7)
  Calculate $\frac{\partial H(\mathbf{A}_t, \mathbf{B}_t)}{\partial \mathbf{A}_t}$ (Equation 6)
  Set $\eta_A = \frac{1}{\lambda_A t}$
  Set $\mathbf{A}_{t+\frac{1}{2}} = (1 - \eta_A \lambda_A)\mathbf{A}_t + \eta_A \frac{\partial H(\mathbf{A}_t, \mathbf{B}_t)}{\partial \mathbf{A}_t}$
  Set $\mathbf{A}_{t+1} = \min\left\{1, \frac{1}{\sqrt{\lambda_A}\|\mathbf{A}\|_F}\right\} \mathbf{A}_{t+\frac{1}{2}}$
  Find the violating indices $J_k(\mathbf{A}_t, \mathbf{B}_t)$ for each k (Equation 10)
  Calculate $\frac{\partial H(\mathbf{A}_t, \mathbf{B}_t)}{\partial \mathbf{B}_t}$ (Equation 9 for each k)
  Set $\eta_B = \frac{1}{\lambda_B t}$
  Set $\mathbf{B}_{t+\frac{1}{2}} = (1 - \eta_B \lambda_B)\mathbf{B}_t + \eta_B \frac{\partial H(\mathbf{A}_t, \mathbf{B}_t)}{\partial \mathbf{B}_t}$
  Set $\mathbf{B}_{t+1} = \min\left\{1, \frac{1}{\sqrt{\lambda_B}\|\mathbf{B}\|_F}\right\} \mathbf{B}_{t+\frac{1}{2}}$
end for
OUTPUT: $\mathbf{A}_{T+1}$ and $\mathbf{B}_{T+1}$

data. Since a subject may appear several times over different tracks, we manually annotated the identities across tracks. We also annotated 16 attributes for each subject, which include 10 attributes for attire description (sleeve length, pants length, hair length, top color, pants color, hair color, top pattern, hat, facial hair, glasses), 3 attributes for non-attire description (gender, build, complexion), and 3 attributes for carry-on objects (bag, backpack, handheld). Some of the collected attributes are multivalued, *e.g.* color is chosen to be the dominant color and is selected out of the 11 universal color names in Berlin and Kay [3]. Figure 2 illustrates some samples from our dataset. To evaluate our approach we split the dataset into a training set and a testing set of equal size with almost identical attribute distributions. At test time we calculate the distances between two tracks as the distance between two randomly sampled representative images, one from each track. The same sampled images are used for all the benchmarks to ensure a fair comparison. We also average the results for both setups over 10 runs.

We extract the same features used in recent benchmarks [1,11,20] for both VIPeR and Indoor-ReID. We divide each image into overlapping patches of size 8×8 with 50% overlap in both directions, which results in 341 patches. From each patch we collect 8-bin color histograms for YUV and HSV color spaces, and additionally LBP histograms using a grayscale representation of the image. We then concatenate each feature type separately for all the patches in each image as in [20] and proceed to perform unsupervised dimensionality reduction using PCA to project the three feature sets to 100, 20, and 40 dimensions respectively. Our approach would then discriminatively project this down to an even lower dimensional subspace of only 50 dimensions.

Fig. 2. Sample images from our dataset (Indoor-ReID). The dataset contains over 28,000 images of over 290 different subjects captured under diverse lighting and viewpoint conditions. The dataset is also annotated with 16 different attributes for each subject.

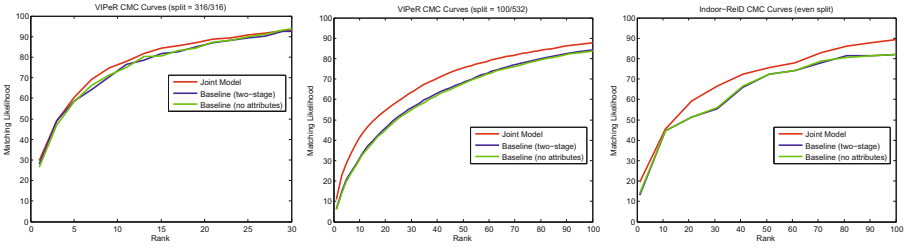

Fig. 3. CMC Curves for VIPeR and Indoor-ReID. The curve for the joint model using the coupled appearance-attribute space dominates the two baseline curves in all graphs. The first graph for VIPeR was created with an even training/test split (316/316), while the second graph figure was created with a split of 100 training subjects and 532 test subjects. For Indoor-ReID the dataset was split evenly.

To evaluate the contribution of the coupling, we report the results against two baselines. The first baseline represents the stripped ranking formulation in Equation 3, where the data is basically just projected to a lower dimensional subspace. As a second baseline we report the results using a two-stage approach, where we augment our extracted features using the output of separately trained attribute classifiers before optimizing the ranking formulation. This baseline validates that the performance gain in the joint model is due to using a coupled appearance-attribute subspace and not just due to utilizing attribute information. We finally report our attribute classification accuracies on both datasets using the 50/50 split. On Indoor-ReID the multi-class attributes were expanded to n-binary valued attributes for the training, and the predicted attribute value at test time is the one with the highest score. We finally compare the attribute classification results on VIPeR to those reported in [15]. The model parameters λ_A, λ_B, and C are set by cross validation to 10^{-2}, 10^{-2}, 10^{-3} respectively for all our experiments.

Table 1. Re-identification quantitative results on VIPeR using two different splits. The numbers are the percentage of correct matches at rank r, *i.e.* in the top ranked r images. In our results we note specifically that augmenting the subspace with the attributes in the joint model significantly improved the results, given a fixed subspace dimensionality.

VIPeR Ranks (split = 316/316)				
Approach	r=1	r=5	r=10	r=20
Joint Model	29.54	60.34	75.95	87.34
Baseline (two-stage)	27.85	58.65	73.42	86.92
Baseline (no attributes)	26.58	58.23	73.00	85.65
RCCA [1]	30.00	-	75.00	87.00
RPLM [11]	27.00	-	69.00	83.00
sLDFV [18]	26.53	56.38	70.88	84.63
eSDC [30]	26.74	50.70	62.37	76.36
LFDA [20]	24.18	-	67.12	-
CPS [6]	21.84	44.64	57.21	71.23
PCCA (χ^2_{RBF}) [19]	19.27	48.89	64.91	80.28
SDLAF+AIR [15]	17.40	39.04	50.84	67.27

VIPeR Ranks (split = 100/532)				
Approach	r=1	r=5	r=10	r=20
Joint Model	11.05	28.91	41.30	54.83
Baseline (two-stage)	6.35	20.81	31.24	46.17
Baseline (no attributes)	5.94	19.89	30.60	45.15
PCCA (χ^2_{RFB}) [19]	9.27	24.89	37.43	52.89
PRDC [31]	9.12	24.19	34.40	48.55

4.2 Results

The Cumulative Matching Characteristic (CMC) curve has been adopted as the standard metric for evaluation for person re-identification systems. The curve illustrates the likelihood of the correct match being in the top r ranked images for each rank r. Our CMC curves for both VIPeR and Indoor-ReID are shown in Figure 3. The first graph for the VIPeR dataset uses the even training/test split (316/316) while the second graph uses the more challenging 100/532 split. The curve for our joint model dominates the two baseline curves, more clearly on the bottom figure, demonstrating the performance gain that the coupling achieves. Similarly, the third graph shows the CMC curves for the new Indoor-ReID dataset, and the joint model with the coupled subspace is also clearly dominating the two baseline curves.

We also report the numbers for comparison in Tables 1 and 2. Our joint model achieves the highest accuracies across all reported ranks. Using better feature descriptors is likely to even increase this gain, as can be seen from the two reported performances for ITML ([31] and [12]). We outperform PCCA [19] using the same kernel (sqrt), number of negative examples (10), and same subspace

Table 2. Re-identification quantitative results on Indoor-ReID. The numbers are the percentage of correct matches at rank r, *i.e.* in the top ranked r images. The joint model projecting to the coupled space significantly outperformed the baseline model, given a fixed subspace dimensionality.

Indoor-ReID Ranks (even split)				
Approach	r=1	r=5	r=10	r=20
Joint Model	19.51	35.77	45.53	59.35
Baseline (two-stage)	13.82	34.15	43.90	50.41
Baseline (no attributes)	13.01	34.15	44.72	50.41

Table 3. Attribute classification results on VIPeR and Indoor-ReID. Our classification accuracies on VIPeR are higher than those of AIR [15] on almost all attributes.

Approach	AIR [15]	Ours
Shorts	79	**88.8**
Sandals	64	**93.3**
Backpacks	**66**	64.2
Jeans	**76**	69.3
Carrying	**75**	71.0
Logo	59	**78.7**
V-neck	44	**91.6**
Open-outer	64	**76.6**
Stripes	41	**92.3**
Sunglasses	66	**76.2**
Headphones	74	**97.5**
Shorthair	**52**	50.0
Longhair	65	**66.9**
Male	**68**	49.3
Skirt	67	**95.2**
Average	64	**77.4**

Approach	Random	Ours
Sleeve Length	50.0	62.7
Pants Length	50.0	86.9
Hair Length	33.3	78.6
Hat	50.0	96.1
Top Color	9.1	22.3
Pants Color	9.1	38.0
Hair Color	9.1	47.5
Top Pattern	25.0	75.1
Bag	50.0	67.0
Backpack	50.0	91.2
Handheld	50.0	58.9
Glasses	50.0	58.2
Gender	50.0	74.0
Facial Hair	50.0	96.4
Build	33.3	62.4
Complexion	33.3	57.8
Average	37.6	67.1

dimensionality (30). Our results also demonstrate that integrating the semantic aspect by coupling attribute classification and matching significantly improved the performance across all experiments. This effect is even more pronounced in the second set of experiments on VIPeR. Similarly, on Indoor-ReID the joint model projecting to the coupled space significantly outperformed the baseline model.

We finally quantify our attribute classification results on both datasets in Table 3. For VIPeR we compare with the reported numbers from Layne *et al*. [15]. We achieve better accuracies for most attributes using the simple linear classifiers in our model. We also report our attribute classification accuracies on Indoor-ReID. Since some of the labeled attributes for Indoor-ReID are multivalued, we also report the random chance performance. During training we

expanded the multivalued attributes to n-binary attributes, and at test time we predict the attribute value with the largest score.

5 Conclusion

We presented a joint learning framework for attribute-consistent matching. We integrate semantic attributes and person re-identification by projecting the input images to lower dimensional coupled appearance-attribute subspace. Matching in this subspace exhibits some invariance to the severe lighting conditions and pose variations that hinder appearance-based matching models. We evaluated our model on VIPeR, a standard benchmark dataset for person re-identification, as well as a new large scale dataset we collected and annotated with attributes relevant to the problem. We report results that outperform several state-of-the-art methods on VIPeR and demonstrate on both datasets that the performance gain by the joint model improves over the baselines and over prior art using the same input features.

Acknowledgments. This research was supported by contract N00014-13-C-0164 from the Office of Naval Research through a subcontract from United Technologies Research Center, and by a grant from Siemens Corporate Research.

References

1. An, L., Kafai, M., Yang, S., Bhanu, B.: Reference-based person re-identification. In: AVSS (2013)
2. Bak, S., Corvee, E., Bremond, F., Thonnat, M.: Person re-identification using spatial covariance regions of human body parts. In: AVSS (2010)
3. Berlin, B., Kay, P.: Basic color terms: Their universality and evolution. University of California, Berkeley (1969)
4. Bourdev, L.D., Maji, S., Malik, J.: Describing people: A poselet-based approach to attribute classification. In: ICCV (2011)
5. Chen, H., Gallagher, A., Girod, B.: Describing clothing by semantic attributes. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part III. LNCS, vol. 7574, pp. 609–623. Springer, Heidelberg (2012)
6. Cheng, D.S., Cristani, M., Stoppa, M., Bazzani, L., Murino, V.: Custom pictorial structures for re-identification. In: BMVC (2011)
7. Farenzena, M., Bazzani, L., Perina1, A., Murino, V., Cristani, M.: Person re-identification by symmetry-driven accumulation of local features. In: CVPR (2010)
8. Gheissari, N., Sebastian, T.B., Tu, P.H., Rittscher, J.: Person reidentification using spatiotemporal appearance. In: CVPR (2006)
9. Gray, D., Brennan, S., Tao, H.: Evaluating appearance models for recognition, reacquisition, and tracking. In: PETS (2007)
10. Gray, D., Tao, H.: Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 262–275. Springer, Heidelberg (2008)

11. Hirzer, M., Roth, P.M., Köstinger, M., Bischof, H.: Relaxed pairwise learned metric for person re-identification. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part VI. LNCS, vol. 7577, pp. 780–793. Springer, Heidelberg (2012)
12. Köstinger, M., Hirzer, M., Wohlhart, P., Roth, P.M., Bischof, H.: Large scale metric learning from equivalence constraints. In: CVPR (2012)
13. Kuo, C.H., Khamis, S., Shet, V.: Person re-identification using semantic color names and rankboost. In: WACV (2013)
14. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: CVPR (2009)
15. Layne, R., Hospedales, T.M., Gong, S.: Person re-identification by attributes. In: BMVC (2012)
16. Layne, R., Hospedales, T.M., Gong, S.: Towards person identification and re-identification with attributes. In: Fusiello, A., Murino, V., Cucchiara, R. (eds.) ECCV 2012 Ws/Demos, Part I. LNCS, vol. 7583, pp. 402–412. Springer, Heidelberg (2012)
17. Liu, C., Gong, S., Loy, C.C., Lin, X.: Person re-identification: What features are important? In: Fusiello, A., Murino, V., Cucchiara, R. (eds.) ECCV 2012 Ws/Demos, Part I. LNCS, vol. 7583, pp. 391–401. Springer, Heidelberg (2012)
18. Ma, B., Su, Y., Jurie, F.: Local descriptors encoded by fisher vectors for person re-identification. In: Fusiello, A., Murino, V., Cucchiara, R. (eds.) ECCV 2012 Ws/Demos, Part I. LNCS, vol. 7583, pp. 413–422. Springer, Heidelberg (2012)
19. Mignon, A., Jurie, F.: Pcca: A new approach for distance learning from sparse pairwise constraints. In: CVPR (2012)
20. Pedagadi, S., Orwell, J., Velastin, S., Boghossian, B.: Local fisher discriminant analysis for pedestrian re-identification. In: CVPR (2013)
21. Prosser, B., Gong, S., Xiang, T.: Multi-camera matching using bi-directional cumulative brightness transfer functions. In: BMVC (2008)
22. Prosser, B., Zheng, W.S., Gong, S., Xiang, T.: Person re-identification by support vector ranking. In: BMVC (2010)
23. Shalev-Shwartz, S., Singer, Y., Srebro, N., Cotter, A.: Pegasos: Primal estimated sub-gradient solver for SVM. Mathematical Programming, Series B **127**(1), 3–30 (2011)
24. Sharmanska, V., Quadrianto, N., Lampert, C.H.: Learning to rank using privileged information. In: ICCV (2013)
25. Torresani, L., Lee, K.: Large margin component analysis. In: NIPS (2007)
26. Vapnik, V., Vashist, A.: A new learning paradigm: Learning using privileged information. In: IJCNN (2009)
27. Vaquero, D.A., Feris, R.S., Tran, D., Brown, L.M.G., Hampapur, A., Turk, M.: Attribute-based people search in surveillance environments. In: WACV (2009)
28. Weinberger, K.Q., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. JMLR (2009)
29. Wu, Y., Mukunoki, M., Funatomi, T., Minoh, M., Lao, S.: Optimizing mean reciprocal rank for person re-identification. In: AVSS (2011)
30. Zhao, R., Ouyang, W., Wang, X.: Unsupervised salience learning for person re-identification. In: CVPR (2013)
31. Zheng, W.S., Gong, S., Xiang, T.: Person re-identification by probabilistic relative distance comparison. In: CVPR (2011)