Pedestrian Detection with 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 rich feature set which gives almost all cues necessary for recognizing pedestrians. Using rich feature set results in higher dimensional space. We have used partial least square regression to work with more discriminative (lower dimensional) features than (higher dimensional) rich feature set. Part model is further applied to deal with occlusions. Our proposed method gives the accuracy of 98% at 10^{-4} false positives per window on INRIA pedestrian database, which is the best result reported so far, under the same false positives per window.**

Keywords-Pedestrian detection; occlusion handling.

I. INTRODUCTION

Recent advances 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. Extensive literature can be found on the problem of object detection [1]. W. R. Schwartz et al. [2] solved the problem of human detection in reduced dimensional space. The Information in the 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 designed for classification. A. Kembhavi et al. [3] also tackled vehicle detection problem in reduced dimensional space. Color properties of the vehicle and its surroundings were captured by using their method named color probability maps. Gradient information of an object was captured using HOG. 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. Q. Wang et al. [4] handled object tracking as a classification problem and worked it out in reduced dimension by creating different PLS subspaces. A model named adaptive appearance model was proposed which used different subspaces to deal with variation of poses, occlusion, and cluttering problems. M. A. Haj et al. [5] used discriminative properties of PLS lower dimensional space to solve the problem of head pose estimation. Different approaches were compared by [5] and the result obtained from PLS regression was reported the best. X. Wang et al. [6] developed a system which can detect pedestrians and also handles partial occlusion. The final feature vector comprises of gradient and texture features. Instead of using whole feature vector the author used contribution of chunks of feature blocks for the final decision. Linear SVM was used as a classifier and its decision functions were translated in terms of small feature blocks. Bias of the linear SVM decision function was learned from the training database under the constant bias distribution scheme proposed by the authors. The system was evaluated on INRIA pedestrian database and provided the state of the art results. P. Dollàr et al. [8] 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 [9] on these features in order to classify the detection window while testing.

Our key contribution includes, the heuristic based integration of the part model with the root model. This integration of both models helped significantly in solving the occlusion cases and decreased the number of false negatives and false positives. This decrease tends to improve the efficiency of the system.

We demonstrate our proposed system on INRIA pedestrian database. INRIA pedestrian database was given by [1] 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) and of 98% at 10^{-4} FPPW. 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.,

 ISBN: 978-1-61208-393-3

pedestrian) in parts rather than to be searched as a whole. PM is helpful in handling occlusions.

II. FEATURE EXTRACTION

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

A. Gradient Features

The first and foremost features which 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 [1]. 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 [8][10][11][12] which is similar to that of [1]. 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 considerable amount of discriminative texture inside human contour.

We have used Local Binary Pattern (LBP) [13] 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 [13]. We have used the implementation of uniform patterns as was given by [14]. 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 [2][8][15] 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 [8], 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 which makes lower dimensional space more discriminative. An idea of constructing latent variables is summarized here, for details reader is encouraged to refer [16][17].

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.

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

 ISBN: 978-1-61208-393-3

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

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 $W=\{w_1,w_2,w_3,\ldots w_k\}$ 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 doesn't consider class labels of data for dimension reduction. PLS is a supervised dimension reduction technique which 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. The performance of our system at sixteen lower dimensional features can be observed in Figure 5.

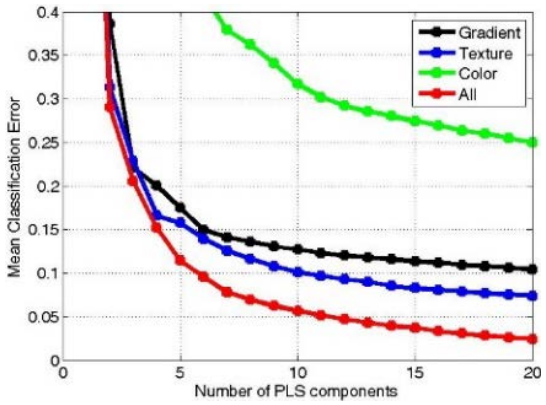

Figure 1. Mean square error vs PLS components

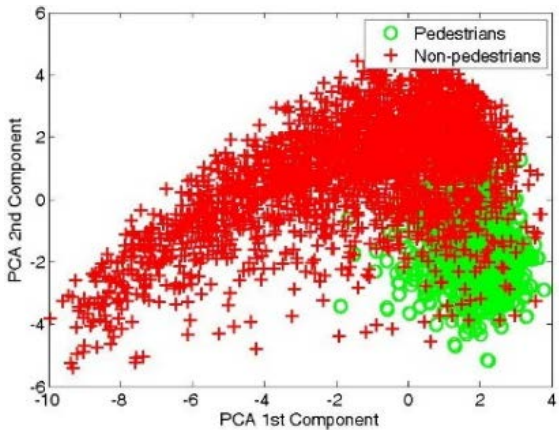

Figure 2. PCA lower dimensional space

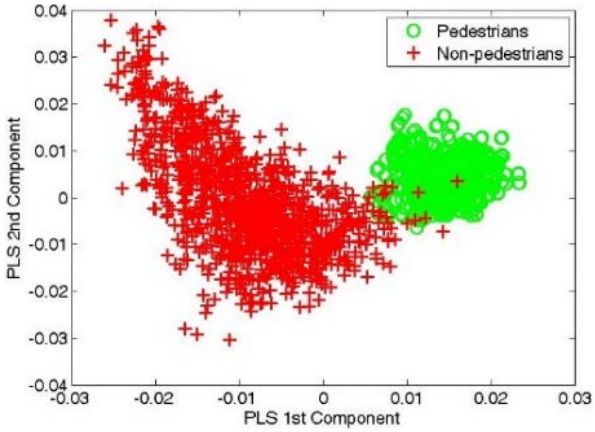

Figure 3. PLS lower dimensional space

Our PLS model gives the accuracy of 94% at 10^{-5} FPPW and accuracy of 96% at 10^{-4} FPPW on INRIA pedestrian database.

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 of the 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 [7], using only upper body as a part model is the best choice.

The upper body part model is designed using gradient, texture, and color features. Adding color features tends to improve the performance of detector, because of similar

 ISBN: 978-1-61208-393-3

variance of color in face and hand regions. Final feature vector contains the information of gradients, texture, and color from head, shoulders, left, and right torso. The performance of our part model on INRIA pedestrian database is shown in Figure 6.

V. 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 which 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 4.

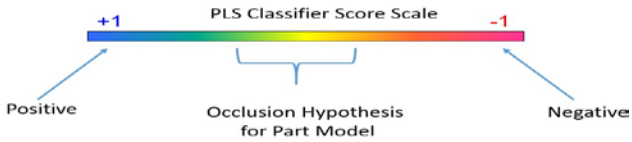

Figure 4. Heuristic for occlusion hypothesis

Figure 7 shows the performance of all our models on INRIA pedestrian database.

VI. EXPERIMENTAL RESULTS

The comparison between HOG, variant of HOG (FHOG) introduced by [7], and our method is shown in Figure 8. Each of the classifier was trained as described in Section 3. 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 [7], 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 which are light insensitive features and remaining features are light sensitive features. As we can see in Figure 8 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 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 corrected that case 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.

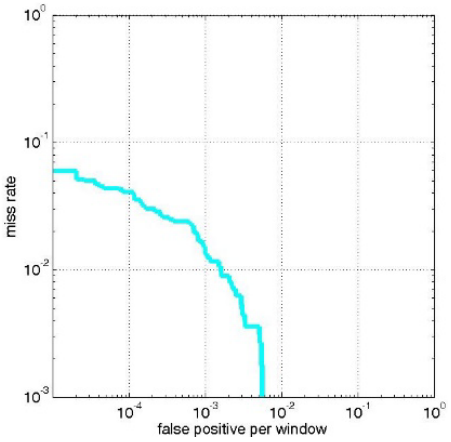

Figure 5. Performance of PLS model

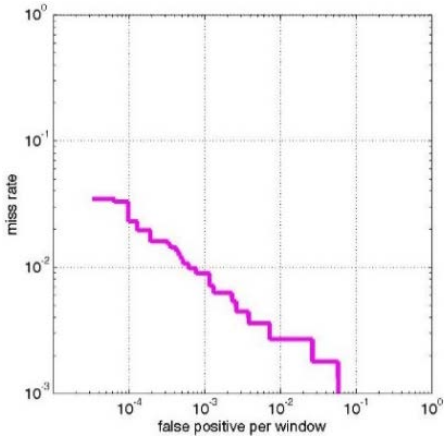

Figure 6. Performance of Part model

Table 1 translates Figure 8 by indicating maximum accuracy achieved by HOG, FHOG and the proposed method. Figure 9 shows some sample results of our proposed system.

VII. CONCLUSION

We have developed a system which is capable of detecting human via monocular camera images efficiently. With the help of PLS, we are able to represent our rich feature set 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 which is the best result reported on INRIA pedestrian database.

 ISBN: 978-1-61208-393-3

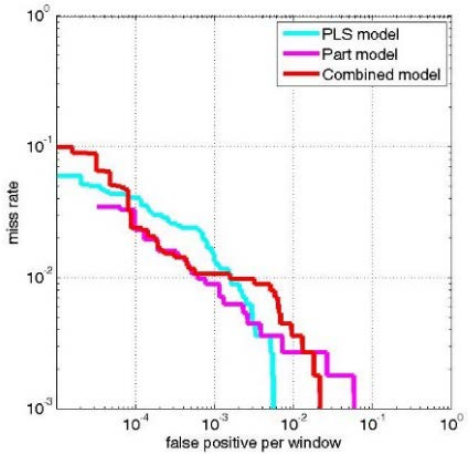

Figure 7. Performance of all models

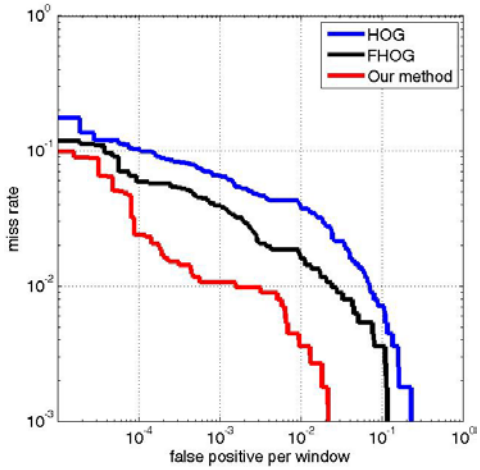

Figure 8. Comparison of our model with HOG and FHOG

TABLE I. (1-MISSRATE) AT FPPW VALUES

1-missrate	*10^{-5} FPPW*	*10^{-4} FPPW*
HOG	82.5 %	90.0 %
FHOG	88.2 %	94.1 %
Ours	90.5 %	98.1 %

We plan to further improve this detection rate by effectively adding another dimension of tracking and between-frames information into our system.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea (NRF) Grant funded by the Korean Government (MOE) (No. NRF-2013R1A1A2004421).

REFERENCES

[1] Dalal, N., Triggs, B., "Histograms of oriented gradients for human detection," Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, June 2005, pp.886-893 vol. 1, 25-25.

[2] Schwartz, W.R.; Kembhavi, A.; Harwood, D.; Davis, L.S., "Human detection using partial least squares analysis," Computer Vision, 2009 IEEE 12th International Conference on, Sept. 29 2009-Oct. 2 2009, vol., no., pp.24-31.

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

[4] Qing Wang; Feng Chen; Wenli Xu; Ming-Hsuan Yang, "Object Tracking via Partial Least Squares Analysis," Image Processing, IEEE Transactions on , Oct. 2012, vol.21, no.10, pp.4454-4465.

[5] Haj, M.A.; Gonzalez, J.; Davis, L.S., "On partial least squares in head pose estimation: How to simultaneously deal with misalignment," Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, June 2012, vol., no., pp.2602-2609, 16-21.

[6] Xiaoyu Wang; Han, T.X.; Shuicheng Yan, "An HOG-LBP human detector with partial occlusion handling," Computer Vision, 2009 IEEE 12th International Conference on , vol., no., pp.32-39.

[7] Felzenszwalb, P.F.; Girshick, R.B.; McAllester, D.; Ramanan, D., "Object Detection with Discriminatively Trained Part-Based Models," Pattern Analysis and Machine Intelligence, Sept. 2010, IEEE Transactions on , vol.32, no.9, pp.1627-1645.

[8] P. Dolla´r, Z. Tu, P. Perona, and S. Belongie, “Integral Channel Features,”, Proc. British Machine Vision Conf. 2009.

[9] P. Viola and M. J. Jones, “Robut Real-Time Face Detection”, International Journal of Computer Vision, 2004, Vol. 57, No. 2.

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

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

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

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

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

[15] Walk, S.; Majer, N.; Schindler, K.; Schiele, B., "New features and insights for pedestrian detection," Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, vol., no., pp.1030-1037, 13-18.

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

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

 ISBN: 978-1-61208-393-3

Figure 9. 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.

 ISBN: 978-1-61208-393-3