Human Detection Using Partial Least Squares Analysis

William Robson Schwartz, Aniruddha Kembhavi, David Harwood, Larry S. Davis

University of Maryland, A.V.Williams Building, College Park, MD 20742

schwartz@cs.umd.edu, anikem@umd.edu, harwood@umiacs.umd.edu, lsd@cs.umd.edu

Abstract

Significant research has been devoted to detecting people in images and videos. In this paper we describe a human detection method that augments widely used edge-based features with texture and color information, providing us with a much richer descriptor set. This augmentation results in an extremely high-dimensional feature space (more than 170,000 dimensions). In such high-dimensional spaces, classical machine learning algorithms such as SVMs are nearly intractable with respect to training. Furthermore, the number of training samples is much smaller than the dimensionality of the feature space, by at least an order of magnitude. Finally, the extraction of features from a densely sampled grid structure leads to a high degree of multicollinearity. To circumvent these data characteristics, we employ Partial Least Squares (PLS) analysis, an efficient dimensionality reduction technique, one which preserves significant discriminative information, to project the data onto a much lower dimensional subspace (20 dimensions, reduced from the original 170,000). Our human detection system, employing PLS analysis over the enriched descriptor set, is shown to outperform state-of-the-art techniques on three varied datasets including the popular INRIA pedestrian dataset, the low-resolution gray-scale DaimlerChrysler pedestrian dataset, and the ETHZ pedestrian dataset consisting of full-length videos of crowded scenes.

1. Introduction

Effective techniques for human detection are of special interest in computer vision since many applications involve people's locations and movements. Thus, significant research has been devoted to detecting, locating and tracking people in images and videos. Over the last few years the problem of detecting humans in single images has received considerable interest. Variations in illumination, shadows, and pose, as well as frequent inter- and intra-person occlusion render this a challenging task. Figure 1 shows an image of a particularly challenging scene with a large number of persons, overlaid with the results of our system.

Two main approaches to human detection have been explored over the last few years. The first class of methods consists of a generative process where detected parts of the human body are combined according to a prior human model. The second class of methods considers purely statistical analysis that combine a set of low-level features within a detection window to classify the window as containing a human or not. The method presented in this paper belongs to the latter category.

Figure 1. Image demonstrating the performance of our system in a complex scene. The image (689×480 pixels) is scanned at 10 scales to search for humans of multiple sizes. We achieve minimal false alarms even though the number of detection windows is $44,996$ (best visualized in color).

Dalal and Triggs [5] proposed using grids of Histograms of Oriented Gradient (HOG) descriptors for human detection, and obtained good results on multiple datasets. The HOG feature looks at the spatial distribution of edge orientations. However, this may ignore some other useful sources of information, thus leading to a number of false positive detections such as the ones shown in Figure 2. Our analysis shows that information such as the homogeneity of human clothing, color, particularly skin color, typical textures of human clothing, and background textures complement the HOG features very well. When combined, this richer set of descriptors helps improve the detection results significantly.

A consequence of such feature augmentation is an extremely high dimensional feature space (more than $170,000$ dimensions), rendering many classical machine learning techniques such as Support Vector Machines (SVM) intractable. In contrast, the number of samples in our training dataset is much smaller (almost 20 times smaller than

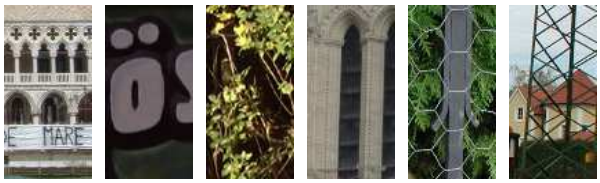


Figure 2. False positives obtained when only edge information (using HOG features) is considered.

the dimensionality). Furthermore, our features are extracted from neighboring blocks within a detection window, which increases the multicollinearity of the feature set. The nature of our proposed feature set makes an ideal setting for a statistical technique known as Partial Least Squares (PLS) regression [23]. PLS is a class of methods for modeling relations between sets of observations by means of latent variables. Although originally proposed as a regression technique, PLS can be also be used as a *class aware* dimensionality reduction tool. We use PLS to project our high dimensional feature vectors onto a subspace of dimensionality as low as 20. In such low dimensional spaces, standard machine learning techniques such as quadratic classifiers and SVMs can be used for our classification task.

Our proposed human detection approach outperforms state-of-the-art approaches on multiple standard datasets. Since the number of detection windows within an image is very high (tens of thousands for a 640×480 image scanned at multiple scales), it is crucial to obtain good detection results at very small false alarm rates. On the popular INRIA person dataset [5], we obtain superior results at false alarm rates as low as 10^{-5} and 10^{-6} false positives per window (FPPW). We also test on the ETHZ pedestrian dataset [7] consisting of full-length videos captured in crowded scenes. Even though we do not retrain our human detector using the provided training set (but use the detector trained on the INRIA training set), our method outperforms other approaches that utilize many more sources of information such as depth maps, ground-plane estimation, and occlusion reasoning [7]. Finally, we also demonstrate our method on detecting humans at very low resolutions (18×36 pixels) using the DaimlerChrysler dataset [18].

2. Related Work

The work of Dalal and Triggs [5] is notable because it was the first paper to report impressive results on human detection. Their work uses HOG as low-level features, which were shown to outperform features such as wavelets [16], PCA-SIFT [11] and shape contexts [2].

To improve detection speed, Zhu et al. [28] propose a rejection cascade using HOG features. Their method considers blocks of different sizes, and to train the classifier for each stage, a small subset of blocks is selected randomly. Also based on HOG features, Zhang et al. [27] propose a multi-resolution framework to reduce the computational cost. Begard et al. [1] address the problem of real-time pedestrian detection by considering different implementations of the AdaBoost algorithm.

Using low-level features such as intensity, gradient, and spatial location combined by a covariance matrix, Tuzel et al. [22] improve the results obtained by Dalal and Triggs. Since the covariance matrices do not lie in a vector space, the classification is performed using LogitBoost classifiers combined with a rejection cascade designed to accommodate points lying on a Riemannian manifold. Mu et al. [17] propose a variation of local binary patterns to overcome some drawbacks of HOG, such as lack of color information. Chen and Chen [4] combine intensity-based rectangle features and gradient-based features using a cascaded structure for detecting humans. Applying combination of edgelets [25], HOG descriptors [5], and covariance descriptors [22], Wu and Nevatia [26] describe a cascade-based approach where each weak classifier corresponds to a subregion within the detection window from which different types of features are extracted. Dollar et al. [6] propose a method to learn classifiers for individual components and combine them into an overall classifier. The work of Maji et al. [14] uses features based on a multi-level version of HOG and histogram intersection kernel SVM based on the spatial pyramid match kernel [12].

Employing part-based detectors, Mikolajczyk et al. [15] divide the human body into several parts and apply a cascade of detectors for each part. Shet and Davis [20] apply logical reasoning to exploit contextual information, augmenting the output of low-level detectors. Based on deformable parts, Felzenszwalb et al. [9] simultaneously learn part and object models and apply them to person detection, among other applications. Tran and Forsyth [21] use an approach that mixes a part-based method and a subwindow-based method into a two stage method. Their approach first estimates a possible configuration of the person inside the detection window, and then extracts features for each part resulting from the estimation. Similarly, Lin and Davis [13] propose a pose-invariant feature extraction method for simultaneous human detection and segmentation, where descriptors are computed adaptively based on human poses.

3. Proposed Method

Previous studies [14, 22, 26] have shown that significant improvement in human detection 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.

Humans in standing positions have distinguishing characteristics. First, strong vertical edges are present along the boundaries of the body. Second, clothing is generally uniform. Clothing textures are different from natural textures observed outside of the body due to constraints on the manufacturing of printed cloth. Third, the ground is composed mostly of uniform textures. Finally, discriminatory color information is found in the face/head regions.

Thus, edges, colors and textures capture important cues for discriminating humans from the background. To capture these cues, the low-level features we employ are the original HOG descriptors with additional color information,

called *color frequency*, and texture features computed from co-occurrence matrices.

To handle the high dimensionality resulting from the combination of features, PLS is employed as a dimensionality reduction technique. PLS is a powerful technique that provides dimensionality reduction for even hundreds of thousands of variables, accounting for class labels in the process. The latter point is in contrast to traditional dimensionality reduction techniques such as Principal Component Analysis (PCA).

The steps performed in our detection method are the following. For each detection window in the image, features extracted using original HOG, color frequency, and co-occurrence matrices are concatenated and analyzed by the PLS model to reduce dimensionality, resulting in a low dimensional vector. Then, a simple and efficient classifier is used to classify this vector as either a human or non-human. These steps are explained in the following subsections.

3.1. Feature Extraction

We decompose a detection window, d_i, into overlapping blocks and extract a set of features for each block to construct the feature vector $\boldsymbol{v_i}$.

To capture texture, we extract features from co-occurrence matrices [10], 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 [10]. 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 captures edge or gradient structures that are characteristic of local shape [5]. Since the histograms are computed for regions of a given size within the detection 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. Although colors may not be consistent due to variability in clothing, certain dominant colors are more often observed in humans, mainly in the face/head regions. In order to incorporate color we used the original HOG to extract a descriptor called *color frequency*. In HOG, the orientation of the gradient for a pixel is chosen from the color band corresponding to the highest gradient magnitude. Some color information is captured by the number of times each color band is chosen. Therefore, we construct a three bin histogram that tabulates the number of times each color band is chosen. In spite of its simplicity, experimental results have shown that color frequency increases detection performance.

Once the feature extraction process is performed for all blocks inside a detection window d_i, features are concatenated creating an extremely high-dimensional feature vector $\boldsymbol{v_i}$. Then, $\boldsymbol{v_i}$ is projected onto a set of weight vectors (discussed in the next section), which results in a low dimensional representation that can be handled by classification methods.

3.2. 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 [19], a brief mathematical description of the procedure is provided below.

Let $\mathcal{X} \subset \mathbb{R}^m$ denote an 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

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

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 [23], constructs a set of weight vectors (or projection vectors) $W = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots \boldsymbol{w}_p\}$ 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 latent vectors had been extracted.

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

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. Fisher Discriminant Analysis (FDA) is, in this way, similar to PLS. However, FDA has the limitation that

after dimensionality reduction, there are only $c-1$ meaningful latent variables, where c is the number of classes being considered. Additionally, when the number of features exceeds the number of samples, the covariance estimates do not have full rank and the weight vectors cannot be extracted.

3.3. Speed Issues

Although detection results can be improved by utilizing overlapping blocks for low-level feature extraction within the detection window, the dimensionality of the feature vector becomes extremely high. As a result, the speed of the human detector decreases significantly due to the time needed to extract features and project them.

To overcome this problem, we employ a two-stage approach. In a fast first stage, based on a small number of features, the majority of detection windows (those with low probability of containing humans) are discarded. The remaining windows are evaluated during a second stage where the complete set of features allows challenging samples to be correctly classified.

The reduced set of features used during the first stage is obtained by selecting representative blocks within the detection window. We use a PLS-based feature selection method called variable importance on projection (VIP) [24] to do this. VIP provides a score for each feature, so that it is possible to rank the features according to their predictive power in the PLS model (the higher the score the more importance a feature presents). VIP for the j-th feature is defined as

$$\text{VIP}_j = \sqrt{m \sum_{k=1}^{p} b_k^2 w_{jk}^2 / \sum_{k=1}^{p} b_k^2}$$

where m denotes the number of features, w_{jk} is the j-th element of vector $\boldsymbol{w}_k$, and b_k is the regression weight for the k-th latent variable, $b_k = \boldsymbol{u}_k^T \boldsymbol{t}_k$.

The speed improvements are twofold: (i) reducing the overall number of feature computations; (ii) reducing the time to create the data structure for a block, i.e. computing a co-occurrence matrix from which features are extracted. If features were selected individually, then a data structure might need to be constructed for a block to compute only one feature. To avoid that, we select features based on blocks. This way, data structures for a block are only built if several features within the block present some importance.

To obtain the relative discriminative power among blocks we build a PLS model for each block, from which only the first latent variable is considered (since PLS considers class labels, the first latent variable can be used as a clue about how well that block contributes to the detection). A global PLS model is built using as input only the first latent variable of every block. Then, VIP scores are computed with respect to this PLS model, in this way, blocks can be ranked according to their importance in detection. Finally, the features used in the first stage of our approach are those computed from blocks having high rank.

4. Experiments

We now present experiments to evaluate several aspects of our proposed approach. First, we demonstrate the need for dimensionality reduction and the advantages of using PLS for this purpose. Second, we evaluate the features used in our system. Third, we compare various classifiers that can be used to classify the data in the low dimensional subspace. Fourth, we discuss the computational cost of our method. Finally, we compare the proposed system to state-of-the-art algorithms on several datasets considering cropped as well as full images.

Experimental Setup. For co-occurrence feature extraction we use block sizes of 16×16 and 32×32 with shifts of 8 and 16 pixels, respectively. 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. 12 descriptors mentioned earlier are then extracted from each co-occurrence matrix. This results in $63,648$ features.

We calculate HOG features similarly to Zhu et al. [28], where blocks with sizes ranging from 12×12 to 64×128 are considered. In our configuration there are $2,748$ blocks. For each block, 36 features are extracted, resulting in a total of $98,928$ features. In addition, we use the same set of blocks to extract features using the color frequency method. This results in three features per block, and the total number of resulting features is $8,244$. Aggregating across all three feature channels, the feature vector describing each detection window contains $170,820$ elements.

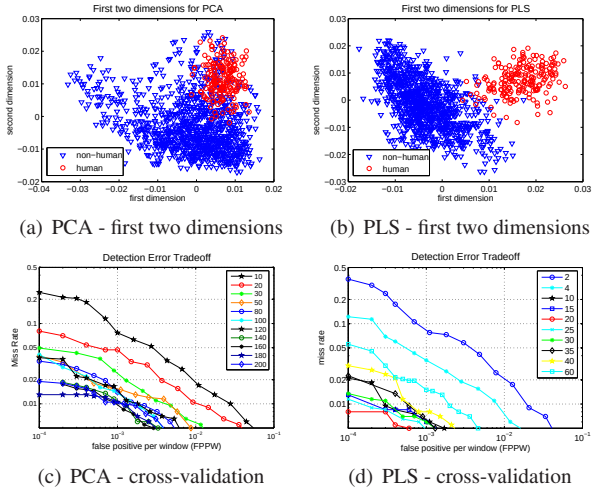
We estimate the parameters of our system using a 10-fold cross-validation procedure on the training dataset provided by INRIA Person Dataset [5]. The INRIA person dataset provides a training dataset containing 2416 positive samples of size 64×128 pixels and images containing no humans, used to obtain negative exemplars. We sample this set to obtain our validation set containing 2000 positive samples and 10000 negative samples. In sections 4.1 to 4.4 our experiments are performed using the INRIA person dataset.

Experimental results using INRIA Person Dataset are presented using detection error tradeoff (DET) curves on log-log scales. The x-axis corresponds to false positives per window (FPPW), defined by *FalsePos/(TrueNeg + FalsePos)* and the y-axis shows the miss rate, defined by *FalseNeg/(FalseNeg + TruePos)*. To clarify the results shown throughout the paper, curves where the lowest FPPW is 10^{-4} are obtained using the training data, while curves where the lowest FPPW is 10^{-6} are obtained using the testing data.

All experiments were conducted on an Intel Xeon 5160, 3 GHz dual core processor with 8GB of RAM running Linux operating system.

4.1. Dimensionality Reduction

PLS+QDA Vs SVM. We first examine the feasibility of applying support vector machines (SVM) directly on



(a) PCA - first two dimensions (b) PLS - first two dimensions

(c) PCA - cross-validation (d) PLS - cross-validation

Figure 3. Comparison of PCA and PLS for dimensionality reduction. (a-b) projection of the first two dimensions of the training samples for one of the models learned in the cross-validation. (c-d) DET curves according to the number of dimensions used to train the classifier (best visualized in color).

the high dimensional feature space ($170,820$ features per sample). Table 1 shows the comparison between time required to train a linear SVM and the time required to train a PLS model along with a Quadratic Discriminant Analysis (QDA) model (we use the QDA classifier, but in later subsections we provide a comparison to other classifiers as well). We used the LIBSVM [3] package for this purpose. As the number of training samples is increased, the training time also increases. For more than 1800 samples we were unable to train a linear SVM since the procedure ran out of memory. In addition, the computational cost to learn a PLS model and train a QDA classifier is an order of magnitude smaller than the cost for training an SVM. These results indicate that for such a high dimensional space, it is more suitable to project the data onto a low dimensional subspace and then learn a classifier.

# samples	PLS + QDA	SVM
200	23.63	131.72
600	62.62	733.63
1000	97.38	1693.50
1400	135.81	2947.51
1800	174.57	4254.63
2200	213.93	-
11370	813.03	-

Table 1. Time, in seconds, to train SVM and PLS + QDA models. The number of features per sample is 170,820. The training time increases with an increase in the number of training samples.

PLS Vs PCA. We now establish a baseline using Principal Component Analysis (PCA) to perform linear dimensionality reduction and compare its results to PLS. Figures 3(c) and (d) show the DET curves obtained for a QDA classifier in the PCA subspace as well as in the PLS subspace. It is interesting to note that while the best results are obtained by using the first 20 PLS latent variables, the performance of the system drops when the number of latent variables is increased beyond 20. This can be attributed to overfitting of the data caused by using a larger number of latent variables. The results achieved while using the first 20 latent variables are the best results obtained over both subspaces (0.8% miss rate at 10^{-4} FPPW). The best performance on the PCA subspace is obtained for a dimensionality of 180 (1.8% miss rate at 10^{-4} FPPW).

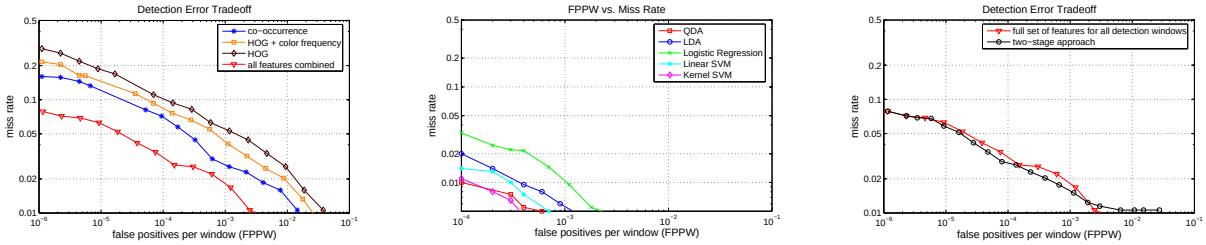
As the dimensionality of the subspace increases, the time required to project the high dimensional feature vectors onto the low dimensional space also increases. On our computer, projecting the feature vector for a single window onto a 180 dimensional PCA subspace takes 0.0264 seconds while it takes 0.0032 seconds to project onto the 20 dimensional PLS subspace. Since an image contains several thousand windows, a computational cost of 0.0264 seconds/window is substantially worse than that for PLS. Thus, in addition to the superior performance, the computational cost of projection makes PLS more suitable for our application than PCA. Figure 3(a) and (b) show the training dataset projected onto the first two dimensions for PLS and PCA. PLS clearly achieves better class separation than PCA.

4.2. Feature Evaluation

Comparing features. Figure 4(a) shows the results of the three classes of features used in our system as well as the combined performance. We show results combining the HOG and color frequency features to demonstrate the positive contribution of the color features. A significant improvement is achieved when all features are combined.

Analysis of the PLS Weight Vectors. In this experiment, we perform an analysis of the contribution of each feature channel based on the weights of the PLS weight vectors used to project the features onto the low dimensional subspace. We use the same idea as described in Section 3.3. For a given block in the detection window, we create a PLS model for each feature channel. Then, using only the first latent variable for every block, we learn a global PLS model. Figure 5 shows the weights for the first five projection vectors of this global PLS model. The features considered are HOG, co-occurrence extracted from color bands H, S and V, and the color frequency.

Figure 5 shows how each feature channel (edge, texture, color) provides information from different regions within the detection window. This supports our claim that the considered features complement each other, leading to an improvement over single-feature-based methods. For example, the first weight vector of the HOG feature set captures information about the body shape due to the presence of edges. Co-occurrence matrix features from color band H extract information around the body silhouette. Color bands S and V provide information about the head and homogeneous parts inside the body, respectively. Except for the first weight vector, color frequency features are able to identify regions located in the head due to similarity of the dominant colors in that region (skin color).



(a) (b) (c)

Figure 4. (a) results obtained by using different features and combination of all three feature channels used by this work; (b) comparison of several classification methods for the low dimensional PLS subspace; (c) results after adding two stages compared to results obtained without speed optimization.

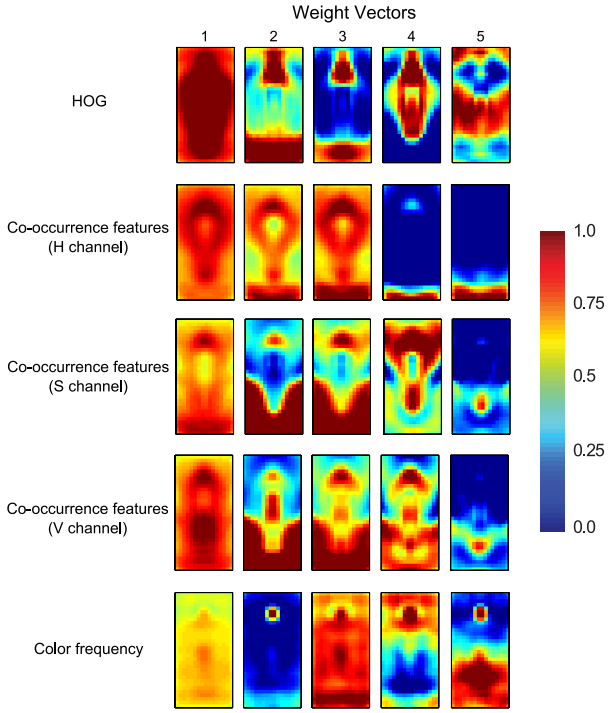


Figure 5. Weight vectors for different features within the detection window. Red indicates high importance, blue low (the plots are in the same scale and normalized to interval $[0, 1]$).

4.3. Classification in Low Dimensional Space

To evaluate the classification in the low dimensional subspace, we compare the performance of several classifiers using the 10-fold cross-validation described earlier. Figure 4(b) shows the results. According to the results, QDA classifier, kernel SVM and linear SVM achieved comparable performance in low dimensional subspace. Due to its simplicity, we have chosen to use QDA in our system. PLS tends to produce weight vectors that provide a good separation of the two classes for the human detection problem, as shown in Figure 3(b). This enables us to use simple classifiers in the low dimensional subspace.

4.4. Computational Cost

We accelerate the process using the two-stage approach described in Section 3.3. To reduce the number of features computed in the first stage, we rank blocks according to their VIP scores and then select only those features in blocks with higher rankings. Using 10-fold cross-validation in the training set, we select a subset of blocks containing $3,573$ features per detection window, together with a probability threshold to decide whether a detection window needs to be considered for the second stage.

It is important to note that the use of the first stage alone achieves poor results for low false alarm rates. Therefore, for the detection windows not discarded in the first stage (approximately 3% for the INRIA person dataset), the complete feature set is computed. For the testing set of the INRIA person dataset, the results shown in Figure 4(c) indicate no degradation in performance at low false alarm rates when the two-stage approach is used, as compared to computing the full set of features for all detection windows. After speeding the process up using our two-stage method, we were able to process 2929 detection windows per second.

4.5. Evaluation and Comparisons

In this section we evaluate the proposed system on different datasets and compare it to state-of-the-art methods.

INRIA Person Dataset. The INRIA person dataset [5] provides both training and testing sets containing positive samples of size 64×128 pixels and negatives images (containing no humans). To estimate weight vectors (PLS model) and train the quadratic classifier we employ the following procedure. First, all 2416 positive training samples and 5000 of the negative detection windows, sampled randomly from training images, are used. Once the first model is created, we use it to classify negative windows in the training set. The misclassified windows are added into the 5000 negative windows and a new PLS model and new classifier parameters are estimated. This process is repeated a few times and takes approximately one hour. Our final PLS model considers 8954 negative and 2416 positive samples, using 20 weight vectors (as discussed in section 4.1).

Figure 6(a) compares results obtained by the proposed approach to methods published previously. Our results were obtained using 1126 positive testing samples and by shifting the detection windows by 8 pixels in the negative testing images, all of which are available in the dataset. While we were able to run the implementations for methods [5, 22], curves for methods [6, 13, 14, 26] were obtained from their

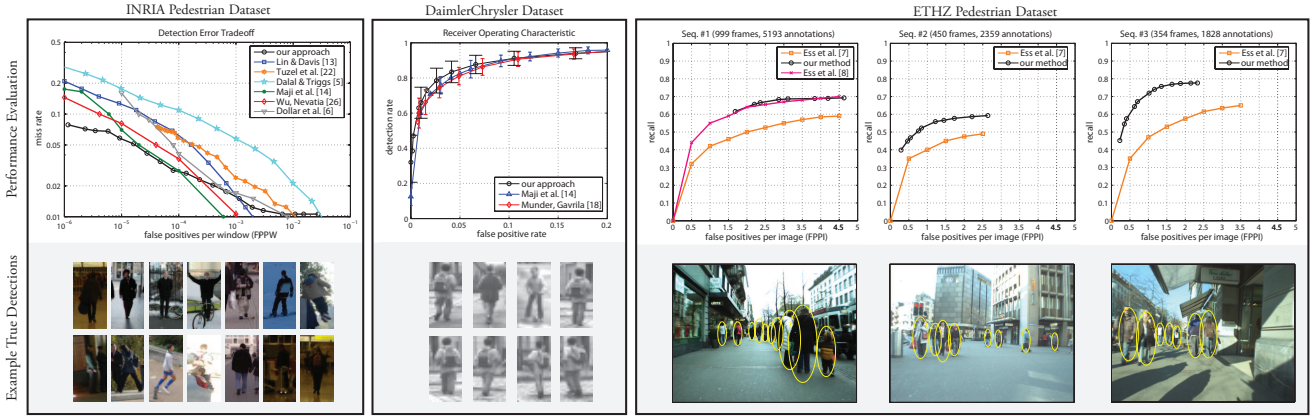


Figure 6. Evaluation of our method on multiple pedestrian datasets. First row shows performance and comparisons with state-of-the-art methods. Second row shows some sample true detections for each dataset (best visualized in color).

reported results. The PLS approach outperforms all methods in regions of low false alarm rates, i.e. 5.8% miss rate at 10^{-5} FPPW and 7.9% miss rate at 10^{-6} FPPW.

DaimlerChrysler Pedestrian Dataset. This dataset provides grayscale samples of size 18×36 pixels [18]. We adapt our feature extraction methods for these image characteristics as follows. For co-occurrence feature extraction, we use block sizes of 8×8 and 16×16 with shifts of 2 pixels for both. Co-occurrence matrices are estimated using the brightness channel quantized into 16 bins. For HOG feature extraction, we adopt the same approach used for the INRIA person dataset; however, block sizes now range from 8×8 to 18×36. Due to the lack of color information, the color frequency feature cannot be considered in this experiment.

The DaimlerChrysler dataset is composed of five disjoint sets, three for training and two for testing. To obtain results that can be compared to those presented by Maji et al. [14] and by Munder and Gavrila [18], we report results by training on two out of three training sets at a time. Therefore, we obtain six curves from which the confidence interval of the true mean detection rate is given by the $t_{(\alpha/2,N-1)}$ distribution with desired confidence of $1 - \alpha = 0.95$ and $N = 6$. The boundaries of this interval are approximated by $\overline{y} \pm 1.05s$, where $\overline{y}$ and s denote the estimated mean and standard deviation, respectively [18].

Figure 6(b) compares results obtained by the proposed method to results reported in [14, 18]. In contrast to previous graphs, this shows detection rates instead of miss rates on the y-axis and both axes are shown using linear scales. Similar to experiments conducted on the INRIA person dataset, the results obtained with the proposed method show improvements in regions of low false alarm rates.

ETHZ Dataset. We evaluate our method for un-cropped full images using the ETHZ dataset [7]. This dataset provides four video sequences, one for training and three for testing (640×480 pixels at 15 frames/second). Even though a training sequence is provided, we do not to use it; instead we use the same PLS model and QDA parameters learned on the INRIA training dataset. This allows us to evaluate the generalization capability of our method to different datasets.

For this dataset we use false positives per image (FPPI) as the evaluation metric, which is more suitable for evaluating the performance on full images [21]. Using the same evaluation procedure described in [7] we obtain the results shown in Figure 6(c), (d) and (e) for the testing sequences provided. We use only the images provided by the left camera and perform the detection for each single image at 11 scales without considering any temporal smoothing. We do not train our detector on the provided training set and we do not use any additional cues such as depth maps, ground-plane estimation, and occlusion reasoning, all of which are used by [7]. Yet, our detector outperforms the results achieved by [7] in all three video sequences.

The work by Ess et al. [8] also considers sequence #1 in their experiments, so we have added their results in Figure 6(c). Even though [8] uses additional cues such as tracking information, our method, trained using the training set of INRIA dataset, achieves very similar detection results.

Additional Set of Images. We present some results in Figure 7 for a few images obtained from INRIA testing dataset and Google. These results were also obtained using the same PLS model and QDA parameters learned on the INRIA training dataset. We scan each image at 10 scales. Despite the large number of detection windows considered, the number of false alarms produced is very low.

5. Conclusions

We have proposed a human detection method using a richer descriptor set including edge-based features, texture measures and color information, obtaining a significant improvement in results. The augmentation of these features generates a very high dimensional space where classical machine learning methods are intractable. The characteristics of our data make an ideal setting for applying PLS to obtain a much lower dimensional subspace where we use simple and efficient classifiers. We have tested our approach

(a) 640 × 480 (41,528 det. windows) (b) 1632 × 1224 (389,350 det. windows) (c) 1600 × 1200 (373,725 det. windows)

Figure 7. Results obtained from images containing people of different sizes and backgrounds rich in edge information. The image size and the total number of detection windows considered are indicated in the caption (best visualized in color).

on a number of varied datasets, demonstrated its good generalization capabilities and shown it to outperform state-of-the-art methods that use additional cues.

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 Ryan Farrell for his useful comments.

References

[1] J. Begard, N. Allezard, and P. Sayd. Real-time human detection in urban scenes: Local descriptors and classifiers selection with adaboost-like algorithms. In *CVPR Workshops*, 2008.

[2] S. Belongie, J. Malik, and J. Puzicha. Matching Shapes. In *ICCV 2001*, volume 1, pages 454–461 vol.1, 2001.

[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] Y.-T. Chen and C.-S. Chen. Fast human detection using a novel boosted cascading structure with meta stages. *Image Processing, IEEE Trans. on*, 17(8):1452–1464, 2008.

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

[6] P. Dollar, B. Babenko, S. Belongie, P. Perona, and Z. Tu. Multiple Component Learning for Object Detection. In *ECCV 2008*, pages 211–224, 2008.

[7] A. Ess, B. Leibe, and L. V. Gool. Depth and appearance for mobile scene analysis. In *ICCV*, October 2007.

[8] A. Ess, B. Leibe, K. Schindler, and L. Gool. A mobile vision system for robust multi-person tracking. *CVPR*, 2008.

[9] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. *CVPR*, pages 1–8, June 2008.

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

[11] Y. Ke and R. Sukthankar. PCA-SIFT: A More Distinctive Representation for Local Image Descriptors. In *CVPR 2004*, volume 2, pages 506–513, 2004.

[12] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR 2006*, pages 2169–2178, 2006.

[13] Z. Lin and L. S. Davis. A pose-invariant descriptor for human detection and segmentation. In *ECCV*, 2008.

[14] S. Maji, A. Berg, and J. Malik. Classification using intersection kernel support vector machines is efficient. In *CVPR*, June 2008.

[15] K. Mikolajczyk, C. Schmid, and A. Zisserman. Human detection based on a probabilistic assembly of robust part detectors. In *ECCV 2004*, volume I, pages 69–81, 2004.

[16] A. Mohan, C. Papageorgiou, and T. Poggio. Example-based object detection in images by components. *PAMI*, 23(4):349–361, 2001.

[17] Y. Mu, S. Yan, Y. Liu, T. Huang, and B. Zhou. Discriminative local binary patterns for human detection in personal album. In *CVPR 2008*, pages 1–8, June 2008.

[18] S. Munder and D. Gavrila. An experimental study on pedestrian classification. *PAMI*, 28(11):1863–1868, 2006.

[19] R. Rosipal and N. Kramer. Overview and recent advances in partial least squares. *Lecture Notes in Computer Science*, 3940:34–51, 2006.

[20] V. Shet, J. Neuman, V. Ramesh, and L. Davis. Bilattice-based logical reasoning for human detection. In *CVPR*, 2007.

[21] D. Tran and D. Forsyth. Configuration estimates improve pedestrian finding. In *NIPS 2007*, pages 1529–1536. MIT Press, Cambridge, MA, 2008.

[22] O. Tuzel, F. Porikli, and P. Meer. Human detection via classification on riemannian manifolds. In *CVPR*, 2007.

[23] 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.

[24] S. Wold, W. Johansson, and M. Cocchi. PLS - Partial Least-Squares Projections to Latent Structures. In H. Kubinyi, editor, *3D QSAR in Drug Design: Volume 1: Theory Methods and Applications*, pages 523–550. Springer Verlag, 1993.

[25] B. Wu and R. Nevatia. Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In *ICCV*, pages 90–97, 2005.

[26] B. Wu and R. Nevatia. Optimizing discrimination-efficiency tradeoff in integrating heterogeneous local features for object detection. In *CVPR 2008*, pages 1–8, June 2008.

[27] W. Zhang, G. Zelinsky, and D. Samaras. Real-time accurate object detection using multiple resolutions. In *ICCV*, 2007.

[28] Q. Zhu, M.-C. Yeh, K.-T. Cheng, and S. Avidan. Fast human detection using a cascade of histograms of oriented gradients. In *CVPR 2006*, pages 1491–1498, 2006.