An Empirical Study of Visual Features for Part Based Model

Junge Zhang, Yinan Yu, Shuai Zheng and Kaiqi Huang
National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences
{jgzhang,ynyu,szheng,kqhuang}@nlpr.ia.ac.cn

***Abstract*—Object detection is a fundamental task in computer vision. Deformable part based model has achieved great success in the past several years, demonstrating very promising performance. Many papers emerge on part based model such as structure learning, learning more discriminative features. To help researchers better understand the existing visual features' potential for part based object detection and promote the deep research into part based object representation, we propose an evaluation framework to compare various visual features' performance for part based model. The evaluation is conducted on challenging PASCAL VOC2007 dataset which is widely recognized as a benchmark database. We adopt Average Precision (AP) score to measure each detector's performance. Finally, the full evaluation results are present and discussed.**

I. Introduction

Object detection receives more and more attention in recent years owing to its fundamental role in many applications such as video surveillance, content based image retrieval (*e.g.* Flickr, Google, Baidu) and driver assistance. Accurate object detection is a very challenging task due to appearance deformation, large intra-class variation and cluttered background [1].

Despite these difficulties, there has been a surge of work and significant advancement in object detection. These progresses can be roughly categorized into three groups: 1) building visual feature [2], [3], [4], [5], [6], [7]. 2) Learning object structure [8], [9], [10], [11], [12]. 3) Learning context information [13], [14]. Visual feature and object structure are the two key factors for general object detection. As for visual feature, one of the most representative work is Histogram of Oriented Gradients (HOG) proposed in [2]. Local Binary Patterns (LBP) proposed in [15] is another widely used feature for object detection especially in face detection and human detection [3]. Besides, there are some other types of features such as edgelet feature [16], Region Covariance [17] and shapelet features [18], which have been proposed for some specific object detection. Meanwhile, based on these visual features, there is much work on addressing the problem of modeling object structure at topological level [8], [9], [12]. One of the most promising methods is part based model [9] due to its good performance when handling deformation and occlusion. Therefore, part based model has attracted more and more researchers' interest. In the past several years, the most representative work should be Felzenszwalb *et al.*'s deformable part based model [9]. The method in [9], [19] has been a key component in many applications such as object classification, action recognition and image segmentation.

This paper aims to provide the empirical evaluations of most frequently used low-level visual features for deformable part based model. To our best knowledge, there is no similar work on this topic. The main contributions of this paper are as follows.

- The evaluation can help researchers better understand the potentials of existing low-level visual features and part based model.
- Furthermore, the evaluations should motivate the researchers to build more discriminative visual feature and learning better object structure model.
- The empirical evaluation experiments are conducted to find both the advantages and shortcomings of different features on different categories. These evaluations enable us to design the proper detector with optimal configuration.

The rest of this paper is organized as follows. Section 2 gives a brief introduction to deformable part based model. Section 3 discusses the detailed configuration of different types of visual features. Section 4 presents the evaluation framework including dataset and criterion. Section 5 gives the evaluation results and discussion. Section 6 concludes this paper and discusses the future work.

II. Deformable Part Based Model

This section primarily focuses on the case of star-structured model (see Fig. 1) proposed by Felzenszwalb *et al.* in [9], [19]. This method keeps state-of-the-art performance in the past several years. In star-structured part based model, an object is represented by a root model plus several part models. Especially, quadratic displacement is used to describe the appearance deformation. In a word, the model consists of three elements: root model, parts models and deformation model. A latent SVM is proposed to train the part based model from partially labeled data (only bounding box). The latent information refers to the locations of parts. Once the latent information is determined, the optimization becomes a traditional SVM optimization problem. Moreover, the stochastic gradient descent technique is used to optimize the parameters. Distance transform and dynamic programming techniques are adopted to solve fast matching problem which reduces the complexity from $O(n^p)$ to $O(np)$, where n is the number of

978-1-4577-0121-4/11/$26.00 ©2011 IEEE

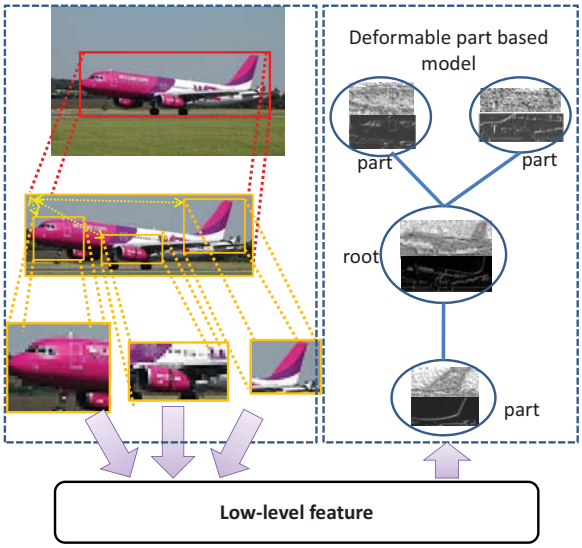

Fig. 1. Overview of deformable part based model.

possible part locations and p is the number of parts. In the final inference step, the score of each sub-window is defined by [9], [20].

$$score\left(w, p_1, \cdots, p_n\right) = m_0\left(w\right) + \sum_{i=1}^{n} m_i\left(p_i\right) - \sum_{i=1}^{n} d_i\left(p_i\right) \quad (1)$$

w specifies a sub-window, p_i and m_i denote the i^{th} part and its score for each hypothesis respectively. d_i represents the deformation cost function penalizing the i^{th} part's displacement.

III. EVALUATION FEATURES

There are a large number of visual features for object detection. Therefore, we adopt a research methodology under which we gradually evaluate those various kinds of features for part based model. Limited by the space of this paper, we choose HOG, LBP, color histogram and their different variants, which are the most frequently used in general object detection to do the evaluation. We'd like to explain why we do not evaluate those salient edge based visual features such as PAS [6]. One one hand, these features are not widely used in general object detection. More important, due to those feature's special computation scheme, they are not flexible enough to be associated with the pixel based feature driven part models [9]. Therefore, considering the generality and flexibility, HOG, LBP and color feature are adopted in this paper.

Histogram of Oriented Gradients (HOG).

HOG is first proposed by Dalal *et al*. in [2] for human detection. Afterwards, HOG has been widely applied in general object detection. In [2], the HOG only includes unsigned gradient information. But the HOG in [9] contains both unsigned gradient and signed information. Let $G_1\left(x, y\right)$ and $G_2\left(x, y\right)$ denotes the signed gradient feature and unsigned gradient feature at pixel $\left(x, y\right)$, respectively. In mathematics

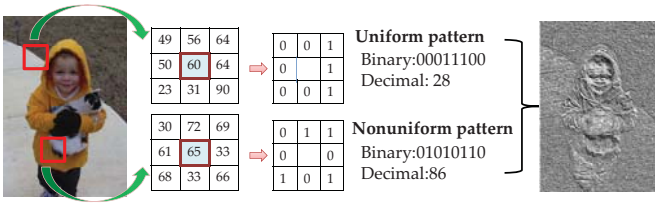

Fig. 2. The computation of LBP. We threshold the neighbor pixels by the center pixel. The larger ones are denoted by "1", others are denoted by "0".

[9],

$$G_1\left(x, y\right) = round\left(\frac{p\theta\left(x, y\right)}{2\pi}\right) \mod p \quad (2)$$

$$G_2\left(x, y\right) = round\left(\frac{p\theta\left(x, y\right)}{\pi}\right) \mod p \quad (3)$$

where p is the quantization level for gradient orientation and $\theta(x, y)$ is the orientation at pixel $\left(x, y\right)$.

Additionally, in [9], they use extra four dimensions computed from the four adjacent cell's gradient energy as texture information. In this paper, we evaluate four variants of HOG.

(1) HOG with signed gradient orientation.

(2) HOG with unsigned gradient orientation.

(3) HOG with both signed and unsigned gradient orientation.

(4) HOG from [9].

Local Binary Patterns (LBP).

LBP was originally developed for texture classification. During the past decade, LBP has been widely applied in general object detection. Similar to HOG, we build a cell structured LBP feature. The original LBP method is illustrated in Fig. 2. It operates with 3×3 neighboring pixels using the center pixel as a threshold. The ones larger than threshold are marked as "1", otherwise as "0". This produces a 8 bit binary number. The histogram of these $2^8 = 256$ is then used as the basic LBP feature. This coding scheme of LBP can be very efficiently implemented and is invariant to monotonic changes in intensity. Another significant extension to the original LBP is the uniform LBP which is designed to reduce affect caused by the non-uniform pattern. One more benefit from uniform LBP is the reduction of length of the feature vector. The uniform pattern is such a concept that has a limited number of $0-1$ transitions [15]. We use the notation $LBP_{p,r}\left(u\right)$ to denote LBP feature that takes p sampling points on sampling circle with radius r, and the limited number of $0-1$ transitions is u. Usually we set $u = 2$. For an example (as is shown in Fig. 2), the patterns 00011100 with 2 transitions are uniform patterns, and 01010110 with 6 transitions is non-uniform pattern. For those good properties, uniform LBP [15] is chosen to construct the feature map at each cell with 59 bins. We can construct the LBP feature at each cell independent from others. On the other hand, the feature map can be built with spatial aggregation as well as [2], [9]. In this paper, we evaluate the two different types of LBP.

(1) LBP without spatial aggregation.

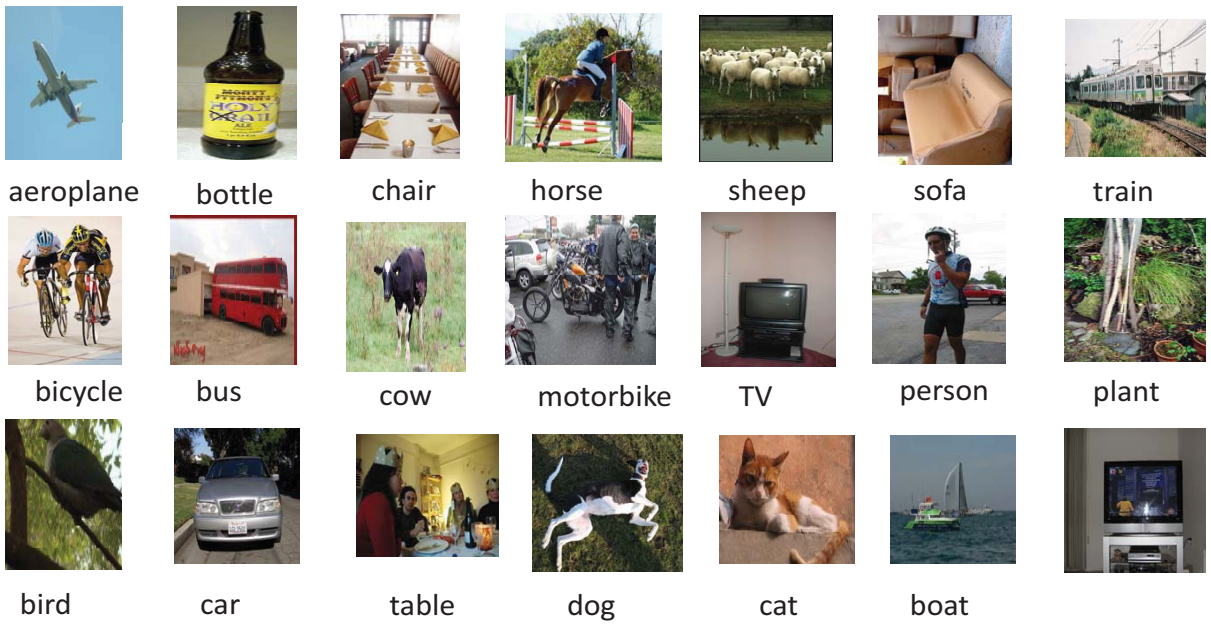

Fig. 3. Some samples from PASCAL VOC2007. There are 20 categories covering vehicles, animals, household objects and people. These images are downloaded from Flickr.

(2) LBP with spatial aggregation.

Color histogram.

The last type of evaluation feature is color histogram. An image is divided into cells and then a color histogram is constructed for each cell. For each color channel, we quantize the color value into 8 bins. Thus, the dimensionality of the color histogram at each cell is $3 \times 8 = 24$. Finally, the color histogram are concatenated and $L2$ normalized. In this paper, two types of color histogram are evaluated: color histogram in HSV color space and color histogram in RGB space.

(1) Color histogram in HSV color space.

(2) Color histogram in RGB color space.

IV. EVALUATION METHODOLOGY

As mentioned previously, we choose the state-of-the-art part based model [9], [19] as our testbed. The goal of the proposed evaluation is to provide an empirical and fair insight into each feature's performance for deformable part based model. To accomplish the goal, we should consider: 1) the evaluation should be conducted on public and well acknowledged challenging dataset. 2) The evaluation should adopt a fair criterion which can predict the real detection performance in practice.

A. Evaluation dataset

There exist many datasets for various kinds of object detection such as MIT LableMe data [21], PASCAL VOC dataset [1] and ImageNet [22]. Among these datasets, PASCAL VOC dataset have been widely applied for the purpose of general object detection. The database increases every year. There are four subclasses [1] in this database including: vehicles, animals, household objects and people. The dataset is divided into two parts: training/validation (`trainval`) and test data (`test`). During the past several years, most of the dominant object detection algorithms have been tested on this dataset. Therefore, to make a convincing evaluation of each visual feature, we choose PASCAL VOC2007 dataset as our evaluation dataset. The other reason why we choose this dataset is that VOC2007 provides full annotations for training and test images. Fig. 3 gives some samples from PASCAL VOC2007.

B. Evaluation criterion

We should follow the principle: the criterion should reflect the detection performance on whole image in practice. This indicates we cannot adopt the commonly used criterion False Positive Per-Window (FPPW) in the area of human detection. FPPW requires the cropped positive windows and assumes that better per-window scores are equal to better detection performance on whole image in practice. But, FPPW only considers cropped positive windows and cannot predict the real detection performance [1], [23]. To measure the real performance, the trade-off between false positives and false negatives should be evaluated. Therefore, we choose the widely used average precision (AP) as the evaluation criterion [1]. AP score is computed from precision/recall curve. Recall represents the correct objects detected from the ground truth. Precision denotes the proportion of correct matches in the all detected objects.

$$recall = \frac{|\{relevant\ objects\} \cap \{retrieved\ objects\}|}{|\{relevant\ objects\}|} \tag{4}$$

$$precision = \frac{|\{relevant\ objects\} \cap \{retrieved\ objects\}|}{|\{retrieved\ objects\}|} \tag{5}$$

abbreviation	description
SHOG	HOG with signed gradient orientation
UHOG	HOG with unsigned gradient orientation
SUHOG	HOG with both signed and unsigned gradient orientation
FHOG	HOG from [9]
LBPS	LBP without spatial aggregation
LBPWS	LBP with spatial aggregation
HSV	Color histogram in HSV color space
RGB	Color histogram in RGB color space

TABLE I
SOME ABBREVIATIONS USED IN THE FOLLOWING EXPERIMENTS.

In PASCAL VOC2007 challenge, the AP score is achieved by the mean precision of eleven equally spaced recall points:

$$AP = \frac{1}{11} \sum_{i \in \{0, 0.1, \cdots, 1\}} p_{in}(i) \quad (6)$$

As explained in [1], we use the maximum precision to represent $p_{in}(i)$, when the corresponding recall exceeds i. This kind of processing has an advantage of stability especially it can reduce the impact of the "wiggles" in precision/recall curve [1].

The next problem is how to determine a detected bounding box is correct or not. The overlap ratio measure is commonly used to measure the correctness [1], [23]. Suppose the detected bounding box is BB_{dt} and a ground truth BB_{gt}. Then, the overlap ratio is defined by

$$\mathrm{O}(BB_{dt}, BB_{gt}) = \frac{area(BB_{dt} \cap BB_{gt})}{aera(BB_{dt} \cup BB_{gt})} \quad (7)$$

In the proposed evaluation, we use the PASCAL VOC measure, which considers the correct detection's overlap ratio must exceed 50%.

C. Evaluation configuration

To make a fair comparison among different features, the object models are all configured with six components, and each component has eight parts. It should be mentioned that we only train three of these six component models, and the other three component models can be inferred from the learnt models according to its left-right horizontal symmetry [9], [19]. Besides, at feature level, the feature map is all configured with 8×8 cell size.

V. EVALUATION RESULTS

To investigate each visual feature's potential performance, we evaluate four variants of HOG , two variants of LBP and two types of color histogram as mentioned in Section3. We use the criterion described in Section 4, plotting the precision/recall curve to obtain the AP score. The symbols are defined in Table 1. The full evaluation results on PASCAL VOC2007 are plotted in Table II.

HOG. As is shown in Table II, FHOG obtains the best AP score in 15 of 20 classes. SUHOG has the best score in 3 of 20, in which train's score is equal to FHOG's. These results indicate that the extra four dimensional texture information helps classification on most categories. This observation can motivates us to build more discriminative low-level feature in a similar way. Compared with UHOG, SHOG exceeds UHOG on 15 categories. Therefore, we can conclude that the signed gradient orientation provides more expressive information than unsigned gradient orientation. It should be noted that the performance of UHOG on `bird` and `dog` exceeds SHOG significantly, which can help researcher build the detector with optimal configuration. Furthermore, SUHOG has better performance on 16 out of 20 categories with a mean improvement by 1.7%, compared with SHOG. Compared with UHOG, SUHOG obtains an increase by 4.4% in mean AP. This means augmented gradient orientation is more discriminative.

LBP. It can be seen from Table II that LBPWS obtains better results in 16 of 20 categories, compared with LBPS. This result proves that spatial aggregation is helpful in cell-structured histogram feature. This can be explained by that spatial aggregation can reduce the negative effect caused by small deformation or aliasing. Therefore, the paradigm can be applied in other block based histogram features. It should be noted that LBPWS performs inferior to SUHOG on most categories except `dog` and `pottedplant`. Based on the results, we can use LBP as the complementary feature for HOG with optimal configuration.

Color histogram. As is shown in Table II, color information performs poor on most categories for detection task except several categories with salient color characteristics such as `horse, car, motorbike and train`, *etc*. This result can be easily understood because color information can be inconsistent for the same category. For example, when people wear different color clothes, the color histogram differs much. However, for those categories with salient consistent color information (*e.g.*, `horse`), color histogram still can be used for recognition.

VI. CONCLUSION AND FUTURE WORK

In this paper, we have evaluated the commonly used visual features for part based model. To help researchers better understand each feature's potential performance, several variants of HOG, LBP and color histogram have been tested on PASCAL VOC2007 dataset. From the results, we can conclude that HOG outperforms LBP feature and color histogram on most classes except `dog`. Specially, signed gradient orientation provides more discriminative power than unsigned gradient orientation. On the other hand, based on the evaluation results, we believe that LBP feature can provide complementary information for HOG feature. Additionally, color information is also a very useful cue for some categories especially in some special application environment. Based on the evaluations, in the future work, we can train a specific detector with optimal configuration. On the other hand, some other visual features will be tested to provide a benchmark performance for further research.

VII. ACKNOWLEDGEMENT

This work is supported by National Natural Science Foundation of China (Grant No.61135002,61175007).

	plane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	meanAP
FHOG [9]	**28.9**	59.5	**10.0**	**15.2**	**25.5**	49.6	**57.9**	**19.3**	**22.4**	**25.2**	**23.3**	11.1	56.8	**48.7**	**41.9**	12.2	**17.8**	**33.6**	**45.1**	**41.6**	32.3
SHOG	24.3	59.5	3.4	11.4	20.4	49.8	57.0	18.9	19.9	21.0	21.5	3.8	**57.1**	44.8	39.2	**12.5**	16.2	29.5	44.1	37.9	29.6
UHOG	20.9	56.7	9.7	10.8	24.0	45.3	53.1	18.2	16.3	20.5	15.9	11.0	54.5	39.6	40.0	6.5	4.7	28.8	25.4	36.9	26.9
SUHOG	26.9	**60.2**	9.8	14.1	22.6	**53.6**	57.5	17.2	20.8	24.1	22.8	10.9	55.9	46.8	40.8	9.9	15.5	32.6	45.1	40.0	31.4
LBPS	23.6	48.4	2.7	11.2	14.0	42.6	50.0	17.0	12.6	17.7	19.4	10.1	51.0	38.3	32.2	9.7	11.7	21.9	35.3	30.1	25.0
LBPWS	23.9	49.7	4.9	8.7	16.6	43.8	51.2	16.6	13.6	20.4	17.1	13.4	52.9	39.9	33.6	10.0	12.3	21.8	35.8	32.4	25.9
LBPL1	27.7	51.1	9.8	10.4	14.8	43.3	50.0	16.6	13.7	21.6	19.2	**14.1**	53.5	43.8	34.0	10.2	17.0	27.5	39.0	33.9	27.6
HSV	13.3	11.0	0.3	0.0	1.3	13.4	22.7	9.6	1.0	4.8	12.3	0.8	33.9	18.0	14.0	5.0	6.7	1.1	18.1	14.0	10.1
RGB	13.3	11.0	0.3	4.7	0.6	12.7	22.9	3.6	0.2	2.1	6.4	1.0	33.1	19.5	17.1	9.6	6.2	1.3	18.2	14.0	9.9

TABLE II
EMPIRICAL EVALUATION RESULTS ON PASCAL VOC2007 DATASETS.

REFERENCES

[1] E. Mark, L. Gool, C. K. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes (VOC) Challenge," *In IJCV*, 2010.

[2] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE International Conference on Computer Vision and Pattern Recognition, 2005.*, 2005, pp. 886–893.

[3] X. Wang, T. X. Han, and S. Yan, "An hog-lbp human detector with partial occlusion handling," in *Computer Vision, 2009 IEEE 12th International Conference on*, 29 2009-oct. 2 2009, pp. 32 –39.

[4] W. R. Schwartz, A. Kembhavi, D. Harwood, and L. S. Davis, "Human detection using partial least squares analysis," in *IEEE International Conference on Computer Vision, 2009.*, 2009.

[5] S.-u. Hussain and B. Triggs, "Feature sets and dimensionality reduction for visual object detection," in *Proceedings of the British Machine Vision Conference*. BMVA Press, 2010, pp. 112.1–112.10, doi:10.5244/C.24.112.

[6] V. Ferrari, L. Fevrier, F. Jurie, and C. Schmid, "Groups of adjacent contour segments for object detection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, no. 1, pp. 36 –51, jan. 2008.

[7] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman, "Multiple kernels for object detection," in *Computer Vision, 2009 IEEE 12th International Conference on*, 29 2009-oct. 2 2009, pp. 606 –613.

[8] P. Schnitzspan, S. Roth, and B. Schiele, "Automatic discovery of meaningful object parts with latent crfs," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010*, june 2010, pp. 121 –128.

[9] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 9, pp. 1627 –1645, sept. 2010.

[10] P. Schnitzspan, M. Fritz, S. Roth, and B. Schiele, "Discriminative structure learning of hierarchical representations for object detection," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, june 2009, pp. 2238 –2245.

[11] Y. Chen, L. Zhu, and A. Yuille, "Active mask hierarchies for object detection," in *Proceedings of the 11th European conference on Computer vision: Part V*, ser. ECCV'10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 43–56.

[12] L. Zhu, Y. Chen, A. L. Yuille, and W. T. Freeman, "Latent hierarchical structural learning for object detection." in *CVPR'10*, 2010, pp. 1062–1069.

[13] S. K. Divvala, D. Hoiem, J. Hays, A. A. Efros, and M. Hebert, "An empirical study of context in object detection," in *Computer Vision and Pattern Recognition*, pp. 1271–1278.

[14] C. Desai, D. Ramanan, and C. Fowlkes, "Discriminative models for multi-class object layout," in *Computer Vision, 2009 IEEE 12th International Conference on*, 29 2009-oct. 2 2009, pp. 229 –236.

[15] T. Ojala, M. Pietik?inen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions." *Pattern Recognition*, pp. 51–59, 1996.

[16] B. Wu and R. Nevatia, "Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors," *International Journal of Computer Vision*, vol. 75, pp. 247–266, 2007.

[17] O. Tuzel, F. Porikli, and P. Meer, "Human detection via classification on riemannian manifolds," in *Computer Vision and Pattern Recognition*, 2007.

[18] P. Sabzmeydani and G. Mori, "Detecting pedestrians by learning shapelet features," in *Computer Vision and Pattern Recognition*, 2007.

[19] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester, "Discriminatively Trained Deformable Part Models, Release 4," 2010.

[20] P. F. Felzenszwalb, R. B. Girshick, and D. Mcallester, "Cascade Object Detection with Deformable Part Models," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[21] B. Russell, A. Torralba, K. Murphy, and W. Freeman, "LabelMe: A Database and Web-Based Tool for Image Annotation," *International Journal of Computer Vision*, vol. 77, no. 1, pp. 157–173, May 2008. [Online]. Available: http://dx.doi.org/10.1007/s11263-007-0090-8

[22] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, june 2009, pp. 248 –255.

[23] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, june 2009, pp. 304 –311.