Learning Discriminative Appearance-Based Models Using Partial Least Squares

William Robson Schwartz
schwartz@cs.umd.edu

Larry S. Davis
lsd@cs.umd.edu
University of Maryland, A.V.Williams Building, College Park, MD 20742, USA

Abstract

Appearance information is essential for applications such as tracking and people recognition. One of the main problems of using appearance-based discriminative models is the ambiguities among classes when the number of persons being considered increases. To reduce the amount of ambiguity, we propose the use of a rich set of feature descriptors based on color, textures and edges. Another issue regarding appearance modeling is the limited number of training samples available for each appearance. The discriminative models are created using a powerful statistical tool called Partial Least Squares (PLS), responsible for weighting the features according to their discriminative power for each different appearance. The experimental results, based on appearance-based person recognition, demonstrate that the use of an enriched feature set analyzed by PLS reduces the ambiguity among different appearances and provides higher recognition rates when compared to other machine learning techniques.

1 Introduction

Appearance-based person recognition has widespread applications such as tracking and person identification and verification. However, the nature of the input data poses great challenges due to variations in illumination, shadows, and pose, as well as frequent inter- and intra-person occlusion. Under these conditions, the use of a single feature channel, such as color-based features, may not be powerful enough to capture subtle differences between different people's appearances. Therefore, additional cues need to be exploited and combined to improve discriminability of appearance-based models.

In general, human appearances are modeled using color-based features such as color histograms [4]. Spatial information can be added by representing appearances in joint color spatial spaces [6]. Also, appearance models of individuals based on nonparametric kernel density estimation have been used [11]. Other representations include spatial-temporal appearance modeling [8] and part-based appearance modeling [10].

Previous studies [12, 17, 18, 20, 22] have shown that significant improvements can be achieved using different types (or combinations) of low-level features. A strong set of features provides high discriminatory power, reducing the need for complex classification methods. Therefore, we augment color-based features with other discriminative cues. We exploit features based on textures and edges, obtaining a richer feature descriptor set as result.

To detect subtle differences between appearances, it is useful to perform a dense sampling for each feature channel, as will be shown on the experiments. However, as a result, the dimensionality of the feature space increases considerably (a feature vector describing an appearance is composed of more than 25,000 features).

Once discriminative appearance-based models have been built, machine learning methods need to be applied so that new samples of the appearances can be correctly classified during a testing stage. Learning methods such as support vector machines (SVM) [2], k-neareast neighbors combined with SVM [21], decision trees [1], learning discriminative distance metrics [11] have been exploited. However, since feature augmentation results in a high dimensional feature space, these machine learning methods may not always be used directly due to high computational requirements and low performance, as we show in the experimental results. The dimensionality of the data needs to be reduced first.

The high dimensionality, the very small number of samples available to learn each appearance and the presence of multicollinearity among the features due to the dense sampling make an ideal setting for a statistical technique known as Partial Least Squares (PLS) regression [19]. PLS is a class of methods for modeling relations between sets of observations by means of latent variables. Although originally

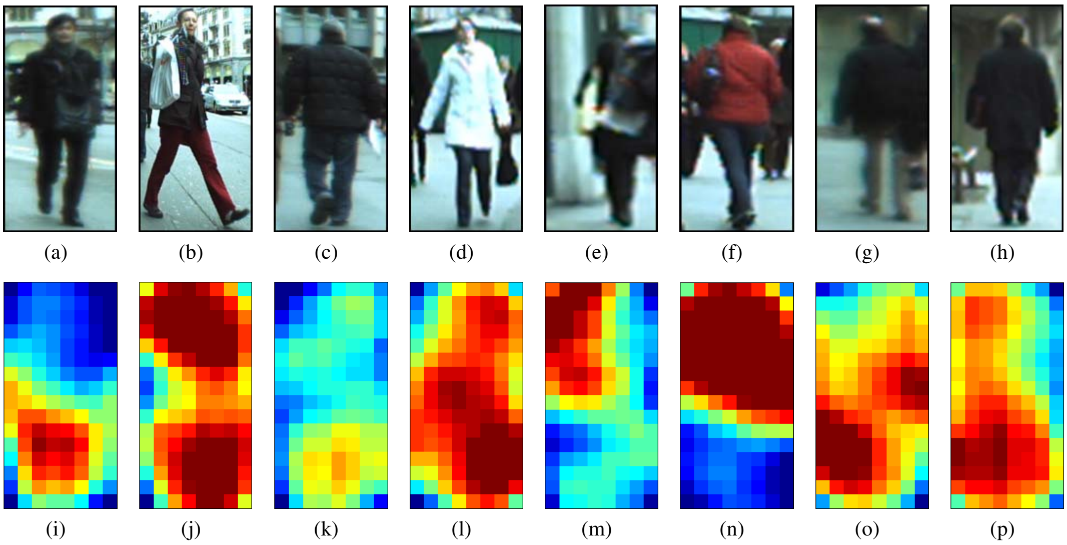


Figure 1. Spatial distribution of weights of the discriminative appearance-based models considering eight people extracted from video sequence #0 of the ETHZ dataset. The first row shows the appearance of each person and the second row the weights estimated by PLS for the corresponding appearance. Models are learned using the proposed method combining color, texture and edge features. PLS is used to reduce the dimensionality and the weights of the first projection vector are shown as the average of the feature weights in each block. Red indicates high weights, blue low.

proposed as a regression technique, PLS can be also be used as a class aware dimensionality reduction tool. This is in contrast to the commonly used Principal Component Analysis (PCA), which does not consider class discrimination during dimensionality reduction.

The projection vectors estimated by PLS provide information regarding the importance of features as a function of location. Since PLS is a class-aware dimensionality reduction technique, the importance of features in a given location is related to the discriminability between appearances. For example, Figure 1 shows the spatial distribution of the weights of the first projection vector when PLS is used to combine the three feature channels. High weights are located in regions that better distinguish a specific appearance from the remaining ones. For example, blacks regions of the homogeneous jackets are not given high weights, since several people wear black jackets. However, the regions where the white and red jackets are located obtain high weights due to their unique appearances.

In this work we exploit a rich feature set analyzed by PLS using an one-against-all scheme [13] to learn discriminative appearance-based models. The dimensionality of the feature space is reduced by PLS and then a simple classification method is applied for each model using the resulting latent variables. This classifier is used during the testing stage to classify new samples. Experimental results based on appearance-based person recognition demonstrate that the feature augmentation provides better results than models based on a single feature channel. Additionally, experiments show that the proposed approach outperforms results obtained by techniques such as SVM and PCA.

2 Proposed Method

In this section we describe the method used to learn the appearance models. The combination of a strong feature set and dimensionality reduction is based on our previous work developed for the purpose of pedestrian detection [16]. The features used are described in section 2.1 and an overview of partial least squares is presented in section 2.2. Finally, section 2.3 describes the learning stage of the discriminative appearance-based models.

2.1 Feature Extraction

In the learning stage, only one exemplar is provided for each appearance i in the form of an image window. This window is decomposed into overlapping blocks and a set of features is extracted for each block to construct a feature vector. Therefore, for each appearance i, we obtain one sample described by a high dimensional feature vector $\boldsymbol{v}_i$.

To capture texture we extract features from co-occurrence matrices [9], a method widely used for texture analysis. Co-occurrence matrices represent second order texture information - i.e., the joint probability distribution of gray-level pairs of neighboring pixels in a block. We use 12 descriptors: angular second-moment, contrast, correlation, variance, inverse difference moment, sum average, sum variance, sum entropy, entropy, difference variance, difference entropy, and directionality [9]. Co-occurrence features are useful in human detection since they provide information regarding homogeneity and directionality of patches. In general, a person wears clothing composed of homogeneous textured regions and there is a significant difference between the regularity of clothing texture and background textures.

Edge information is captured using histograms of oriented gradients (HOG) [5]. This method captures edge or gradient structures that are characteristic of local shape. Since the histograms are computed for regions of a given size within a window, HOG is robust to some location variability of body parts. HOG is also invariant to rotations smaller than the orientation bin size.

The last type of information captured is color. In order to incorporate color we use color histograms computed for blocks. To avoid artifacts obtained by monotonic transformation in color and linear illumination changes, before calculating the histogram the value of pixels within a block are transformed to the relative ranks of intensities for each color channel R, G and B, similarly to [11]. Finally, each histogram is normalized to have unit L_2 norm.

Once the feature extraction process is performed for all blocks inside an image window, features are concatenated creating a high dimensional feature vector $\boldsymbol{v}_i$.

2.2 Partial Least Squares for Dimension Reduction

Partial least squares is a method for modeling relations between sets of observed variables by means of latent variables. The basic idea of PLS is to construct new predictor variables, latent variables, as linear combinations of the original variables summarized in a matrix $\boldsymbol{X}$ of descriptor variables (features) and a vector $\boldsymbol{y}$ of response variables (class labels). While additional details regarding PLS methods can be found in [15], a brief mathematical description of the procedure is provided below.

Let $\mathcal{X} \subset \mathbb{R}^m$ denote a m-dimensional space of feature vectors and similarly let $\mathcal{Y} \subset \mathbb{R}$ be a 1-dimensional space representing the class labels. Let the number of samples be n. PLS decomposes the zero-mean matrix $\boldsymbol{X}$ $(n \times m)$ and zero-mean vector $\boldsymbol{y}$ $(n \times 1)$ into

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^T + \boldsymbol{E}$$
$$\boldsymbol{y} = \boldsymbol{U}\boldsymbol{q}^T + \boldsymbol{f}$$

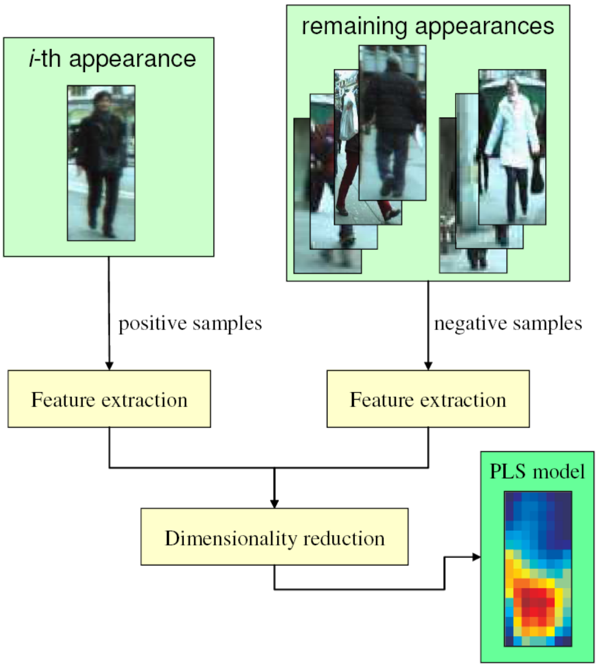


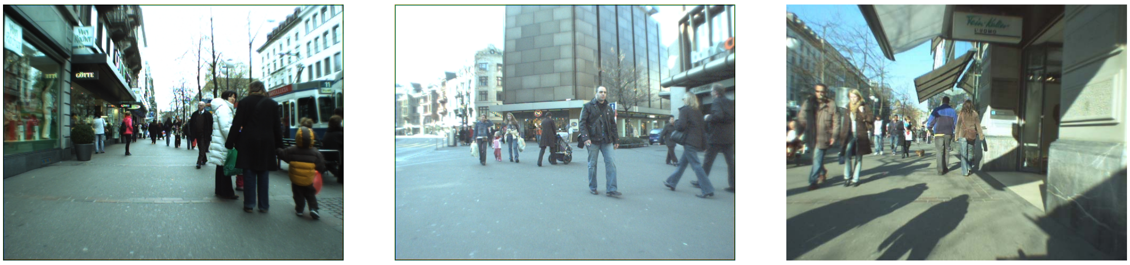
Figure 2. Proposed method. For each appearance represented by an image window, features are extracted and PLS is applied to reduce dimensionality using a one-against-all scheme. Afterwards, a simple classifier is used to match new samples to models learned.

where $\boldsymbol{T}$ and $\boldsymbol{U}$ are $n \times p$ matrices containing p extracted latent vectors, the $(m \times p)$ matrix $\boldsymbol{P}$ and the $(1 \times p)$ vector $\boldsymbol{q}$ represent the loadings and the $n \times m$ matrix $\boldsymbol{E}$ and the $n \times 1$ vector $\boldsymbol{f}$ are the residuals. The PLS method, using the nonlinear iterative partial least squares (NIPALS) algorithm [19], constructs a latent subspace composed of a set of weight vectors (or projection vectors) $W = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \dots \boldsymbol{w}_p\}$ such that

$$[cov(\boldsymbol{t}_i, \boldsymbol{u}_i)]^2 = \max_{|\boldsymbol{w}_i|=1} [cov(\boldsymbol{X}\boldsymbol{w}_i, \boldsymbol{y})]^2$$

where $\boldsymbol{t}_i$ is the i-th column of matrix $\boldsymbol{T}$, $\boldsymbol{u}_i$ the i-th column of matrix $\boldsymbol{U}$ and $cov(\boldsymbol{t}_i, \boldsymbol{u}_i)$ is the sample covariance between latent vectors $\boldsymbol{t}_i$ and $\boldsymbol{u}_i$. After the extraction of the latent vectors $\boldsymbol{t}_i$ and $\boldsymbol{u}_i$, the matrix $\boldsymbol{X}$ and vector $\boldsymbol{y}$ are deflated by subtracting their rank-one approximations based on $\boldsymbol{t}_i$ and $\boldsymbol{u}_i$. This process is repeated until the desired number of weight vectors had been extracted.

The dimensionality reduction is performed by projecting a feature vector $\boldsymbol{v_i}$ onto the weight vectors $W = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \dots \boldsymbol{w}_p\}$, obtaining the latent vector $\boldsymbol{z_i}$ $(1 \times p)$ as a result. This latent vector is used in the classification.



(a) sequence #1 (b) sequence #2 (c) sequence #3

Figure 3. Samples of the video sequences used in the experiments. (a) sequence #1 is composed of 1,000 frames with 83 different people; (b) sequence #2 is composed of 451 frames with 35 people; (c) sequence #3 is composed of 354 frames containing 28 people.

Similarly to PCA, in the dimensionality reduction using PLS, after relevant weight vectors are extracted, an appropriate classifier can be applied in the low dimensional subspace. The difference between PLS and PCA is that the former creates orthogonal weight vectors by maximizing the covariance between elements in $\boldsymbol{X}$ and $\boldsymbol{y}$. Thus, PLS not only considers the variance of the samples but also considers the class labels.

2.3 Learning Appearance-Based Models

The procedure to learn the discriminative appearance-based models for a training set $t = \{\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_k\}$, where $\boldsymbol{u}_i$ represents a subset of exemplars of each person (appearance) to be considered, is illustrated in Figure 2 and described in details as follows. Each subset $\boldsymbol{u}_i$ is composed of feature vectors extracted from image windows containing examples of the i-th appearance.

In this work we exploit one-against-all scheme to learn a PLS discriminatory model for each person. Therefore, when the i-th person is considered, the remaining samples $t \setminus \boldsymbol{u}_i$ are used as counter-examples of the i-th person.

For the one-against-all scheme, PLS gives higher weights to features located in regions containing discriminatory characteristics, as shown in Figure 1. Therefore, this process can be seen as a feature selection process depending on the feature type and the location.

Once the PLS model has been estimated for the i-th appearance, the feature vectors describing this appearance are projected onto the weight vectors. The resulting low-dimensional features are used during the testing stage to match a query samples.

When a sample is presented during the testing stage, its feature vector is projected onto the latent subspace estimated previously for each one of the k appearances and has its Euclidean distance to the samples used in training are computed. Then, this sample is classified as belonging to the appearance with the smallest Euclidean distance.

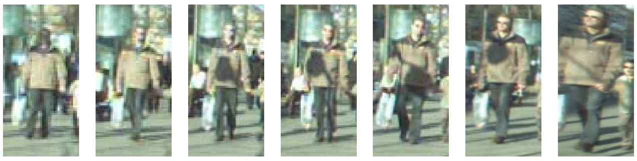


Figure 4. Samples of a person's appearance in different frames of a video sequence belonging to ETHZ dataset.

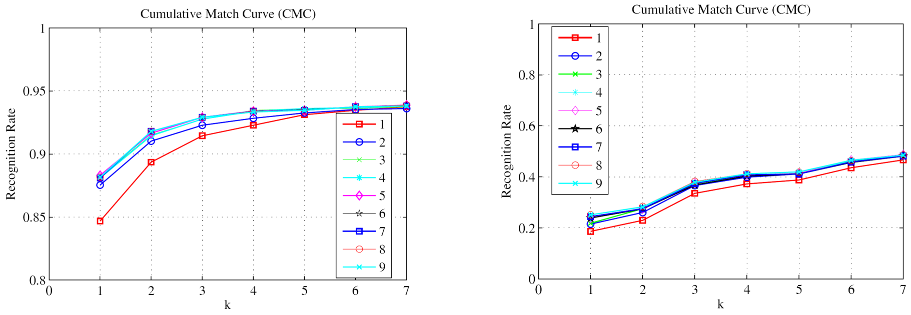
3 Experimental Results

In this section we present experiments to evaluate our approach. Initially, we describe the parameter settings and the dataset used. Then, we evaluate several aspects of our method, such as the improvement provided by using a richer feature set, the reduction in computational cost and improvement in performance compared to PCA and SVM.

Dataset. To obtain a large number of different people captured in uncontrolled conditions, we choose the ETHZ dataset [7] to perform our experiments. This dataset, originally used for human detection, is composed of four video sequences, where the first (sequence #0) is used to estimate parameters and the remaining three sequences are used for testing. Samples of testing sequence frames are shown in Figure 3.

The ETHZ dataset presents the desirable characteristic of being captured from moving cameras. This camera setup provides a range of variations in people's appearances. Figure 4 shows a few samples of a person's appearance extracted from different frames. Changes in pose and illumination conditions take place and due to the fact that the appearance model is learned from a single sample, a strong set of features becomes important to achieve robust appearance matching during the testing stage.

To evaluate our approach, we used the ground truth in-



(a) PLS

(b) PCA

Figure 5. Recognition rate as a function of the number of factors (plots are shown in different scales to better visualization).

formation regarding people's locations to extracted samples from each video (considering only people with size higher than 60 pixels). Therefore, a set of samples is available for each different person in the video. The learning procedure presented in Section 2.3 is executed using one sample chosen randomly per person. Afterwards, the evaluation (appearance matching) considers the remaining samples.

Experimental Setup. To obtain the experimental results we have considered windows of 32×64 pixels. Therefore, either to learn or match an appearance, we rescale the person size to fit into a 32×64 window.

For co-occurrence feature extraction we use block sizes of 16×16 and 32×32 with shifts of 8 and 16 pixels, respectively, resulting in 70 blocks per detection window for each color band. We work in the HSV color space. For each color band, we create four co-occurrence matrices, one for each of the ($0°$, $45°$, $90°$, and $135°$) directions. The displacement considered is 1 pixel and each color band is quantized into 16 bins. The 12 descriptors mentioned earlier are then extracted from each co-occurrence matrix. This results in $10,080$ features.

We calculate HOG features considering blocks with sizes ranging from 12×12 to 32×64. In our configuration there are 326 blocks. As in [5], 36 features are extracted from each block, resulting in a total of $11,736$ features.

The color histograms are computed from overlapping blocks of 32×32 and 16×16 pixels extracted from the image window. 16-bin histograms are computed for the R, G and B color bands, and then concatenated. The resulting number of features extracted by this method is $5,472$. Aggregating across all three feature channels, the feature vector describing each appearance contains $27,288$ elements.

To evaluate the approach described in Section 2.3, we compare the results to another well-know dimensionality reduction technique, PCA, and to SVM. With PCA, we first reduce the dimensionality of the feature vector and then we use the same classification approach described for PLS. However, with SVM the data is classified directly in the original feature space.

We consider four setups for the SVM: linear SVM with one-against-all scheme, linear multi-class SVM, kernel SVM with one-against-all scheme, and kernel multi-class SVM. A polynomial kernel with degree 3 is used. In the experiments we used the LIBSVM [3].

Since the high dimensionality of the feature space poses difficulties to compute the covariance matrix for PCA, we use a randomized PCA algorithm [14]. In addition, the classification for PCA uses the same scheme described in Section 2.3 for PLS, where a query sample is classified as belonging to the model presenting the smallest Euclidean distance in the low dimensional space.

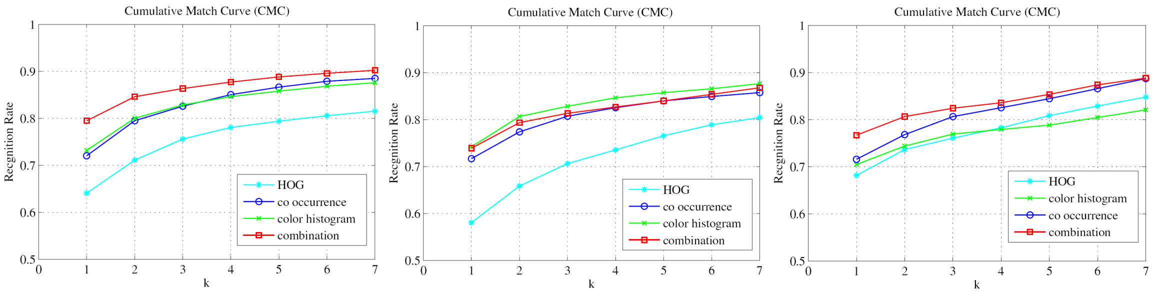
Experimental results are reported in terms of the cumulative match characteristic (CMC) curves. These curves show the probability that a correct match is within the k-nearest candidates (in our experiments k varies from 1 to 7).

Before performing comparisons, we use the video sequence #0 to evaluate how many dimensions (number of weight vectors) should be used in the low dimensional latent space for PLS and PCA. Figure 5 shows the CMC curves for both when the number of factors is changed. The best results are obtained when 3 and 4 factors are considered for PLS and PCA, respectively. These parameters will be used throughout the experiments.

All experiments were conducted on an Intel Xeon, 3 GHz quad-core processor with 4GB of RAM running Linux operating system. The implementation is based on MATLAB.

Evaluation. Figure 6 shows the recognition rates obtained for each feature individually and their combination. In both cases the dimensionality is reduced using PLS. In general, the combination of features outperforms the results obtained when individual features are considered. This justifies the use of a rich set of features.

Figure 8 compares the PLS method to PCA and different setups of the SVM. We can see that the PLS approach



(a) sequence #1 (b) sequence #2 (c) sequence #3

Figure 6. Recognition rates obtained by using individual features and combination of all three feature channels used in this work.

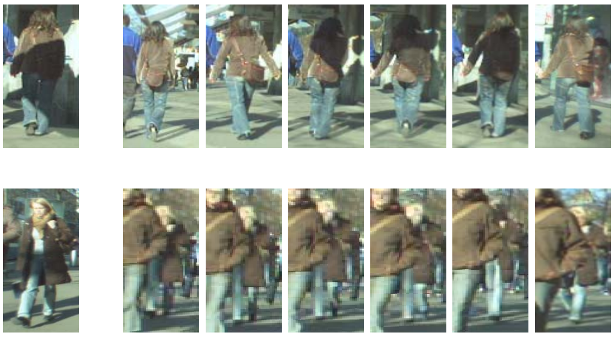


Figure 7. Misclassified samples of sequence #3. The images on the left show the training samples used to learn each appearance model. Images on the right contain samples misclassified by the PLS method.

obtains high recognition rates on the testing sequences of the ETHZ dataset. The results demonstrate, as one would expect, that PLS-based dimensionality reduction provides a more discriminative low dimensional latent space than PCA. In addition, we see that classification performed by SVM in high dimensional feature space when the number of training samples is small might lead to poor results. Finally, compared to the other methods, our approach achieves better results mainly when the number of different appearances being considered is high, i.e. sequences #1 and #2.

In terms of computational cost, Figure 8 shows that the proposed method, is in general, between PCA and SVM. The training and testing computational costs depend on the number of people and number of testing samples. Sequence #1 has $4,857$ testing samples amongst the 83 different people and sequences #2 and #3 have $1,961$ and $1,762$, respectively. The number of different people in each sequence is described in Figure 3.

Figure 7 shows some of the misclassified samples of sequence #3 together with the samples used to learn the PLS models. We see that the misclassifications are due to changes in the appearance, occlusion and non-linear illumination change. This problem commonly happens when the appearance models are not updated over time. However, if integrated into a tracking framework, for example, the proposed method could use some model update scheme that might lead to higher recognition rates.

Finally, samples used to learn the appearance-based models for sequence #1 are shown in Figure 9. The large number of people and high similarity in their appearances increases the ambiguity among the models.

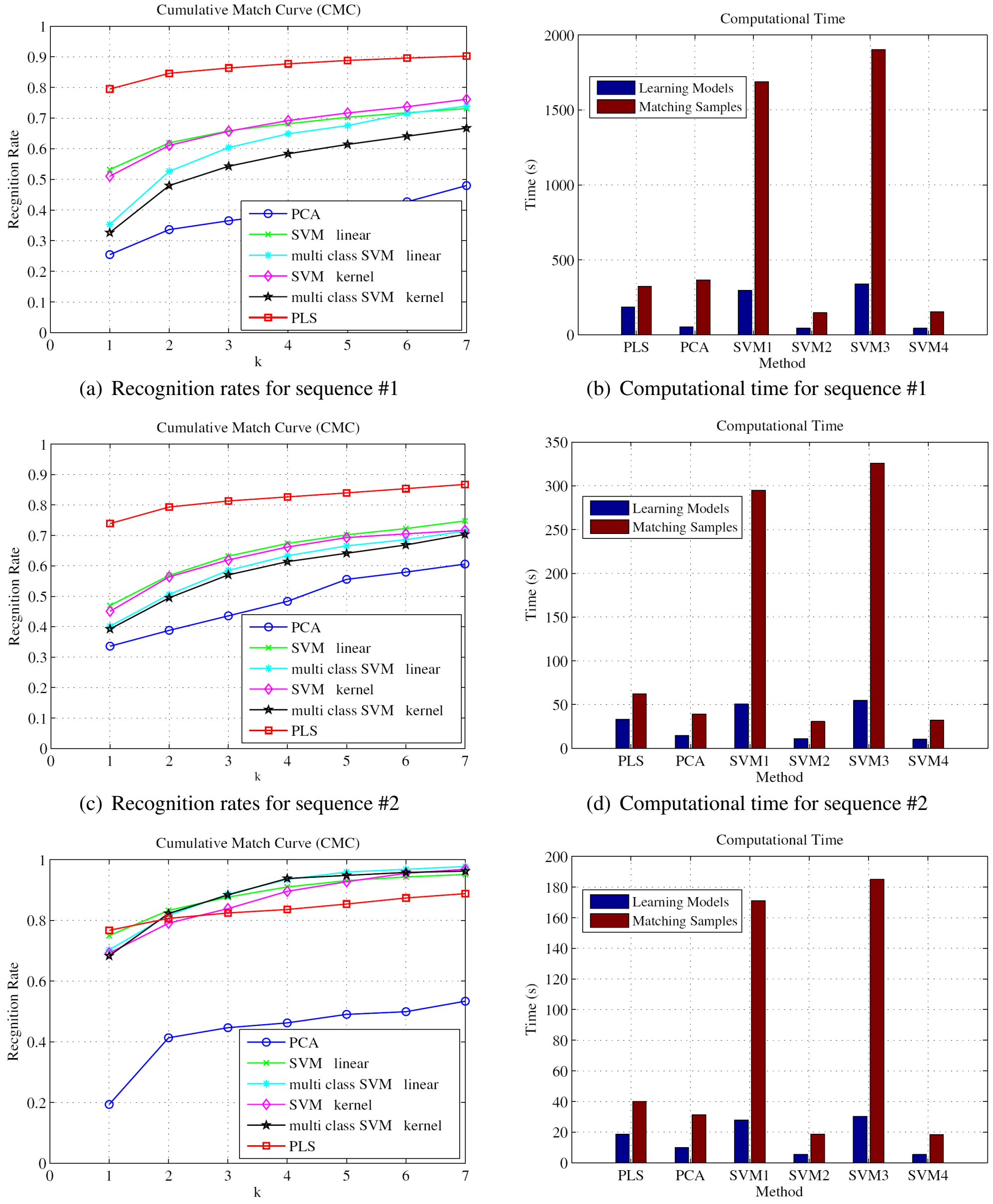
4 Conclusions and Future Work

We described a framework to learn discriminative appearance-based models based on PLS analysis. The results show that this method outperforms other approaches considering an one-against-all scheme. It has also been demonstrated that the use of a richer set of features leads to improvements in results.

As a future direction, we intend to incorporate the use of the richer set of features and the high discriminative dimensionality reduction provided by PLS into a pairwise-coupling framework aiming at further reduction of ambiguity when the number of appearances increases.

Acknowledgements

This research was partially supported by the ONR MURI grant N00014-08-10638 and the ONR surveillance grant N00014-09-10044. W. R. Schwartz acknowledges “Coordenação de Aperfeiçoamento de Pessoal de Nível Superior” (CAPES - Brazil, grant BEX1673/04-1). The authors also thank Aniruddha Kembhavi for his comments.



(a) Recognition rates for sequence #1

(b) Computational time for sequence #1

(c) Recognition rates for sequence #2

(d) Computational time for sequence #2

(e) Recognition rates for sequence #3

(f) Computational time for sequence #3

Figure 8. Performance and time comparisons considering the PLS method, PCA and SVM. SVM1: linear SVM (one-against-all), SVM2: linear SVM (multi-class), SVM3: kernel SVM (one-against-all), SVM4: kernel SVM (multi-class).

References

[1] Y. Amit, D. Geman, and K. Wilder. Joint Induction of Shape Features and Tree Classifiers. *PAMI*, 19(11):1300–1305, 1997.

[2] A. Bosch, A. Zisserman, and X. Muoz. Image Classification using Random Forests and Ferns. In *ICCV*, pages 1–8, 2007.

[3] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at www.csie.ntu.edu.tw/ cjlin/libsvm.

[4] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based Object Tracking. *PAMI*, 25(5):564–577, 2003.

[5] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. In *CVPR*, 2005.

[6] A. Elgammal, R. Duraiswami, and L. Davis. Probabilistic Tracking in Joint Feature-Spatial Spaces. In *CVPR*, volume 1, pages 781–788, 2003.

[7] A. Ess, B. Leibe, and L. V. Gool. Depth and Appearance for Mobile Scene Analysis. In *ICCV*, 2007.

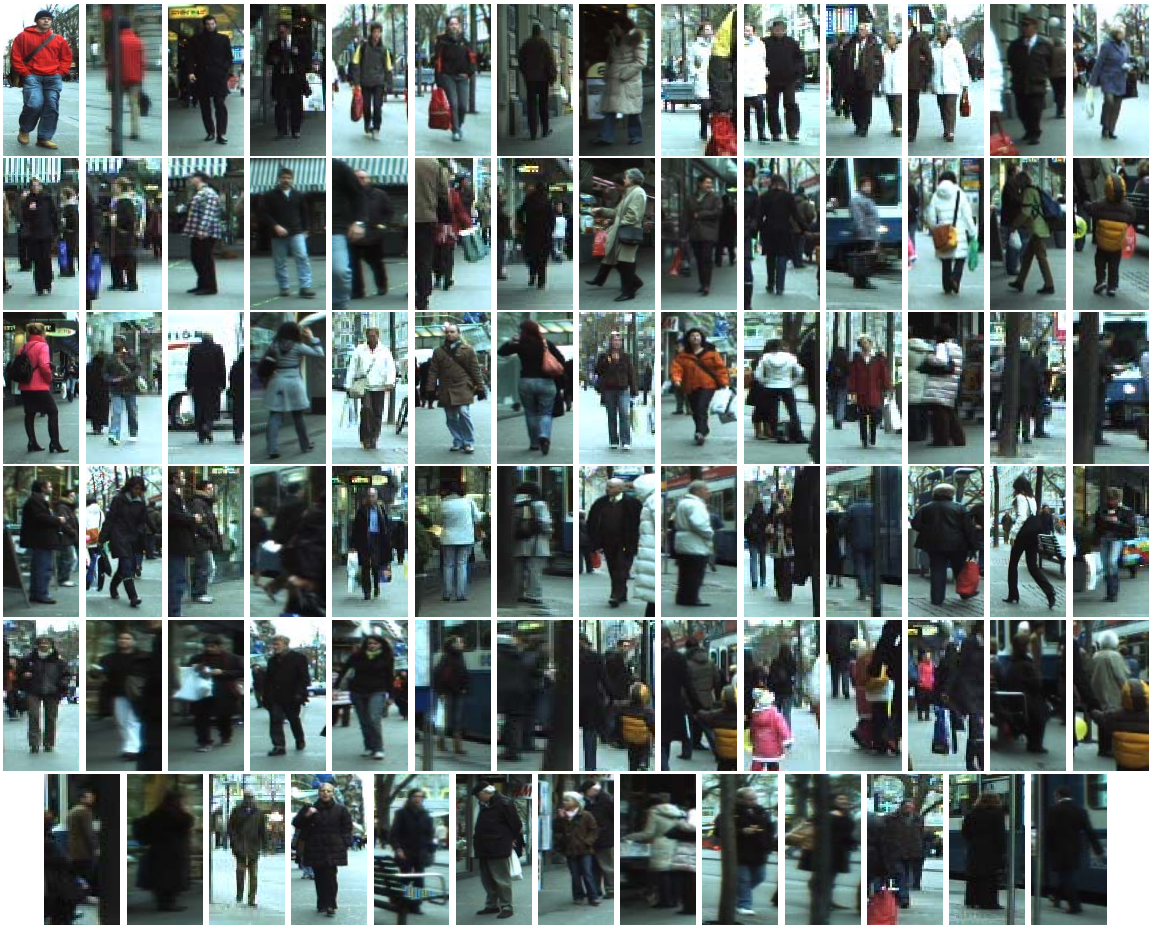


Figure 9. Samples of different people in sequence #1 used to learn the models.

[8] N. Gheissari, T. B. Sebastian, and R. Hartley. Person Reidentification Using Spatiotemporal Appearance. In *CVPR*, pages 1528–1535, 2006.

[9] R. Haralick, K. Shanmugam, and I. Dinstein. Texture Features for Image Classification. *IEEE Transactions on Systems, Man, and Cybernetics*, 3(6), 1973.

[10] J. Li, S. Zhou, and R. Chellappa. Appearance Modeling Under Geometric Context. In *ICCV*, volume 2, pages 1252–1259, 2005.

[11] Z. Lin and L. S. Davis. Learning Pairwise Dissimilarity Profiles for Appearance Recognition in Visual Surveillance. In *International Symposium on Advances in Visual Computing*, pages 23–34, 2008.

[12] S. Maji, A. Berg, and J. Malik. Classification Using Intersection Kernel Support Vector Machines is Efficient. In *CVPR*, 2008.

[13] C. Nakajima, M. Pontil, B. Heisele, and T. Poggio. Full-body Person Recognition System. *Pattern Recognition*, 36(9):1997–2006, 2003.

[14] V. Rokhlin, A. Szlam, and M. Tygert. A Randomized Algorithm for Principal Component Analysis. *ArXiv e-prints*, 2008.

[15] R. Rosipal and N. Kramer. Overview and Recent Advances in Partial Least Squares. *Lecture Notes in Computer Science*, 3940:34–51, 2006.

[16] W. R. Schwartz, A. Kembhavi, D. Harwood, and L. S. Davis. Human Detection Using Partial Least Squares Analysis. In *ICCV*, 2009.

[17] M. Varma and D. Ray. Learning the Discriminative Power-Invariance Trade-Off. In *ICCV*, pages 1–8, 2007.

[18] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu. Shape and Appearance Context Modeling. In *ICCV*, pages 1–8, 2007.

[19] H. Wold. Partial Least Squares. In S. Kotz and N. Johnson, editors, *Encyclopedia of Statistical Sciences*, volume 6, pages 581–591. Wiley, New York, 1985.

[20] B. Wu and R. Nevatia. Optimizing Discrimination-Efficiency Tradeoff in Integrating Heterogeneous Local Features for Object Detection. In *CVPR*, pages 1–8, 2008.

[21] H. Zhang, A. Berg, M. Maire, and J. Malik. SVM-KNN: Discriminative Nearest Neighbor Classification for Visual Category Recognition. In *CVPR*, volume 2, pages 2126–2136, 2006.

[22] W. Zhang, G. Zelinsky, and D. Samaras. Real-time Accurate Object Detection using Multiple Resolutions. In *ICCV*, 2007.