Pedestrian Detection with Cascaded Part Model for Occlusion Handling

Yawar Rehman[1], Irfan Riaz[2], Fan Xue[3], Jingchun Piao[4], Jameel Ahmed Khan[5] and Hyunchul Shin[6]
Department of Electronics and Communication Engineering,
Hanyang University (ERICA Campus), South Korea
e-mail:{yawar[1], irfancra[2], fanxue[3], jameel[5]}@digital.hanyang.ac.kr, {kcpark1011[4], shin[6]}@hanyang.ac.kr

***Abstract*—Pedestrian detection in a crowded environment under occlusion constraint is a challenging task. We have addressed this task by exploiting the properties of a rich feature set, which gives almost all cues necessary for recognizing pedestrians. Such rich feature set results in higher dimensional feature space. We have used partial least square regression to map these higher dimensional features to a lower dimensional yet discriminative feature space. Part model is further applied to deal with occlusions. The proposed method gives the best reported results on INRIA pedestrian dataset with detection accuracy of 98% at 10^{-4} False Positives Per Window (FPPW) and a miss rate of 31.62% at 10^{-1} False Positives Per Image (FPPI). We have also demonstrated the effectiveness of our part model under partial and heavily occluded conditions. Our proposed system outperforms several state of the art techniques under various evaluation conditions of INRIA pedestrian database.**

Keywords-Pedestrian detection; occlusion handling.

I. INTRODUCTION

Recent advancements in computer vision show researchers interest in developing a system to detect pedestrians efficiently. Detecting pedestrian is a challenging problem and various methods have been proposed. The performance of the detector depends on how well the method works in complex environments such as crowded scenes, illumination variation, occlusion, and cluttering [1]. Extensive literature can be found on the problem of object detection. It all started with the revolutionary work of Viola and Jones (VJ) [2][3]. VJ used integral sums as features and developed adaboost as classifiers. VJ achieved 20% of miss rate at over 10 FPPI. Histogram of Oriented Gradients (HOG) [4] achieved the same miss rate at ~ 1 FPPI and more recent methods [5][6] at equivalent miss rate achieved under 10^{-1} FPPI (data obtained from [7][8]). With the passage of time, researchers have given importance to the rich feature sets. Wang [9] cascaded texture and HOG features and trained a linear Support Vector Machine (SVM) so that small feature blocks of SVM weights can handle occlusions. In contrast to Wang's approach (HogLbp), we combined gradient, texture and color features and trained the linear SVM in lower dimensional feature space using partial least squares regression. In addition to it, we have cascaded a part model instead of breaking the final feature vector into small feature blocks for occlusion handling. Felzenszwalb's [10] deformable part model (DPM) achieved two fold improvement over 2006 PASCAL best performance for pedestrian detection. Schwartz [11] solved the problem of human detection in reduced dimensional space. Their feature vector was composed of three concatenated features, i.e., Co-occurrence matrices for texture information, Histogram of Oriented Gradients (HOG) for gradient information, and color information. Concatenation of these three features resulted in a feature vector of 170,820 dimensions. Partial Least Square (PLS) regression was used to reduce high dimensional feature space into discriminative reduced dimensional feature space. Quadratic Discriminant Analysis (QDA) model was used for classification. Kembhavi [12] also tackled vehicle detection problem in reduced dimensional space. They captured the color properties of the vehicle and its surroundings through color probability maps. Gradient information of the vehicle was captured using HOG and the pair of pixels method was used to extract structural properties. Concatenation of all these features resulted in the final feature vector of 70,000 dimensions. PLS regression was used for lower dimensional feature space and QDA model was trained as a classifier for finding objects of interest. Wang [13] handled object tracking as a classification problem and worked it out in reduced dimension by creating different PLS subspaces. They proposed an adaptive appearance model, which used different subspaces to handle variation of poses, occlusion, and cluttering. Haj [14] used discriminative properties of PLS lower dimensional space to solve the problem of head pose estimation. The author also compared different dimensionality reduction approaches and the result obtained from PLS regression was reported the best.

Dollàr [15] proposed Aggregate Channel Features (ACF) and emphasized the importance of color features in the task of pedestrian detection. Author used boosted features for the task of pedestrian detection. Feature pool was created by using multiple channels such as gradient, intensity, and color features. Several first and second order features were calculated on a patch inside a detection window on different channels. Boosted classifier was trained as [2] on these features in order to classify the detection window while testing. Benenson [16] transferred computations from testing time to training time and proposed a pedestrian detector named "Very Fast". Lim [17] proposed a method "Sketch Tokens" to detect contour of the objects. Sketch tokens were used with ACF as additional features for pedestrian detection. Benenson [18] reported a set of experiments in the quest of strongest rigid detector, and proposed "Roerei" pedestrian detector. The detectors proposed in [16][17][18] followed the frame work of ACF for pedestrian detection and provided state of the art results. Our part model is also based on the structural design of [15].

The key contributions of the proposed method are the cascaded integration of the part model with the root model (which in terms suppresses the false positives and handle

occlusions) and the proposed formulation for switching between the root model and the part model (when occlusion hypothesis is verified) on the fly. The integration of both models help significantly in solving the occlusion cases and decreases the number of miss classifications, which improve the detection accuracy of the proposed system. And the switching formulation helps to commute the time for feature calculations of both models.

We demonstrate our proposed system on INRIA pedestrian database. INRIA pedestrian database was introduced by Dalal & Triggs [4] when their detector performed almost ideal on the first ever MIT pedestrian database. INRIA dataset is still not fully explored and rigorously used in pedestrian detection evaluation. It contains 2,416 training positive windows cropped from 614 frames and 1,126 testing positive windows cropped from 288 frames. Both windows and frames are included in INRIA database. Training and testing negative frames are provided separately in INRIA database. Our system achieved the accuracy of 91% at 10^{-5} false positive per window (FPPW), 98% at 10^{-4} FPPW [1] and a miss rate of 31.62% at 10^{-1} FPPI. Our system consists of two main models, Partial Least Square (PLS) model and Part model (PM). Partial Least Square is a dimension reduction technique, which emphasizes supervised dimension reduction. PLS is helpful in providing discriminative lower dimensional feature space and avoiding the calculations containing thousands of extracted features. Part model ensures the search of a subject (i.e., pedestrian) in parts rather than to be searched as a whole. PM is helpful in handling occlusions. We have designed our part model as was described in [15].

II. FEATURE EXTRACTION

We have used three types of features in PLS model, i.e., gradient features, texture features, and color features.

A. Gradient Features

The first and foremost features that we have added in our feature set are gradient features. It is due to the fact that the research in object detection, specifically in human detection has increased significantly after the advent of HOG feature descriptor [4]. HOG was dedicated to human detection and it also provided the best results of its time.

For computing gradient features, we have used heavily optimized implementation of [15][19][20][21], which is similar to that of [4]. An image window is divided into 8x8 pixel blocks and each block is divided into 4 cells of 4x4 pixels. 9 bin HOG features per cell was then calculated obtaining 36 dimensional features per block. Each block is L2 normalized, which resulted 4 different normalizations per cell. It is useful because it makes HOG descriptor illumination invariant. HOG also shows rotation invariant properties as long as rotation is within the bin size. Clipping value of histogram bin is set to 0.2 and trilinear interpolation is used for the placement of gradients into their respective bins.

B. Texture Features

The texture information provides better results particularly in case of face detection because of discriminative texture on face (i.e., eyes, nose, mouth, etc). Including texture information in the pedestrian feature set will tend the system towards improvement in terms of detection because of the fact that there is a considerable amount of discriminative texture inside human contour.

We have used Local Binary Pattern (LBP) [22] to estimate texture features. LBP is a simple yet efficient technique for calculating texture in an image. It assigns the value '1' in 3x3 pixel neighborhood if each pixel's intensity value in the neighborhood is greater than or equal to the center pixel's intensity value, '0' is assigned, otherwise. There are many variants of LBP but we have used the most stable one, which was reported to achieve good results by many authors. 3x3 neighborhood produces 256 possible binary patterns, which are too many for making reliable texture feature descriptor but in 256 possible binary patterns there exist total of 58 patterns, which exhibit at most two bit-wise transitions from '0' to '1' or from '1' to '0'. These patterns are known as uniform patterns. Using uniform patterns instead of 256 patterns will remarkably reduce the texture feature vector size with marginal decrease in performance [22]. We have used the implementation of uniform patterns as was given by [23]. An image window is divided into the blocks of 8x8 pixels and for each block a 58 texture feature descriptor is calculated. The final texture feature set is obtained by concatenating features obtained from several blocks.

C. Color Features

Color features play an important role in providing discriminative identities to objects. The dilemma is when talking about pedestrian detection, better recognition rates and efficiency by including color information is doubted by some researchers because of the variability in clothing color. Instead [11][15][24] showed the importance of color features in pedestrian detection.

We have taken the samples of pedestrians and non-pedestrians (i.e., non-humans) from INRIA database and converted into LUV color space. Our intuition of selecting LUV came from the result reported by [15], that LUV outperformed other color spaces by achieving an accuracy of 55.8% alone (i.e., not combined with other features) in pedestrian detection. PLS regression is applied on L, U, and V space separately. PLS regression components shows maximum inter-class and intra-class variance. Human contour can be seen as silhouette by plotting them. U space showed dominant (red) peak at head region in all three PLS components. It is because variance of the head region in an image with respect to the surrounding region was maximum. During experimentation, we tried to include only U space as color information, but accuracy has decreased. In our opinion, the decrease in accuracy was due to lack of color information, which also points to the fact that including color information plays a significant role in detection. We have exploited this by including LUV color space representation in our system.

The final feature vector reflecting different extracted information from an image window looks like:

$$F = [Gradient\ Texture\ Color] \quad (1)$$

III. PARTIAL LEAST SQUARES MODEL

We have accumulated rich feature set for all possible cues of pedestrians, which resulted in high dimensional feature space. In our experiments, the number of samples used for training the classifier are less than the dimension of rich feature space. The phenomenon when data dimensions remains greater than the number of samples is known as multicollinearity. Partial least squares regression addresses the problem of multicollinearity and reduces data dimensions. PLS regression uses class labels for producing latent components that makes lower dimensional space more discriminative. An idea of constructing latent variables is summarized here, for details reader is encouraged to refer [25][26].

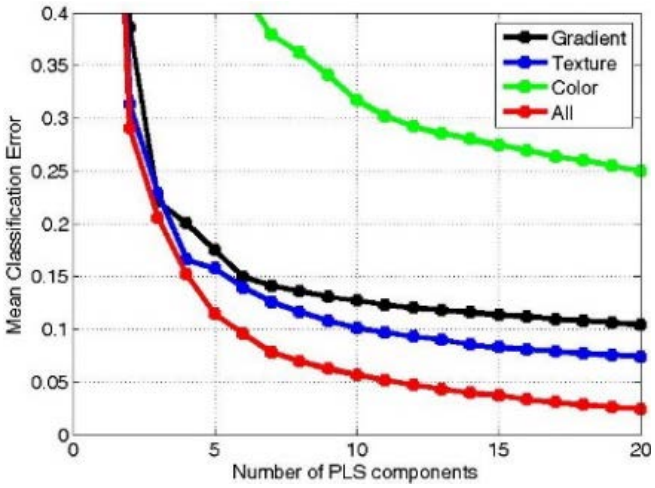

Figure 1. Mean square error vs PLS components

There are two popular variants of PLS, Non-iterative partial least square (NIPALS) and Simple partial least square (SIMPLS). They both differ in matrix deflation process. We used SIMPLS regression in our experiments. Let $X^{N \times m}$ and $Y^{N \times n}$ be the two blocks of variables. PLS models the relationship between the sets of variables by maximizing the covariance between them through latent variables.

$$X = TP^T + E \quad (2)$$

$$Y = UQ^T + F \quad (3)$$

Where T^{Nxp} and U^{Nxp} are score matrices; P^{mxp} and Q^{nxp} are loading matrices and E^{Nxm} and F^{Nxn} are residuals. The weight matrix in first iteration is calculated as,

$$w_1 = \bar{X}^T \bar{Y} / \|\bar{X}^T \bar{Y}\| \quad (4)$$

and till k^{th} iteration it is calculated as,

$$\bar{X}_k = \bar{X}_{k-1} - t_{k-1} p_{k-1}^T \quad (5)$$

Where t and p are the column vectors of matrix T^{Nxp} and P^{mxp}, respectively, and k represents the number of PLS factors. The dimension of an input image x is reduced by projecting its feature vector on to the weight matrix obtained after k iterations, where columns of $W=\{w_1,w_2,w_3,\ldots w_k\}$

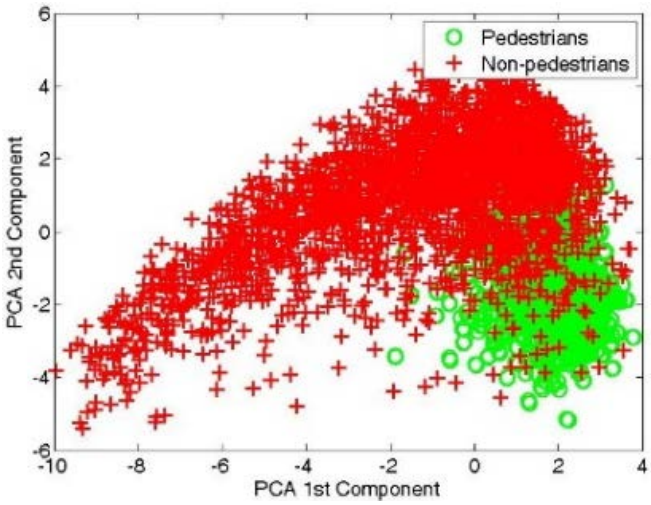

Figure 2. PCA lower dimensional space

represents PLS components. After projection, a low dimensional vector z^{1xk} is obtained.

Principal component analysis (PCA) is a well-known technique for dimension reduction. It also addresses multicollinearity problem, but does not consider class labels of data for dimension reduction. PLS is a supervised dimension reduction technique that considers class labels for dimension reduction. This enables PLS to produce highly discriminative reduced dimensional data as it is evident from Figures 2 and 3. We have plotted first two components of both dimension reduction techniques to show their discriminative power in lower dimensional space.

Our system extracts three cues from an image patch, which makes our high dimensional feature set. The total number of features extracted from an image patch are approximately fourteen thousand. With the help of PLS, we have reduced our feature set to only sixteen dimensions, which are the best representation of our high dimensional data. Figure 1 shows the mean classification error at different dimensions.

IV. PART MODEL

Part models are generally used in pedestrian detection to handle occlusions. It is a common practice to divide human body into five parts (i.e., head, left torso, right torso, upper limbs, and lower limbs) and detect each part separately. Deformation schemes were also introduced by several authors in order to keep different parts glued together. In our case, we have used upper body part model. The model includes head, left torso, and right torso.

We argue that, using upper body parts as a whole will give more discrimination among features because hardly any other object is represented with this structure. The structure of head, shoulders, arms, and torso (all connected) gives more discriminative feature property rather than to search them individually. Furthermore, to avoid complex deformation schemes [10], using only upper body as a part model is the best choice.

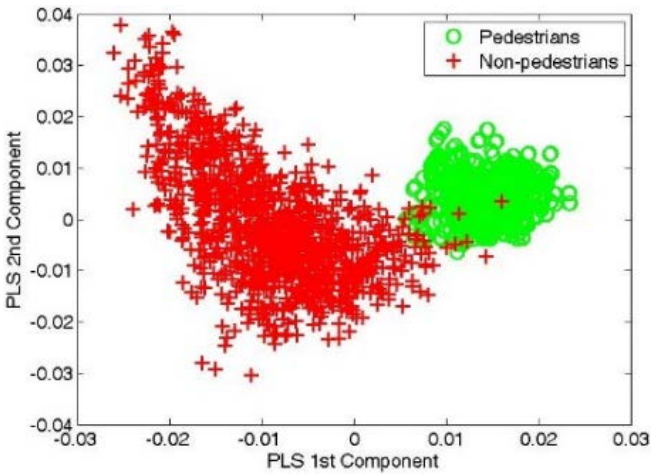

Figure 3. PLS lower dimensional space

P. Dollàr reported in [27] that over 53% of the pedestrians are occluded in some frames and 19% of the pedestrians are occluded in all frames. The author underscores the importance of detecting occluded pedestrians by reporting that over 70% of the pedestrians are occluded in all frames. Author further reports that 97% of the occlusion belongs to a small subset out of hundreds of possible occlusion types. The pedestrians in this subset are occluded from lower torso to limbs region, which also seconds our rationale of using upper body model for occlusion handling.

In order to design the upper body model, we have used the frame work of Aggregate Channel Features (ACF). For an input image I, we compute gradient histogram, gradient magnitude and color channels. After computing channels, we sum block of pixels to make the aggregate channels. Thus, the aggregate channels are the single pixel look up tables of the computed channels. We finally vectorize the aggregate channels and give them to the decision tree classifier to differentiate upper body from the background.

We have used total of ten channels, six gradient histogram channels, one gradient magnitude channel and three color channels of LUV color space as shown in Figure 4. Six gradient histogram channels contain the high pixel values of only those pixel, which lie in the respective span of the gradient angles, the values of remaining pixels are assigned zero. First gradient channel contains the high values of pixels that lie in the range of 0 ~ 30 degrees, second gradient channel contains the high values of pixels that lie in the range of 31 ~ 60 and so on. Six gradient histogram channels covers the span of 0 ~ 180 degrees. Magnitude channel contains the magnitude values of all the pixels in an input image I. Gradient magnitude channel basically gives the information of a sudden change in intensities or edges. Color channels contain the three channels of LUV color space. The reason of using LUV color space was discussed in Section II. The gradient angles and magnitudes were calculated by using the following equations.

$$\nabla I = \begin{bmatrix} g_x \\ g_y \end{bmatrix} = \begin{bmatrix} \partial I / \partial x \\ \partial I / \partial y \end{bmatrix} \approx \begin{bmatrix} I(x+1,y) - I(x,y) \\ I(x,y+1) - I(x,y) \end{bmatrix} \quad (6)$$

$$M(x,y) = \sqrt{g_x^2 + g_y^2} \approx |g_x| + |g_y| \quad (7)$$

$$\emptyset(x,y) = \tan^{-1}\left(\frac{g_y}{g_x}\right) \quad (8)$$

This type of channel frame work is often called modern HOG features with color. Since features are designed in such a way that they should be easier for the classifier to learn. It is a common practice to choose a classifier first, and based on the properties of that classifier features are designed, and not vice versa. Adding color features along with gradient histogram and gradient magnitude channel tends to improve the performance of detector, because of similar variance of color in face and hand regions. Final feature vector contains the information of gradients histogram, magnitude, and color from head, shoulders, left, and right torso. Discussion on the training of the classifier for part model is presented in Section V.

V. CLASSIFIERS

A. Linear Support Vector Machine

We have used linear support vector machine (SVM) for training our root model (i.e., full body detector) and for detection purpose the famous sliding window technique have been used. The sliding window technique checks for the object of interest at every possible location. Because we have trained our linear SVM on a fixed scale, so we made the pyramid of test image, to make our detector scale invariant and reduced the search problem to a binary classification problem. We trained our linear SVM classifier as was described in [4] and using a template size of 128 x 64 pixels.

For a robust binary classifier it is a common practice to use 'bootstrapping', which was introduced in [28]. The main idea behind this technique is to minimize the training data of negative images. First, a classifier is loaded with the cache of all positive and some negative examples and it is trained. Then the trained classifier is applied on the negative images from the natural dataset, i.e., the dataset does not contains object of interest. This is also known as "mining hard negatives". Classifier will produce some false positives in the current round, which are stored. Then the classifier is again trained with the cache of false positives in addition to its previous cache of positive and negative examples in the next round. This process may be repeated few times but over fitting should be avoided.

We have trained linear SVM with the help of afore mentioned discussion. We have loaded all the positive samples of the upper body and 5000 negative samples extracted randomly from the training negative images and positive windows provided in INRIA pedestrian database. We have set the bucket size of collecting negatives as 5000, and maximum negatives that can be collected as 10,000. In the 1st round of bootstrapping, our classifier has collected 5000 negative samples, making the total number of negatives 10,000. We have trained the classifier again with the positives and 10,000 negative samples, and ran it on the testing negative images. In the 2nd round of bootstrapping, our classifier has collected around 3000 negative samples, making the total number of hard negatives ~13,000. We

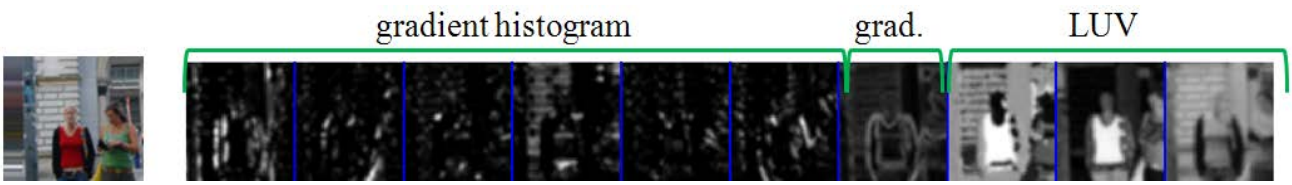

Figure 4. The computed gradient histogram, gradient magnitude and LUV color channels. The channels are computed on 128x64 pixel image. Six histogram channels, each containing high values of those pixels, which lie in the respective range. Gradient magnitude channel captures the edges as shown by the person upper body silhouette. L, U and V channel shows different shades of color in its domain. The input image on the left most is bigger in dimension and contains more background [15].

randomly choose 10,000 hard negatives out of 13,000 and train our classifier to initiate the next round of bootstrapping. In round 3, our classifier hardly collected ~100 negative samples. We stop the training of the classifier after three rounds of bootstrapping.

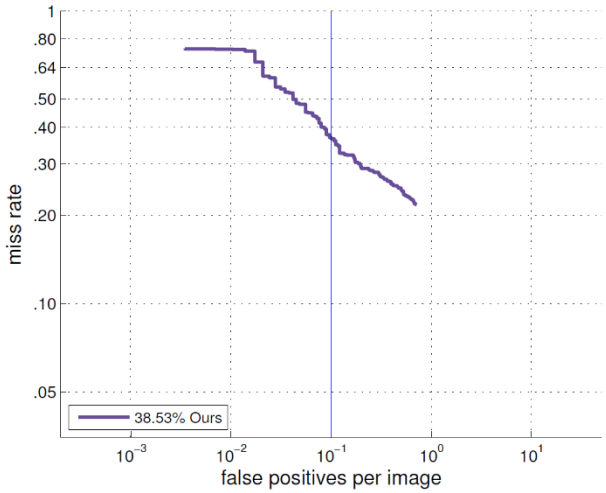

Figure 5. Missrate of our trained PLS person detector on INRIA dataset with 5000 bucket size and 10,000 total no. of negatives using FPPI evaluation metric.

It is necessary to fix the total number of negative that can be collected to avoid the over fitting problem. As SVM is a good memorizer, it over fits on the data when large number of negatives are added. An over fitted classifier will perform better on the training data but the accuracy of the same classifier will fall drastically if tested on a data other than training. We have fixed the total number of negative samples that can be collected to 10,000 and also introduced randomness in the selection of negative samples, which ensures that the classifier will only learn the features of object rather than to learn a particular pattern that yields high false positive rate in testing. We have achieved the miss rate of 38.53% at 10^{-1} FPPI with our PLS person detector on INRIA dataset as shown in Figure 5. The reason for reporting our results in FPPI evaluation metric and FPPW metric is discussed in the Section VII.

By increasing the total number negatives that can be collected to 12,000 we have achieved the miss rate of 38.42% at 10^{-1} FPPI as shown in Figure 6. By further altering the bucket size and increasing it to 6000, we have achieved the miss rate of 36.98% at 10^{-1} FPPI as shown in Figure 7.

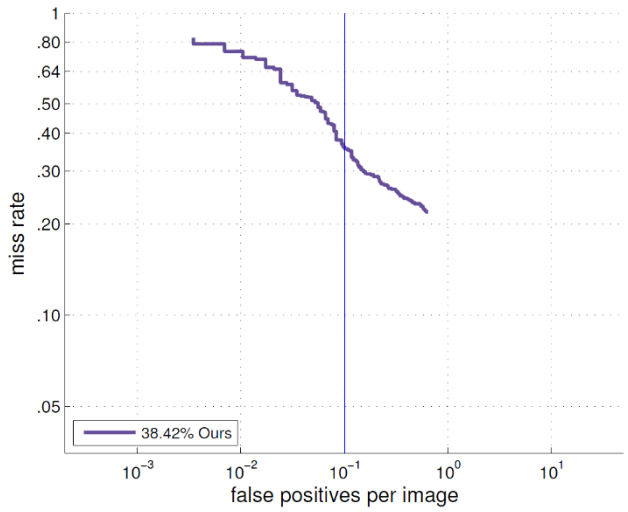

Figure 6. Missrate of our trained PLS person detector on INRIA dataset with 5000 bucket size and 12,000 total no. of negatives using FPPI evaluation metric.

B. *Adaboost*

We have trained our upper body model for handling occlusions using adaptive boosting classifier (adaboost). It is a combination of several weak learner that we add up gradually at each stage to make a strong classifier in the end. Among various variants of adaboost, we have used discrete adaboost classifier.

Discrete adaboost is a technique for constructing strong classifier as a combination of several weak classifiers. We have used depth-2 decision trees as our weak classifiers, where each node is a decision stump, which is defined by a rectangular region on the aggregated channels. We perform four rounds of training, in the first round we have loaded 2416 positive samples and 5000 randomly collected negative samples from INRIA pedestrian dataset. We apply bootstrapping in other three rounds and increased the number of weak classifiers in each round (100, 400, 1000, and 2500).

Total of 5000 negative samples were allowed to be added in each round of bootstrapping and the maximum number of negative sample that can be added was set to 15,000. With these settings, the result of our upper body model combined with the PLS root model on INRIA pedestrian dataset is shown in Figure 9.

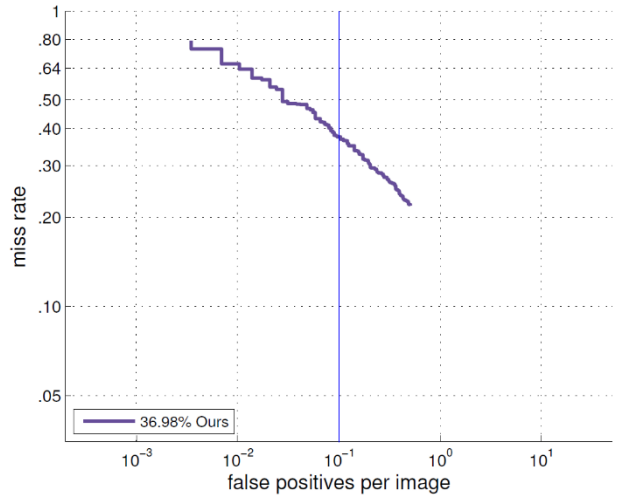

Figure 7. Missrate of our trained PLS person detector on INRIA dataset with 6000 bucket size and 12,000 total no. of negatives using FPPI evaluation metric.

C. Model switching

Our proposed system demands the use of two different types of features and classifiers. We can calculate both type of features of an input image and then apply both the classifiers on their respective features to get the results. This is a simple but computationally expensive, exhaustive and sluggish way of achieving our goal. What we needed was a system that can switch between the feature calculation and weights of two different models (i.e., PLS & PM) at any location or at any scale on a testing image. We designed a simple conditional model for this task as follows.

We consider that our PLS root model linear SVM calculates the score using the following equation.

$$Y = \beta . \varphi(x_p^s) + bias \quad (9)$$

Where 'β' represents SVM weight vector, 'x' represents the extracted features at a pyramid scale 's' and position 'p' from an input image, '$\varphi(.)$' represents PLS dimension reduction function, '$bias$' represents SVM bias term, and 'Y' is the calculated score by linear SVM.

We altered this equation in order to fulfil our needs of speed and to avoid useless computation. We introduced a variable '$flag$', which will decide what features to calculate and, which models to activate using the following equation.

$$Y = \left(\beta . \varphi\left(x_p^s\right) + bias\right) * flag \quad (10)$$

where

$$flag = \begin{cases} 0, & th1 < Y < th2 \\ 1, & th1 > Y > th2 \end{cases}$$

'$flag$' is a function of occlusion hypothesis and it can take only two values i.e., 0 or 1. We generate our occlusion hypothesis when the PLS root model classifier's score at scale 's' and position 'p' lies between thresholds '$th1$' and '$th2$'. Whenever this condition is satisfied, part model (upper body model) is activated and the aggregated channel features of the part model will be calculated at scale 's' and position 'p'. The decision of the part model classifier will be taken as the final decision. After that part model will be deactivated, and it will wait for the call of '$flag$' to again activate.

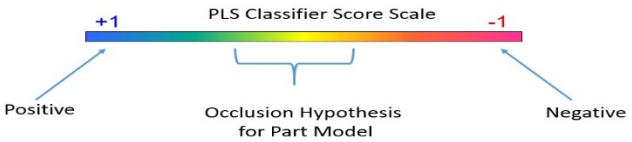

Figure 8. Heuristic for occlusion hypothesis

Combination heuristic of both models (PLS and part model) is discussed in Section VI.

VI. PLS + Part Model (Combined Model)

Our approach for combining both models is based on simple heuristic. We have trained our classifier for PLS model on lower dimensional space, which is very discriminative in nature. Linear SVM trained on lower dimensional data classifies efficiently and separates humans from non-humans almost accurately. Upon careful analysis, we came to know that the samples that were incorrectly classified by linear SVM either positives or negatives, their score lie in the vicinity of '0'. We generate our occlusion hypothesis that if a sample 'q' whose predicted score value 'v' lie between '$th1$' and '$th2$', then it is considered to be an occlusion and upon meeting this condition our part model will be activated and final score 'm' returned by part model will be taken as true value of the sample 'q'. The heuristic for occlusion hypothesis is shown in Figure 8. Figure 9 shows the miss rate of our combined model at 10^{-1} FPPI on INRIA pedestrian database.

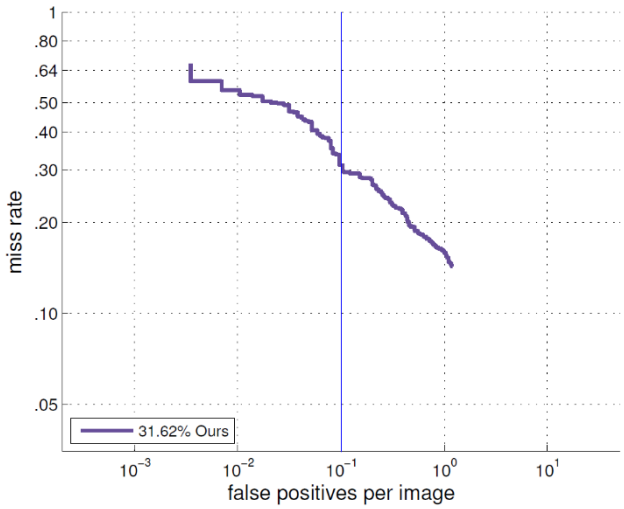

Figure 9. Missrate of our combined model at INRIA pedestrian dataset using FPPI evaluation metic.

VII. Evaluation Metrics

It is in common practice the results of object detection are reported in false positives per image (FPPI). As the INRIA dataset comes with test images as well as test windows, it will be unjust to report the results only in FPPI evaluation metrics. We have also reported our results in false positives per window (FPPW) metrics [1] as results in FPPW metrics were also reported by the dataset authors.

A. FPPW

Several runs of the trained classifiers (on HOG, FHOG & our method) were stored by varying decision threshold. Then the plot between false positive rate *(FPr)* and miss rate *(Mr)* is drawn by using the following equations.

$$FPr = FP/(FP + TN) \quad (11)$$

$$Mr = FN/(FN + TP) \quad (12)$$

Where *FPr* represents false positive rate, *FP* represents false positives, *TN* represents true negatives, *Mr* represents miss rate, *FN* represents false negatives and *TP* represents true positives. The comparison between HOG, variant of HOG (FHOG) introduced by [10], and our method in terms of FPPW, is shown in Figure 10. Each of the classifier was trained as described in Section V. Our system gives accuracy of90.5% at 10^{-5} false positive per window (FPPW) and accuracy of 98.1% at 10^{-4} FPPW. Testing was done on 1,126 positive cropped windows and 105,500 negative cropped windows from negative images provided by INRIA dataset.

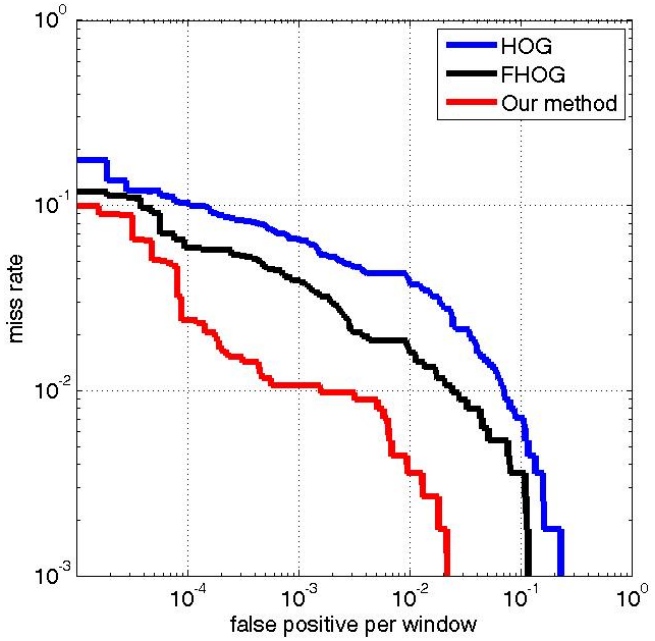

Figure 10. Comparison of our model with HOG and FHOG in FPPW evaluation metrics

B. FPPI

We report the results of our detector in FPPI evaluation metrics on INRIA dataset. INRIA dataset consists of 288 testing images containing pedestrians varying in sizes, postures and occlusions. Some of the images contains pedestrians with challenging posture, cluttered environment and crowded scenes, which tends to induce occlusions. Performance of our detector under these challenging conditions can viewed in Figure 11. For the FPPI evaluation on INRIA dataset, we have used the improved labelling proposed in [29]. In improved labelling, pedestrians are mainly divided into seven classes. Three of those classes deals with occlusions, other three of them deals with the pedestrian sizes and the remaining class includes all pedestrians irrespective of their occlusions and sizes. This labelling gives a good estimate of detector's performance in different conditions and scenarios. Table I shows the distribution of pedestrians according to their height in pixels and visibility.

TABLE I. DISTRIBUTION OF INRIA TESTING PEDESTRIAN SAMPLES

Type	Height (pixels)	Visibility (%)
Reasonable	50 ~ ∞	0.65 ~ ∞
Partially occ.	50 ~ ∞	0.65 ~ 1
Heavily occ.	50 ~ ∞	0.2 ~ 0.65
Large scale	100 ~ ∞	∞ ~ ∞
Near scale	80 ~ ∞	∞ ~ ∞
Medium scale	30 ~ 80	∞ ~ ∞
All	20 ~ ∞	0.65 ~ ∞

Height of a pedestrian in an image is inversely proportional to distance between camera and the pedestrian. This can be modeled by using the following equation [27].

$$h \approx H\ f/d \quad (13)$$

Where *h* represents the pixel height of pedestrians, *H* is the actual height, *f* is the focal length of the camera used for recording images and *d* is the distance between camera and the pedestrian. Assuming pedestrian actual height $H \approx 1.80$m, camera focal length $f \approx 1000$ we obtain $d \approx 1800/h$ m. A camera mounted on a vehicle moving with the speed of 15 m/s (≈55 km/h) on an urban road, a pedestrian of 100 pixels would be 18 m away and a pedestrian of 50 pixels would be 36 m away. It means that, if a pedestrian is standing 36 m away, it would take 2.4 seconds for a driver of a car moving at a speed of 55 km/h to react, which is reasonably good on an urban road. For the evaluation of our proposed detector, we fix the height range from 50 pixel to infinity, which is a reasonable choice under urban conditions. We have evaluated our detector's performance under different height and visibility range settings as shown in Table I. We have reported our results and its comparison with other techniques on INRIA dataset at 10^{-1} FPPI (as indicated with blue line). We have adopted evaluation code from [30] available online.

VIII. EXPERIMENTAL RESULTS

The comparison between HOG, variant of HOG (FHOG) introduced by [10], and our method is shown in Figure 10. Our system gives accuracy of 91% at 10^{-5} false positive per window (FPPW) and accuracy of 98% at 10^{-4} FPPW. Testing was done on 1,126 positive cropped windows and 105,500 negative cropped windows from negative images provided by INRIA dataset. According to the observations of [10], there are some cases in, which the use of light insensitive features will give benefit and in other cases the use of light sensitive features will give benefit. FHOG consists of 32 features. 13 of them are the representations of 36 HOG features in reduced dimensional space that are light insensitive features and remaining features are light sensitive features.

We can see in Figure 10 that FHOG clearly dominates HOG. On the other hand, our method achieved the best accuracy in comparison to HOG and FHOG. To our

knowledge, our system gives the best state of the art results at 10^{-4} FPPW [1] on INRIA pedestrian database. In our opinion, the reason for achieving the best results on INRIA dataset in FPPW evaluation metrics is that our system was able to solve occluded cases with high confidence values, which in case of other state of the art detectors either produced false negatives or they might have corrected those cases with a lower confidence value. The time cost of projecting high dimensional feature vector onto the weight matrix and lacking of vertical occlusion handling can be counted as the limitations of the proposed system.

We also report our results in FPPI evaluation metrics. Our proposed system clearly dominates other state of the art methods, yielding the lowest miss rate at 10^{-1} FPPI as shown in the following graphs.

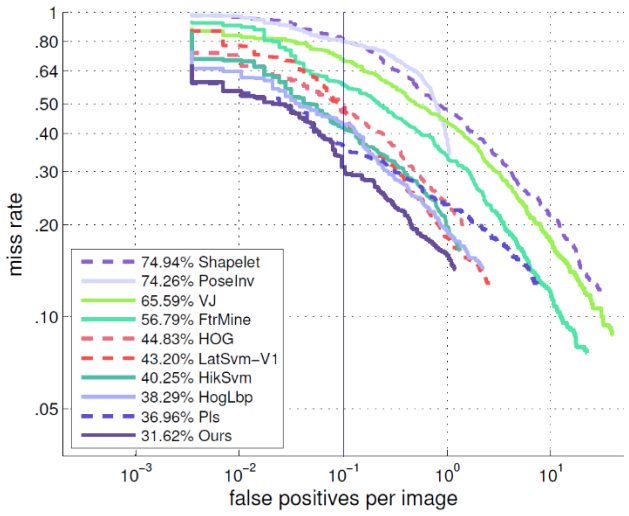

Figure 11. Miss rate vs FPPI under reasonable condition

As shown in Figure 11, our proposed detector achieved the miss rate of 31.62% at 10^{-1} FPPI and dominated other state of the art detectors in reasonable condition. Under reasonable conditions, only those pedestrians are considered for evaluation whose height starts from 50 pixels to as much as maximum height of the frame with a visibility range of 65% and higher.

Figure 12 shows our proposed detector achieved the miss rate of 38.79% at 10^{-1} FPPI and dominated other state of the art detectors in partial occlusion condition. Under partial occlusions, only those pedestrians are considered for evaluation whose height starts from 50 pixels to as much as maximum height of the frame with a visibility range from 65% to 100%.

Figure 13 shows our proposed detector achieved the miss rate of 69.60% at 10^{-1} FPPI and dominated other state of the art detectors except HOG-LBP detector in heavy occlusion condition. HOG-LBP detector uses texture features, and detects the whole pedestrian in chunks. Whereas, our proposed upper body detector neither uses texture features nor uses full body information for occlusion handling. Under heavy occlusions, only those pedestrians are considered for evaluation whose height starts from 50 pixels to as much as

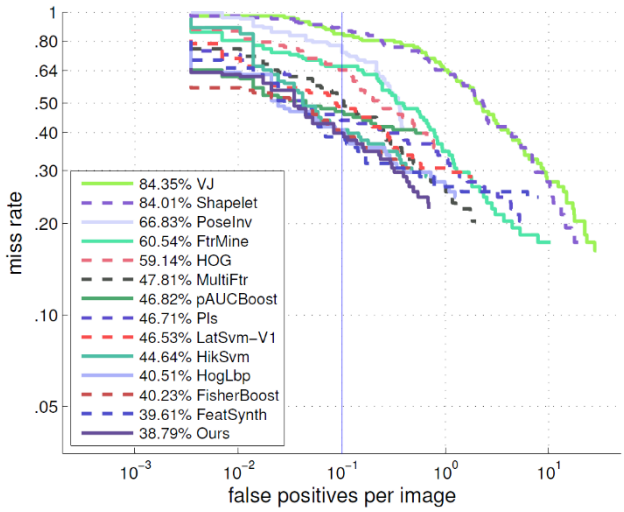

Figure 12. Miss rate vs FPPI under partial occlusions

maximum height of the frame with a visibility range from 20% to 65%.

Figure 14 shows our proposed detector achieved the miss rate of 21.38% at 10^{-1} FPPI and dominated other state of the art detectors in large scale pedestrian condition. Under large scale pedestrian condition, only those pedestrians are considered for evaluation whose height starts from 100 pixels to as much as maximum height of the frame with a visibility of 100%.

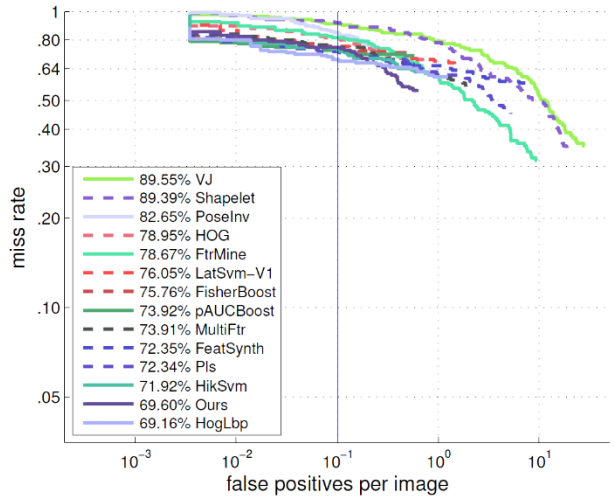

Figure 13. Miss rate vs FPPI under heavy occlusions

Figure 15 shows our proposed detector achieved the miss rate of 22.13% at 10^{-1} FPPI and dominated other state of the art detectors in near scale pedestrian condition. Under near scale pedestrian condition, only those pedestrians are considered for evaluation whose height starts from 80 pixels to as much as maximum height of the frame with a visibility of 100%.

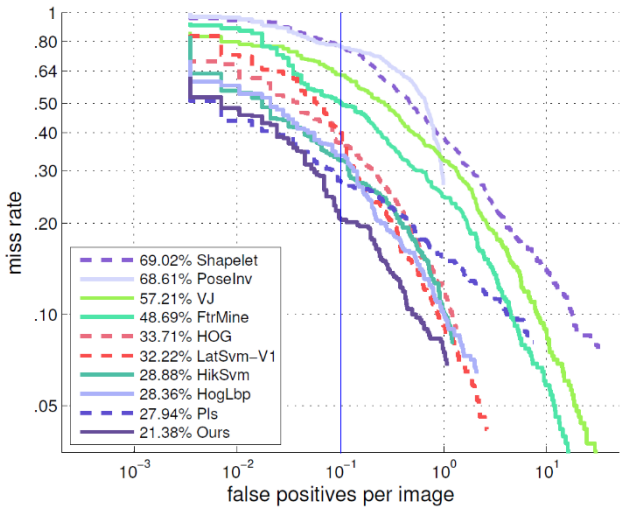

Figure 14. Miss rate vs FPPI of large scale pedestrians

Figure 16 shows our proposed detector achieved the miss rate of 61.11% at 10^{-1} FPPI and dominated other state of the art detectors in medium scale pedestrian condition. Results of only three detectors were available in medium scale pedestrian category. Under medium scale pedestrian condition, only those pedestrians are considered for evaluation whose height ranges from 30 pixels to 80 pixels with a visibility of 100%.

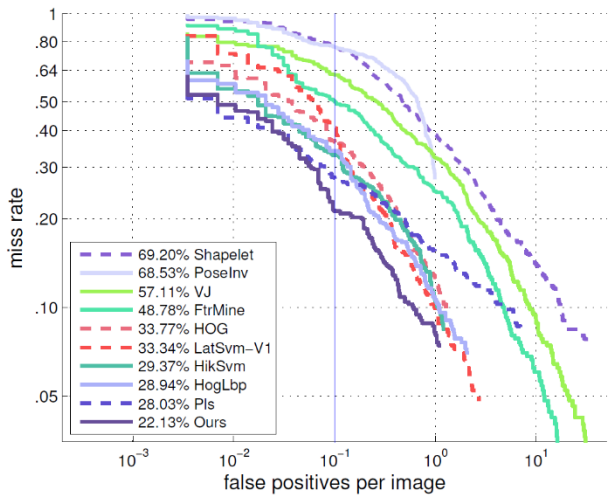

Figure 15. Miss rate vs FPPI of near scale pedestrians

Figure 17 shows our proposed detector achieved the miss rate of 45.57% at 10^{-1} FPPI and dominated other state of the art detectors in all condition. Under all condition, only those pedestrians are considered for evaluation whose height range is 20 pixels and higher with a visibility range of 65% and higher.

IX. CONCLUSION

We have developed a system that is capable of detecting pedestrians via monocular camera images efficiently. With the help of PLS, we are able to represent our rich feature set

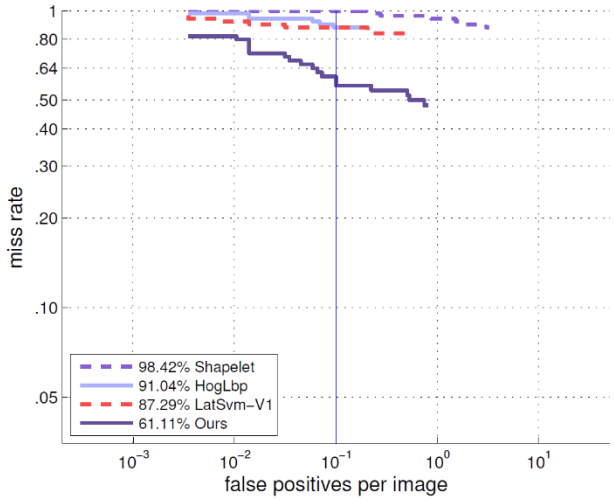

Figure 16. Miss rate vs FPPI of medium scale pedestrians

in more discriminative lower dimensional space. Part model is also integrated with our system for handling occlusions. We have achieved the detection rate of 98.1% at 10^{-4} FPPW and 31.62% miss rate at 10^{-1} FPPI on INRIA pedestrian database. We plan to further improve this detection rate in terms of FPPI by effectively adding another dimension of tracking and between-frames information into our system.

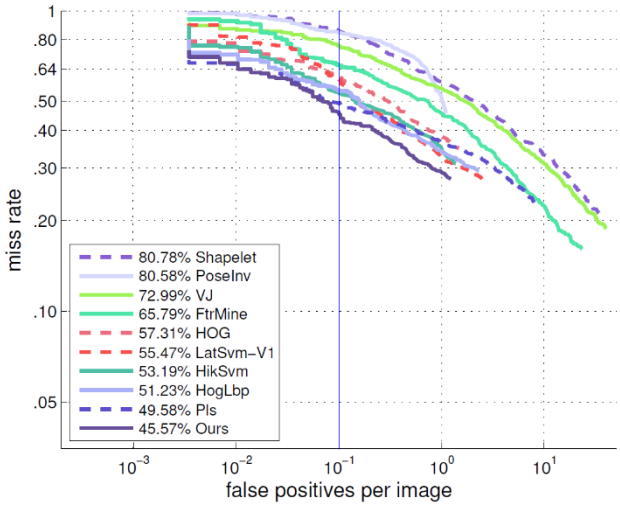

Figure 17. Miss rate vs FPPI of all condition

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea (NRF) Grant funded by the Korean Government (MOE) (No. NRF-2013R1A1A2004421). Moreover, the authors Yawar Rehman, Irfan Riaz and Jameel Ahmed Khan are sponsored by "Higher Education Commission" (HEC) of the Government of Pakistan.

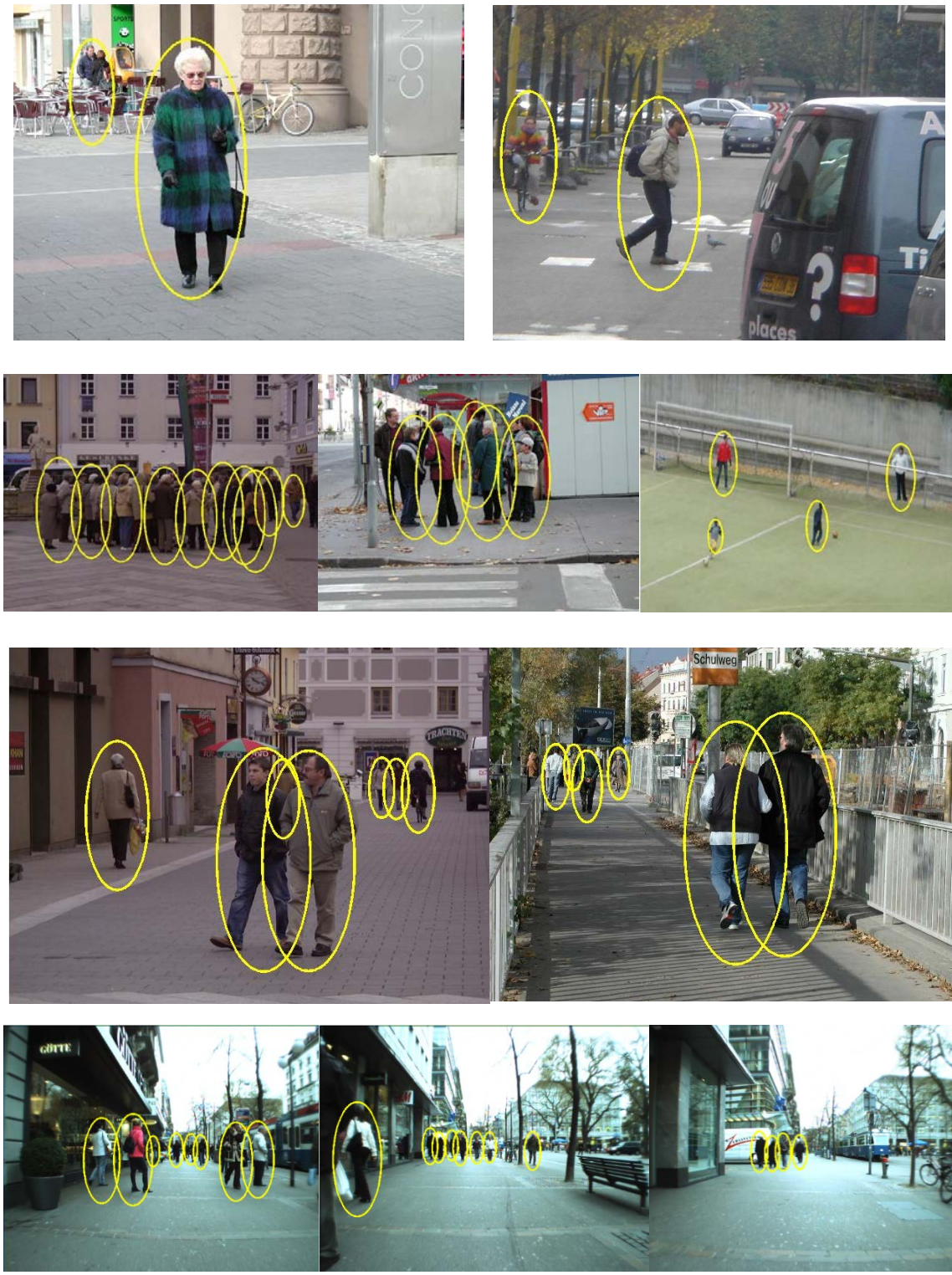

Figure 18. First three rows shows the results obtained from INRIA database and fourth row shows some results from ETH pedestrian database. The performance of our detector in occlusions, cluttered scenes, and pose variations, should be noted.

REFERENCES

[1] Y. Rehman, et al., “Pedestrian detection with occlusion handling,” in Proc. PATTERNS 2015, pp. 15-20.

[2] P. Viola and M. Jones, “Rapid Object Detection Using a Boosted Cascade of Simple Features,” Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2001.

[3] P. Viola, M. Jones, and D. Snow, “Detecting Pedestrians Using Patterns of Motion and Appearance,” Int’l J. Computer Vision 2005, vol. 63, no. 2, pp. 153-161.

[4] N. Dalal and B. Triggs, “Histograms of Oriented Gradients for Human Detection,” Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2005.

[5] P. Dollár, Z. Tu, P. Perona, and S. Belongie, “Integral Channel Features,” Proc. British Machine Vision Conf. (BMVC), 2009.

[6] R. Benenson, M. Mathias, R. Timofte, and L.V. Gool, “Pedestrian Detection at 100 Frames per Second,” Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2012.

[7] P. Dollár, C. Wojek, B. Schiele, and P. Perona, “Pedestrian Detection: An Evaluation of the State of the Art,”IEEE Trans. Pattern Analysis and Machine Intelligence 2012, vol. 34, no. 4, pp. 743–761.

[8] www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/, 2014.

[9] X. Wang, T. X. Han, and S. Yan, "An HOG-LBP human detector with partial occlusion handling," Computer Vision and Pattern Recognition (CVPR) 2009, pp. 32-39.

[10] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in Computer Vision and Pattern Recognition (CVPR) 2008, pp. 1-8.

[11] W.R. Schwartz, A. Kembhavi, D. Harwood, and L.S. Davis, "Human detection using partial least squares analysis," Computer Vision and Pattern Analysis (CVPR) 2009, pp. 24-31.

[12] A. Kembhavi, D. Harwood, and L.S. Davis, "Vehicle Detection Using Partial Least Squares," IEEE Trans. Pattern Analysis and Machine Intelligence 2011, vol.33, no.6, pp. 1250-1265.

[13] Q. Wang, F. Chen, W. Xu, and M.-H. Yang, "Object Tracking via Partial Least Squares Analysis," IEEE Trans. on Image Processing 2012, pp. 4454-4465.

[14] M.A. Haj, J. Gonzalez, and L.S. Davis, "On partial least squares in head pose estimation: How to simultaneously deal with misalignment," Computer Vision and Pattern Recognition (CVPR), 2012, pp. 2602-2609.

[15] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast Feature Pyramids for Object Detection," IEEE Trans. on Pattern Analysis and Machine Intelligence 2014, vol.36, no.8, pp. 1532-1545.

[16] R. Benenson, M. Mathias, R. Timofte, and L.V. Gool, "Pedestrian detection at 100 frames per second," in Computer Vision and Pattern Recognition (CVPR) 2012, pp. 2903-2910.

[17] J.J. Lim, C.L. Zitnick, and P. Dollár, "Sketch Tokens: A Learned Mid-level Representation for Contour and Object Detection," in Computer Vision and Pattern Recognition (CVPR) 2013, pp. 3158-3165.

[18] R. Benenson, M. Mathias, T. Tuytelaars, and L.V. Gool, "Seeking the Strongest Rigid Detector," in Computer Vision and Pattern Recognition (CVPR) 2013, pp. 3666-3673.

[19] P. Dollár, Z. Tu, P. Perona, and S. Belongie, “Integral Channel Features,” Proc. British Machine Vision Conf. 2009.

[20] P. Dollár, S. Belongie, and P. Perona, “Fastest Pedestrian Detector in the West,” Proceedings of BMVC, 2010.

[21] P. Dollár, R. Appel, and W. Kienzle, “Crosstalk Cascades for Frame-Rate Pedestrian Detection,” Proceedings of ECCV, 2012.

[22] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," IEEE Trans. On Pattern Analysis and Machine Intelligence 2002, vol.24, no.7, pp. 971-987.

[23] A. Vedaldi and B. Fulkerson, “VLFeat: An Open and Portable Library of Computer Vision Algorithms,” [Accessed from http://www.vlfeat.org], 2008.

[24] S. Walk, N. Majer, K. Schindler, and B. Schiele, "New features and insights for pedestrian detection," Computer Vision and Pattern Recognition (CVPR) 2010, pp. 1030-1037.

[25] R. Rosipal and N. Kramer, “Overview and recent advances in partial least squares in Latent Structures Feature Selection,” Springer Verlag, 2006.

[26] S. Wold, “PLS for Multivariate Linear Modeling QSAR,” Chemometric Methods in Molecular Design, 1994.

[27] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian Detection: An Evaluation of the State of the Art," IEEE Trans. on Pattern Analysis and Machine Intelligence 2012, vol.34, no.4, pp. 743-761.

[28] S. Kah-Kay and T. Poggio, "Example-based learning for view-based human face detection," IEEE Trans. on Pattern Analysis and Machine Intelligence 1998, vol.20, no.1, pp. 39-51.

[29] M. Taiana, J. Nascimento, and A. Bernardino, “An Improved Labelling for the INRIA Person Data Set for pedestrian Detection,” Pattern Recognition and Image Analysis 2013, vol. 7887, pp. 286-295.

[30] Accessed from, www.vision.caltech.edu/, 2015.