VISUAL OBJECT DETECTION BY PARTS-BASED MODELING USING EXTENDED HISTOGRAM OF GRADIENTS

Amit Satpathy[1,2] *Xudong Jiang*[1] *How-Lung Eng*[2]

[1]Nanyang Technological University
School Of EEE
Nanyang Avenue, Singapore 639798
{amit0010,exdjiang}@ntu.edu.sg

[2]Institute for Infocomm Research
Agency for Science, Technology and Research
1 Fusionopolis Way, Connexis, Singapore 138632
hleng@i2r.a-star.edu.sg

ABSTRACT

In this paper, we present a parts-based modeling framework using Extended Histogram of Gradients (EXHOG) for *object detection*. Visual object detection is a challenging issue in computer vision where objects need to be detected in varying illumination and contrast environments. Furthermore, objects belonging to the same class exhibit large intra-class variations. Here, we propose using EXHOG with the discriminatively trained deformable part models of Felzenszwalb *et. al.* [1]. This framework is based on mixtures of multiscale deformable part models. EXHOG is a novel feature proposed earlier for the purpose of human detection and has shown promising results against other state-of-the-art approaches. The proposed approach is tested on INRIA Human data set and the PASCAL VOC 2007 data set. Results demonstrate superior performance on INRIA compared to existing state-of-the-art approaches and improved performance on PASCAL VOC 2007.

Index Terms— extended histogram of gradients, parts-based, pascal, object detection, feature extraction

1. INTRODUCTION

Object detection has been the focus of researchers for years in computer vision. One of the main challenges in developing a reliable object detection method arises from the large intra-class variations. Objects within the same class often have large variations in appearance. Furthermore, objects need to be detected in different illumination and contrast environments. An object detection approach should be able to represent objects in a way that minimizes intra-class and maximizes inter-class variations.

Object detection approaches can be divided into two groups - approaches that use holistic feature representation and approaches that use parts-based feature representation. Holistic feature representation typically include dense overlapping sampling of features at various locations and scales of the image which are then concatenated to represent the *whole* object. Various holistic feature representations such as raw pixel intensities [2], dense Scale-Invariant Feature Transform (SIFT) [3, 4, 5], Wavelet [6], Haar-like features [7], Histogram of Oriented Gradients (HOG) [8, 9], Extended Histogram of Gradients (EXHOG) [10, 11], Feature Context [12], Local Binary Pattern (LBP) and its variants [13, 14, 15, 16], Local Ternary Pattern (LTP) [17], Geometric-blur [18] and Local Edge Orientation Histograms [19] have been proposed over recent years. Holistic feature representations underperform when there are heavy occlusions or clutter in the image. Furthermore, they cannot fully represent the variable deformations of objects.

In order to alleviate the issues of holistic feature representations, parts-based feature representations have been presented in several works. Some use deformable part-template models [20, 21, 22]. Others use geometric constraints to localize parts of objects for feature representation [23, 24, 25]. The discriminatively trained deformable part models of Felzenszwalb *et. al.* [1] remains as one of the popular state-of-the-art approaches for parts-based object detection. The approach uses a mixture of star-structured part-based models defined by a "root" filter and a set of part filter and deformation models. A latent Support Vector Machine (SVM) is then used for training and testing. The features used in their work are "analytic" versions of HOG and HG.

In earlier works [10, 11], EXHOG, in a holistic feature representation, was proposed for human detection to alleviate the issues of Histogram of Gradients (HG) and Histogram of Oriented Gradients. It did so by considering both the sum and absolute difference of HG with the opposite gradients. In comparison to other holistic feature representations, EXHOG consistently performed better. However, in the presence of occlusions and variability of human articulations, it performed sub-optimally in comparison to parts-based feature approaches [1, 26]. Furthermore, the application of EXHOG was only restricted to human detection.

In this paper, we adopt the framework of [1] to do a parts-based modeling using EXHOG for object detection. The HOG feature in [1] is replaced with EXHOG and a discriminative deformable part model is then trained using the latent SVM. The contribution of the paper is two-fold: 1) The mixture of star-structured part-based models, defined by a "root" filter and a set of part filter and deformation models, is used for parts-based feature representation using EXHOG. Hence, the problems associated with holistic EXHOG representation is alleviated; 2) The application of EXHOG is extended to detection of other objects besides humans. Therefore, the generality of EXHOG is demonstrated and is shown to be not only limited to the detection of humans. Experiments on INRIA Human data set and PASCAL VOC 2007 object detection demonstrate superior performance over other approaches.

2. PARTS-BASED EXTENDED HISTOGRAM OF GRADIENTS REPRESENTATION

2.1. Extended Histogram of Gradients

HG differentiates a situation whereby a bright object is against a dark background or a dark object is against a bright background. This differentiation makes the intra-class variation of objects larger. HOG alleviates this problem by considering the gradients of direction θ

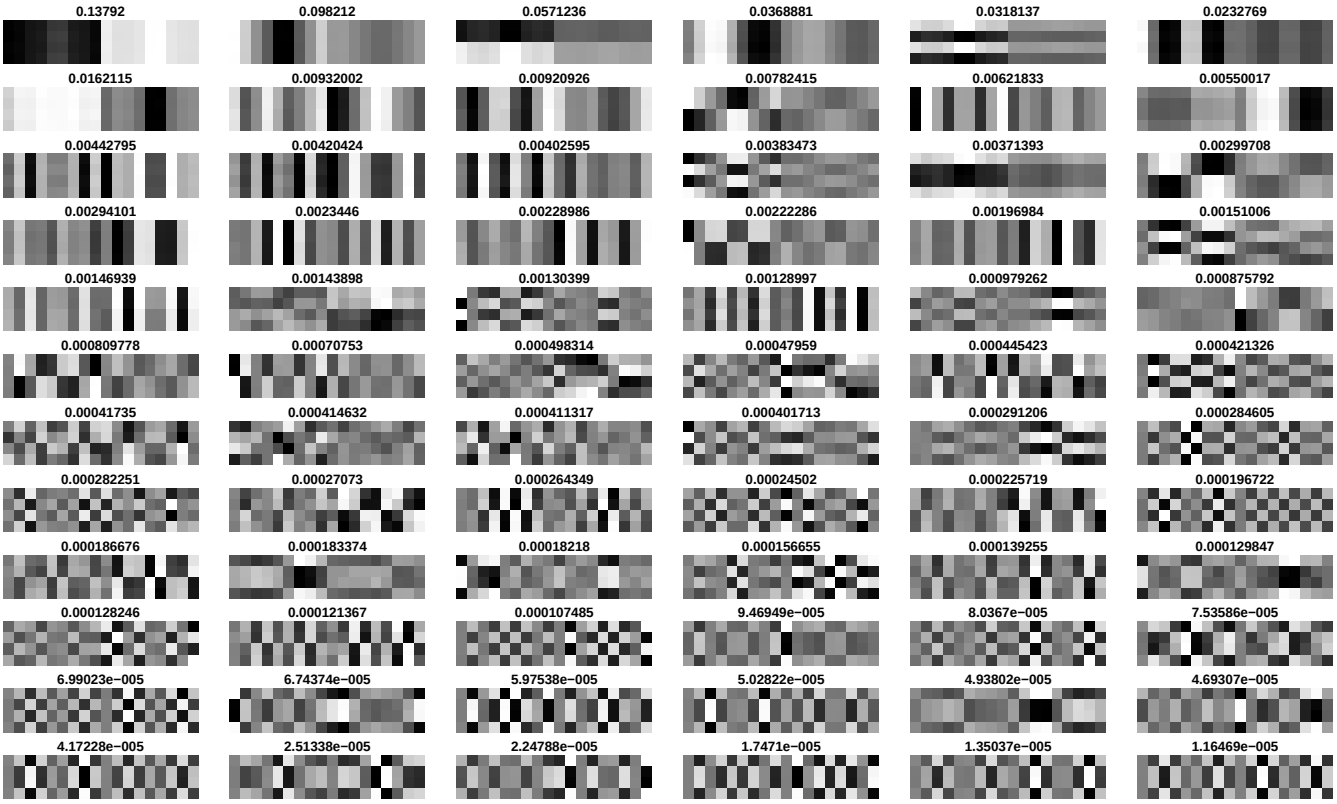


Fig. 1: PCA of ExHoG features. Each eigenvector is displayed as a 4×18 matrix so that each row corresponds to one normalization factor and each column to one orientation bin. The eigenvalues are displayed on top of the eigenvectors.

and $\theta + 180^\circ$, $0^\circ \leq \theta < 180^\circ$, to be of the same orientation θ only. In order to solve the problem of HG, HOG treats all gradients of opposite directions as gradients of a same orientation. However, this causes HOG to be unable to discern some local structures that are different from each other as gradients of opposite directions *in the same cell* are mapped to the same bin in a histogram. It is possible for 2 different structures to have similar feature representation.

In order to solve both issues of HG and HOG, Extended Histogram of Gradients (EXHOG) [10, 11] has been proposed. The steps in creating EXHOG are explained as follows. The HG block feature is created as in [8] (without normalization) and is normalized as follows:

$$h_{gn_k}(i) = \frac{h_{g_k}(i)}{\sqrt{\sum_{k=1}^{N}\sum_{i=1}^{L}\left(h_{g_k}(i)\right)^2}} \tag{1}$$

where $h_{g_k}(i)$ is the i^{th} bin value of HG from [8] without normalization.

$$h_{gc_k}(i) = \begin{cases} h_{gn_k}(i), & h_{gn_k}(i) < T \\ T, & h_{gn_k}(i) \geq T, \end{cases} \tag{2}$$

$$h_{gcn_k}(i) = \frac{h_{gc_k}(i)}{\sqrt{\sum_{k=1}^{N}\sum_{i=1}^{L}\left(h_{gc_k}(i)\right)^2}} \tag{3}$$

where N is the number of cells in the block and T is the clipping threshold.

HOG is created from $h_{gcn_k}(i)$ as follows:

$$h_{og_k}(i) = \quad h_{gcn_k}(i) + h_{gcn_k}(i + \tfrac{L}{2}), \quad 1 \leq i \leq \tfrac{L}{2} \tag{4}$$

where $h_{og_k}(i)$ is the i^{th} bin value of HOG.

Difference of HG (DHG) is created as follows:

$$h_{dg_k}(i) = \quad |h_{gcn_k}(i) - h_{gcn_k}(i + \tfrac{L}{2})|, \quad 1 \leq i \leq \tfrac{L}{2} \tag{5}$$

where $h_{dg_k}(i)$ is the i^{th} bin value of DHG. DHG produces the same feature as HOG for patterns that contain no gradients of opposite directions. It differentiates these patterns from the ones that contain opposite gradients by assigning small values to the bins that the gradients are mapped to.

The concatenation of these 2 histograms produces EXHOG as follows:

$$h_{eg_k}(i) = \begin{cases} h_{gcn_k}(i) + h_{gcn_k}(i + \frac{L}{2}), & 1 \leq i \leq \frac{L}{2} \\ |h_{gcn_k}(i) - h_{gcn_k}(i - \frac{L}{2})|, & \frac{L}{2} + 1 \leq i \leq L, \end{cases} \tag{6}$$

where $h_{eg_k}(i)$ is the i^{th} bin value of ExHoG.

2.2. PCA and Analytic Dimensionality Reduction

In [1], Principal Component Analysis (PCA) [27] was performed on the HOG cell features. It was discovered that the top few eigenvectors had some special properties that allowed for some dimension reduction without loss of information. Since HOG is being replaced with EXHOG in this work, PCA is also performed to analyze the top few eigenvectors. The eigenvectors are displayed in Fig. 1.

EXHOG is defined using four different normalizations of a 18-dimensional histogram over orientations for each cell. Therefore, it can be viewed as 4×18 matrix for a cell. The top 28 eigenvectors in Fig. 1 exhibit a special structure similar to the ones in [1]: They are each (approximately) constant along each row or column of their matrix representation. Hence, they lie (approximately) in a linear subspace defined by sparse vectors that have ones along a single row

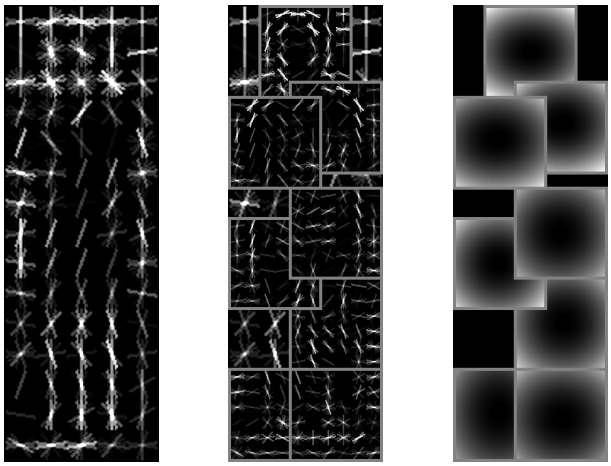


Fig. 2: One-component model with 8 parts learned using INRIA. The model is defined by a coarse root filter (left most), several higher resolution part filters (middle) and a spatial model for the location of each part relative to the root (right most). The visualization of the spatial models reflects the cost of placing the center of a part at different locations relative to the root.

or column of their matrix representation. Therefore, the analytic dimensionality reduction as defined in [1] can be used to reduce the dimensions of ExHoG in a similar manner.

In [1], a concatenation of Hg and Hog features was used. In our implementation, ExHoG is concatenated with Hg. Hence, for each cell, following the analytic dimensionality reduction in [1], the cell feature is recomputed by summing the 4 histograms of ExHoG and Hg and concatenating the four normalization factors (of Hg) to form a 40-dimensional vector. An ExHoG feature pyramid is then defined by computing ExHoG features of each level of a standard image pyramid. Features at the top of the pyramid capture coarse gradients histogrammed over fairly large areas of the input image while features at the bottom of the pyramid capture finer gradients histogrammed over small areas.

2.3. Parts-based representation

The parts-based representation adopted for ExHoG is described in brief as follows. For complete details on the representation, readers are referred to [1]. Every object is represented by a mixture of models, $M = (M_1, ..., M_m)$ where M_c is the model for the c-th component and m is the number of components. A singular object model consists of a global "root" filter (which covers the entire object) and several higher resolution part models covering smaller parts of the object. Each part model specifies a spatial model and a part filter. The spatial model defines a set of allowed placements for a part relative to a detection window, and a deformation cost for each placement. All of the root and part models involve linear filters which are rectangular templates specifying weights for subwindows of an ExHoG pyramid. They are applied to a dense ExHoG array whose entries are feature vectors computed from a dense grid of locations in an image.

Let F be a $w \times h$ filter. Let E be a ExHoG feature pyramid and $p = (x, y, l)$ specify a position (x, y) in the l-th level of the pyramid. Let $\phi(E, p, w, h)$ denote the vector obtained by concatenating the feature vectors in the $w \times h$ subwindow of E at p. The score of F at p is $F' \cdot \phi(E, p, w, h)$, where F' is the vector obtained by concatenating the weight vectors in F. Subsequently, w and h are dropped from the notations here on as these dimensions are already

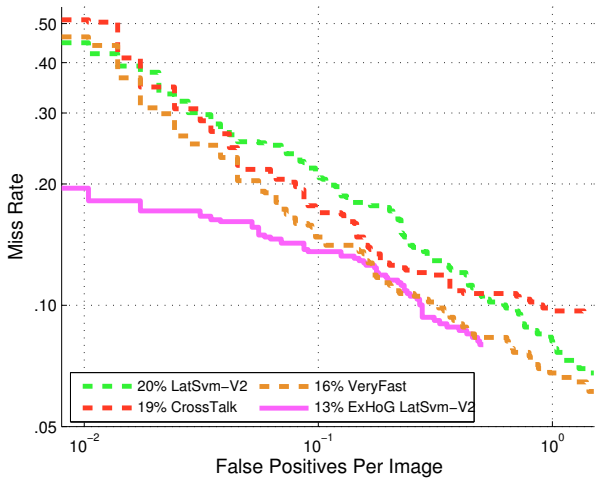


Fig. 3: Performance of ExHoG LatSvm-V2 on INRIA against other existing state-of-the-art methods [Best viewed in colour]. ExHoG LatSvm-V2 outperforms all other methods.

defined by the dimensions of F.

An object model M_c with n_c parts is defined by a root filter, F_0, a set of part models $(P_1, ..., P_{n_c})$ where $P_i = (F_i, v_i, d_i)$ and a bias term, b. F_i is a filter for the i-th part, v_i is a two-dimensional vector specifying a position for part i relative to the root position, and d_i is a four-dimensional vector representing coefficients of a part deformation cost quadratic function.

The placement of a mixture model component M_c in an ExHoG pyramid is defined as $z = (p_0, ..., p_{n_c})$, where $p_i = (x_i, y_i, l_i)$ specifies the level and position of the i-th filter. The score of a placement z can be expressed as a dot product, $\beta_c \cdot \varphi(E, z)$, of a vector of model parameters β_c and a vector $\varphi(E, z)$. β_c and a vector $\varphi(E, z)$ are defined as follows:

$$\beta_c = (F'_0, ..., F'_{n_c}, d_1, ..., d_{n_c}, b), \tag{7}$$

$$\begin{aligned} \varphi(E, z) = (&\phi(E, p_0), ..., \phi(E, p_{n_c}), \\ &- \phi_d(dx_1, dy_1), ..., -\phi_d(dx_{n_c}, dy_{n_c}), 1), \end{aligned} \tag{8}$$

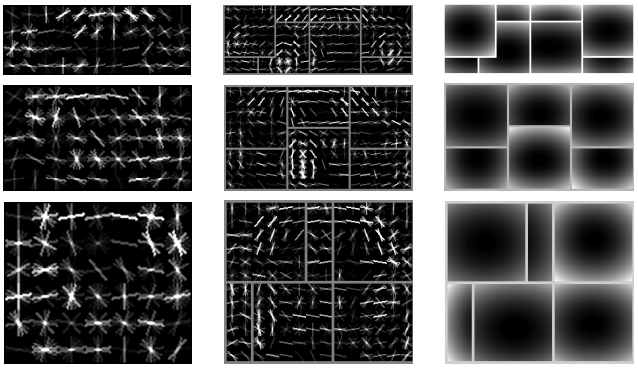
where (dx_i, dy_i) is the displacement of the i-th part relative to v_i and $\phi_d(dx_i, dy_i)$ are the deformation features. This relationship is used to learn the model parameters with the latent SVM framework. To detect objects in an image, the overall score is computed for each root location according to the best possible placement of the parts. High-scoring root locations define detections, while the locations of the parts that yield a high-scoring root location define a full object placement.

3. EXPERIMENTAL STUDY

Experiments are performed on two data sets - INRIA Human [8] and PASCAL VOC 2007 [28] - to demonstrate the effectiveness of parts-based modeling using ExHoG. Furthermore, we also show that ExHoG can be used for general object detection with improved results on PASCAL VOC 2007 which contains 20 different object categories. Results are reported for INRIA using the per-image methodology suggested in [29] as the authors have shown it to be a better evaluation method. For PASCAL VOC 2007, the average precision (AP) of a precision-recall curve for each object class is reported. For both data sets, a cell size of 8×8 pixels is used with a block size

Table 1: Summary of PASCAL VOC 2007 Results.

		aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv
ExHoG LatSvm-V2	without context	35.5	61.6	10.8	16.1	28.0	49.8	58.8	19.6	22.4	24.7	23.9	13.4	59.3	48.7	43.5	13.6	22.7	35.3	48.1	41.9
	with context	**37.9**	**62.6**	**13.3**	**17.6**	**29.9**	49.8	**61.1**	24.1	**23.6**	**27.6**	25.0	**15.1**	**62.1**	**51.8**	**45.3**	**15.6**	**22.2**	**38.6**	**50.2**	43.0
LatSvm-V2	without context	33.2	60.3	10.2	16.1	27.3	54.3	58.2	23.0	20.0	24.1	26.7	12.7	58.1	48.2	43.2	12.0	21.1	36.1	46.0	43.5
	with context	36.6	62.2	12.1	17.6	28.7	54.6	60.4	25.5	21.1	25.6	26.6	14.6	60.9	50.7	44.7	14.3	21.5	38.2	49.3	43.6



(a) car

(b) aeroplane

Fig. 4: Three-component models with 8 parts learned for (a) car and (b) aeroplane classes.

of 2 $\times$ 2 cells. The number of bins for each cell for ExHoG is 18. A 50% overlap of blocks are used in the construction of the feature vectors. The HG block feature is normalized using a clipping value of 0.2. The number of parts is set at 8. A latent SVM [1] classifier is used for both data sets. The training procedure in [1] is followed closely to train our root and part-detectors which includes mining for hard negatives to re-train the detectors. The overall detector is referred to as ExHoG LatSvm-V2.

3.1. INRIA Human

A one-component model for detecting humans is trained for INRIA using the INRIA training set which consist of 614 positives images and 1218 negative images. The sliding window size is 128 $\times$ 64 pixels. Fig. 2 shows the model learned. The INRIA test set consist of 288 images. The images are scanned over multiple scales at a scale step of 1.07. The window stride is 8 pixels in the x and y directions. The miss rate against false positives per image is plotted to compare between different detectors. The *log-average miss rate* (LAMR) [29] is used to summarize the detector performance.

The performance of ExHoG LatSvm-V2 is compared with LatSvm-V2 [1], CrossTalk [30] and VeryFast [31] in Fig. 3. ExHoG LatSvm-V2 achieves a LAMR of 13% which is significantly lower than all the methods being compared with.

3.2. PASCAL VOC 2007

The PASCAL VOC 2007 data set contains 20 object classes and 9,963 images, containing 24,640 annotated objects. The data set is split into two parts: 5011 images for training and 4952 images for testing. A three component-model is trained for each class. Fig. 4 shows the model learned for 2 object classes.

During testing, the images are scanned over multiple scales at a scale step of 1.07. The window stride is 8 pixels in the x and y directions. The bounding boxes of all objects of a given class in an image are predicted. A bounding box is considered correct if it overlaps more than 50 percent with a ground-truth bounding box. Multiple detections are penalized. If several bounding boxes that overlap with a single ground-truth bounding box are predicted, only one prediction is considered correct. The others are considered false positives. We rescore detections using contextual information similar to [1]. This procedure has been noted to improve AP on several object classes. Table 1 summarizes the results (AP score in %) of ExHoG LatSvm-V2 against LatSvm-V2[1]. Out of the 20 object categories, ExHoG LatSvm-V2 performs better or same (only for boat category) than LatSvm-V2 in 16 object categories. This demonstrates that ExHoG can also be applied for detection of other objects besides humans and can achieve a better average detection performance.

4. CONCLUSION

This paper presents a parts-based modeling approach using Extended Histogram of Gradients (ExHoG) for object detection. This approach alleviates the problems of holistic ExHoG modeling which include poor performances in the presence of clutter, occlusions and variable articulations of objects. The parts-based approach of Felzenszwalb *et. al.* [1] is adopted and analysis of ExHoG cell features using PCA is conducted to determine appropriate dimension reduction to reduce complexity analysis. Furthermore, ExHoG is extended to other object detection problems to demonstrate that ExHoG can be used for general visual object detection problem. Results indicate that using parts-based modeling and ExHoG improves detection performance significantly for INRIA and, in comparison to LatSvm-V2, achieves better average precision for 16 out of 20 object categories in PASCAL VOC 2007.

[1]These results are obtained from the website of the authors of [1] at http://people.cs.uchicago.edu/ rbg/latent/. They are the most updated based on the latest version of their code.

5. REFERENCES

[1] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627 –1645, Sept. 2010.

[2] F. Jurie and B. Triggs, "Creating efficient codebooks for visual recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2005, vol. 1, pp. 604–610.

[3] O. Boiman, E. Shechtman, and M. Irani, "In defense of nearest-neighbor based image classification," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, June 2008, pp. 1–8.

[4] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2006, vol. 2, pp. 2169–2178.

[5] T. Tuytelaars and C. Schmid, "Vector quantizing feature space with a regular lattice," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1 –8.

[6] C.P. Papageorgiou and T. Poggio, "A trainable system for object detection," *Int. J. Comput. Vis.*, vol. 38, no. 1, pp. 15–33, June 2000.

[7] P. Viola, M. J. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," *Int. J. Comput. Vis.*, vol. 63, no. 2, pp. 153–161, 2005.

[8] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. I: 886–893.

[9] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, June 2010, pp. 3360 –3367.

[10] A. Satpathy, X.D. Jiang, and H-L. Eng, "Extended histogram of gradients feature for human detection," in *Proc. IEEE Int. Conf. Image. Process.*, Sept. 2010, pp. 3473 –3476.

[11] A. Satpathy, X.D. Jiang, and H-L. Eng, "Extended histogram of gradients with asymmetric principal component and discriminant analyses for human detection," in *Proc. IEEE Canad. Conf. Comput. Robot. Vis.*, May 2011, pp. 64 –71.

[12] X. Wang, X. Bai, W. Liu, and L.J. Latecki, "Feature context for image classification and object detection," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, June 2011, pp. 961 –968.

[13] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 2037 – 2041, Dec. 2006.

[14] C. K. Heng, S. Yokomitsu, Y. Matsumoto, and H. Tamura, "Shrink boost for selecting multi-lbp histogram features in object detection," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, June 2012, pp. 3250 –3257.

[15] A. Satpathy, H-L. Eng, and X.D. Jiang, "Difference of gaussian edge-texture based background modeling for dynamic traffic conditions," in *Advances in Visual Computing*, vol. 5358 of *Lecture Notes in Computer Science*, pp. 406–417. Springer Berlin Heidelberg, 2008.

[16] A. Satpathy, X.D. Jiang, and H-L. Eng, "Human detection using discriminative and robust local binary pattern," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process.*, May 2013, p. pp.

[17] X. Tan and B. Triggs, "Enhanced local texture feature sets for face recognition under difficult lighting conditions," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1635 –1650, June 2010.

[18] H. Zhang, A.C. Berg, M. Maire, and J. Malik, "Svm-knn: Discriminative nearest neighbor classification for visual category recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2006, vol. 2, pp. 2126–2136.

[19] K. Levi and Y. Weiss, "Learning object detection from a small number of examples: the importance of good features," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, June 2004, vol. 2, pp. 53 – 60.

[20] Y. Amit and A. Trouv, "Pop: Patchwork of parts models for object recognition," *Int. J. Comput. Vis.*, vol. 75, pp. 267–282, 2007.

[21] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial structures for object recognition," *Int. J. Comput. Vis.*, vol. 61, pp. 55–79, 2005.

[22] S. Ioffe and D.A. Forsyth, "Probabilistic methods for finding people," *Int. J. Comput. Vis.*, vol. 43, pp. 45–68, 2001.

[23] A. Mohan, C.P. Papageorgiou, and T. Poggio, "Example-based object detection in images by components," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 4, pp. 349–361, Apr. 2001.

[24] J. Sivic, B.C. Russell, A.A. Efros, A. Zisserman, and W.T. Freeman, "Discovering objects and their location in images," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2005, vol. 1, pp. 370 – 377.

[25] D. Crandall, P. Felzenszwalb, and D. Huttenlocher, "Spatial priors for part-based recognition using statistical models," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, June 2005, vol. 1, pp. 10 – 17.

[26] A. Bar-Hillel, D. Levi, E. Krupka, and C. Goldberg, "Part-based feature synthesis for human detection," in *Proc. Eur. Conf. Comput. Vis.*, vol. 6314, pp. 127–142. 2010.

[27] X.D. Jiang, "Linear subspace learning-based dimensionality reduction," *IEEE Signal Process. Mag.*, vol. 28, no. 2, pp. 16–26, Mar. 2011.

[28] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results," http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[29] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743 –761, Apr. 2012.

[30] P. Dollr, R. Appel, and W. Kienzle, "Crosstalk cascades for frame-rate pedestrian detection," in *Proc. Eur. Conf. Comput. Vis.*, Lecture Notes in Computer Science, pp. 645–659. 2012.

[31] R. Benenson, M. Mathias, R. Timofte, and L. Van Gool, "Pedestrian detection at 100 frames per second," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, June 2012, pp. 2903 –2910.