

Object Detection via Random Subspace Classifiers in Partially Occluded Images

Irshad Ul Haq Najar
Dept. of Computer Science
H.K.B.K. College of Engineering
Bangalore - 45

Abstract— Object detection plays an important role in image processing. Several methods have been developed for detecting objects but they do not provide satisfactory result in cases of partially occluded objects in image frames. This paper describes a general method for object detection when there is a partial occlusion in still images. In this method we have developed classifiers using Random Subspace Method (RSM) to handle the partially occluded image frames. The classifiers handle images on both global level as well as local level to improve the detection performance. Using such classifiers increases detection rate when partial occlusions are present without compromising the detection rate of non-occluded data. The input image window is described as a block-based feature vector where each block is passed through a holistic level. On the basis of holistic responses for each block, local classifiers can be invoked to produce the final output. In contrast to many recent approaches, we propose a method which does not require manual labeling, defining any semantic spatial components, color based features, or any additional data coming from motion or stereo.

Index Terms—Object detection, partial occlusion, random subspace method, classifiers.

INTRODUCTION

Object detection is a challenging problem in vision based computer applications. Object detection involves identifying whether a known object is in an image frame and, if so, determining the location of the object. A good object-detection system should be able to determine the presence or absence of objects in arbitrary scenes and be invariant to object scaling and rotation, the camera viewpoint, and changes in illumination.

Object detection has got many applications in everyday life like robot sensing, surveillance systems and airport security, automatic driving and driver assistance systems in high-end cars, human-robot interaction and immersive, interactive entertainments, smart homes and assistance for senior citizens that live alone, and people-finding for military applications. The wide range of applications and underlying intellectual challenges of object detection have attracted many researchers' and developers' attention from the very early age of computer vision and image processing; and they continue to act as hot research topics in these fields. It is not an easy task to detect an object in an image frame because of wide variability of difficulties coming from the shape, size, colour and some other factors like illumination, visibility and partial occlusion.

Recent works address detection problem with different objectives, which broadly fall into two categories: "specific" and "conceptual" object detection. Specific detection involves the detection of a known object (such as a specific pillow or bottle), while the conceptual detection involves the detection of an object class of interest (such as faces and vehicles). The objective of this paper is to detect any type of object when there is a partial occlusion.

All the methods for object detection rely on the machine learning approach. There are two different issues in machine learning approach: extracting features [1]-[4], and classification through machine learning algorithms [1], [2], [5], [6].

The feature extraction scheme can be roughly classified into two categories. The first category models object shapes globally or densely over image locations, e.g., rectangular features in [1], histograms of oriented gradients (HOGs) in [2], an overcomplete set of Haar wavelet features in [7], or covariance descriptors in [8]. Global, dense feature-based approaches such as [2], [8] are designed to tolerate some degree of occlusions and shape articulations with a large number of samples and have been demonstrated to achieve excellent performance with well-aligned, more or less fully visible training data.

The second category models an object shape using sparse local features or as a collection of visual parts. Local feature based approaches learn body part and/or full-body detectors based on sparse interest points and descriptors from predefined pools of local curve segments [9], [10], a contour segment network [11], k-adjacent segments [12], or edgelets [13]. Part-based approaches model an object shape as a rigid or deformable configuration of visual parts. Part-based representations have been shown to be very effective for handling partial occlusions.

Global methods offer robustness with respect to illumination, background and texture changes, whereas part-based methods are advantageous for different poses. In all cases, the presence of partial occlusions causes a significant degradation of performance, even for part-based methods which are supposed to be robust in that respect.

Current methods for handling occlusion are not generalized, either because additional information is required (coming from manual annotations of the parts or from other sensors), or they are tied to a specific object class [3], [14]. Therefore, our aim is to introduce a general method for

automatic, accurate and robust detection of objects in the presence of partial occlusion.

Here we proposed a method for detecting objects in still images, which can handle occlusion automatically. Manual annotation or defining specific parts/regions of the window are not needed. Our method is based on an ensemble of random subspace classifiers obtained through a selection process. It is worth mentioning that, as the random subspace classifiers use the original feature space, there is no additional feature extraction cost. Similar to [3] and [15], the proposed approach uses a segmentation process to find the unoccluded part of a candidate-window. An ensemble is applied only in uncertain cases. In particular, the proposed method generalizes the inference process presented in [3] by extending it to multiple descriptors.

The proposed approach brings several benefits: 1) the approach is generic, therefore applicable to any class of objects; 2) as the random subspace classifiers are trained in the original space, no further feature extraction is required; 3) the detection is done on monocular intensity images, unlike other methods for which stereo and motion information are mandatory; and 4) during training, we only require a subset of images with and without partial occlusion; other detection methods require delineation of the occluded area.

I. RELATED WORK

So far various methods have been developed for object detection but they lack the detection performance for partially occluded images.

Various features have been applied to detect objects, e.g., Haar features for faces [1], and edgelets for pedestrians [13]. However, HOG is probably the most popular feature in object detection [2], [3], [6], [5]. The distribution of edge strength in various directions seem to efficiently capture objects in images, especially for pedestrians. Recently, variants of Local Binary Pattern (LBP) also show high potentials [3]. A recent trend in human detection is to combine multiple information sources, e.g., color, local texture, edge, motion, etc. [3]. Introducing more information channels usually increases detection accuracy rates, at the cost of increased detection time.

Very few methods from the literature handle occlusions explicitly. Dai et al. [16] propose a part-based method for face and car detection. The method consists of a set of substructure detectors, each of which is composed of detectors related to the different parts of the object. The disadvantage of this method is that the different parts of the object need to be manually labeled in the training dataset, in particular, eight parts for face detection and seven parts for cars.

Wang et al. [3] propose a new scheme to handle occlusions. More concretely, the response at a local level of the histograms of oriented gradients (HOG) [2] descriptor is used to determine whether or not such local region contains a human figure. Then, by segmenting the binary responses over the whole window, the algorithm infers the possible occlusion. If the segmentation process does not lead to a consistent positive or negative response for the entire window, an upper/lower-body classifier is applied. The drawback of this method is that it makes use of a pre-defined spatial layout that characterizes a pedestrian but not any other object class.

In terms of classifiers, linear SVM is widely used, probably for its fast testing speed. With the fast method to evaluate Histogram Intersection Kernel (HIK) [5], HIK SVM was used to achieve higher accuracies with slight increase in testing / detection time. Sophisticated machine learning algorithms also play important roles in various object detection systems. A very influential approach is the part-based, discriminatively trained latent SVM classifier by Felzenszwalb et al. [6]. This detector is a general object detection framework that have been applied to detect tens of object types. The cascade classifier [1] has been successfully used for detecting faces and pedestrians.

Another important research topic in object detection is to improve the speed of detection systems. It is a common practice to use extra hardware like the GPU to distribute the computing task to hundreds of GPU cores in parallel, e.g., in [3]. The cascade classifier framework is an algorithmic approach that first makes face detection run in real-time [1], by using a data structure called integral images. The cascade classifier in [1], however, is only applicable to objects that has a fixed aspect ratio.

II. OBJECT DETECTION METHOD

The block diagram of this method is shown in fig 1. In this method, the window is described by a block-based feature vector. The resulting feature vector is evaluated by the global classifier. If the confidence given by the global classifier falls into an ambiguous range, then an occlusion inference process is applied by using the block responses. Finally, if the inference process determines that there is a partial occlusion, an ensemble classifies the window. Otherwise the final output is given by the global classifier.

In the following, we explain in detail the components shown in Fig. 1.

In this method the original image is presented using block based representation. Fig. 2 gives an idea of such type of representation.

The window descriptor $\mathbf{x} \in \mathbf{R}^n$ is defined as the concatenation of the features extracted from every predefined block \mathbf{B}_i , $i \in \{1, \dots, m\}$. A block is a fixed sub region of the window as shown in Fig. 2. In this method blocks can overlap. The descriptor is denoted as $\mathbf{x} = (\mathbf{B}_1, \dots, \mathbf{B}_m)^T$.

The feature vector \mathbf{x} is passed to a global classifier G

$$G : \mathbf{R}^n \rightarrow (-\infty, +\infty) \quad (1)$$

$$\mathbf{x} \mapsto G(\mathbf{x})$$

where the feature space dimension, n , is $n = m \cdot q$, being q the number of features per block.

The higher the value returned by the function G , the higher the confidence that there is a given object in the window.

In order to detect if there is a partially occluded object in the image, following procedure is used. First, we determine whether the score of the classifier is ambiguous. For example,

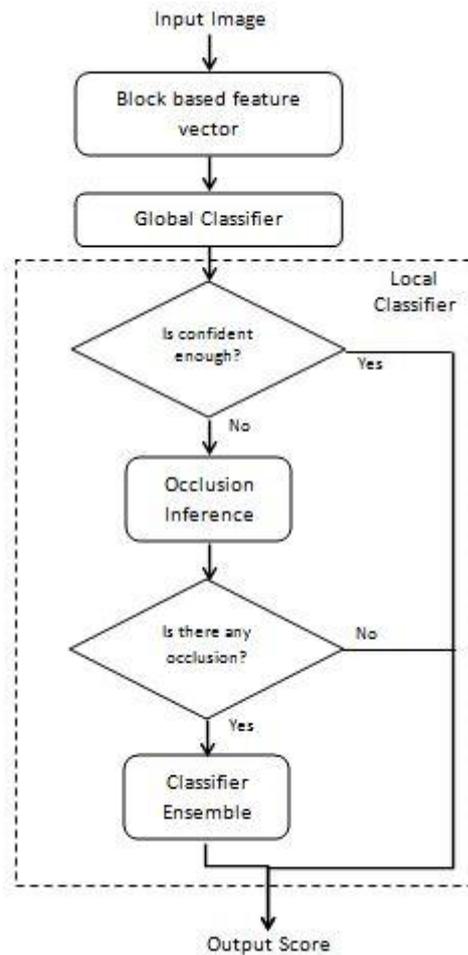


Fig. 1. Block diagram of object detection scheme with occlusion handling

the response from an SVM classifier can be perceived as ambiguous if it is close to 0. When the output is ambiguous, an occlusion inference process is applied. This is based on the responses obtained from the features computed in each block. In particular, for every block B_i , $i \in \{1, \dots, m\}$ we define a local classifier l_i

$$l_i : \mathbf{R}^q \rightarrow (-\infty, +\infty) \quad (2)$$

$$\mathbf{B}_i \mapsto l(\mathbf{B}_i)$$

where the classifier l_i takes as input the i -th block \mathbf{B}_i of the window, and provides as output the likelihood that the block \mathbf{B}_i is object or, otherwise, is an occluding block or background.

The algorithm for the occlusion inference is described in Alg. 1. For each block \mathbf{B}_i we obtain a discrete label s_i by thresholding the local response $l_i(\mathbf{B}_i)$ (1). The discrete label s_i indicates whether the block \mathbf{B}_i is object ($s_i = 1$) or is an occluding block or background ($s_i = -1$). Once we have determined this for all the blocks, we can define a binary map as illustrated in Fig. 3, and then apply a segmentation algorithm

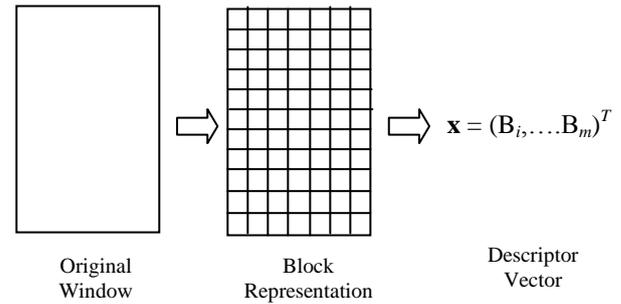


Fig. 2. Block-based representation

Algorithm 1: Pseudo-code for occlusion inference.

Input: $\mathbf{B}_1, \dots, \mathbf{B}_m$
Output: Found partial occlusion
Procedure:
foreach $i \in 1, \dots, m$ **do**
 Calculate $l_i(\mathbf{B}_i)$;
 $s_i := \text{sign}(l_i(\mathbf{B}_i))$;
end
 $(s'_1, \dots, s'_m) := \text{seg}(s_1, \dots, s_m)$;
if $|\sum s'_i| \neq m$ **then**
 return true; // There are occluded blocks
else
 return false; // Object or Background
end

on this binary map. The objective of applying segmentation is to remove spurious responses and to obtain spatially coherent regions. As a result of this segmentation, we obtain spatially coherent block labels s'_i (Fig. 3), and we can determine if there is actually an occlusion or not.

In Algorithm 1, (s_1, \dots, s_m) represents the binary image given by the sign of the local responses $(l_1(\mathbf{B}_1), \dots, l_m(\mathbf{B}_m))$, being $s_i \in \{-1, 1\}$, $\forall i \in \{1, \dots, m\}$. After obtaining the local responses s_i , the algorithm returns (s'_1, \dots, s'_m) as the result of applying a segmentation process over the binary image, where again $s'_i \in \{-1, 1\}$ $\forall i$. Finally, the algorithm returns a Boolean confirming whether there is a partial occlusion depending on the responses. More concretely, if all the responses s'_i are negative, we interpret that such window only contains background. If the responses are all positive, then we consider that there is an object with no occlusions. Finally, if there are both, positive and negative values, we consider that there is a partial occlusion (Fig. 3).

In this method we have adapted random subspace method [19] to develop a flexible model. In particular, we propose to use classifiers trained on random locally distributed blocks; the collection of such classifiers is subsequently browsed to find an optimal combination. Our adapted RSM is introduced next

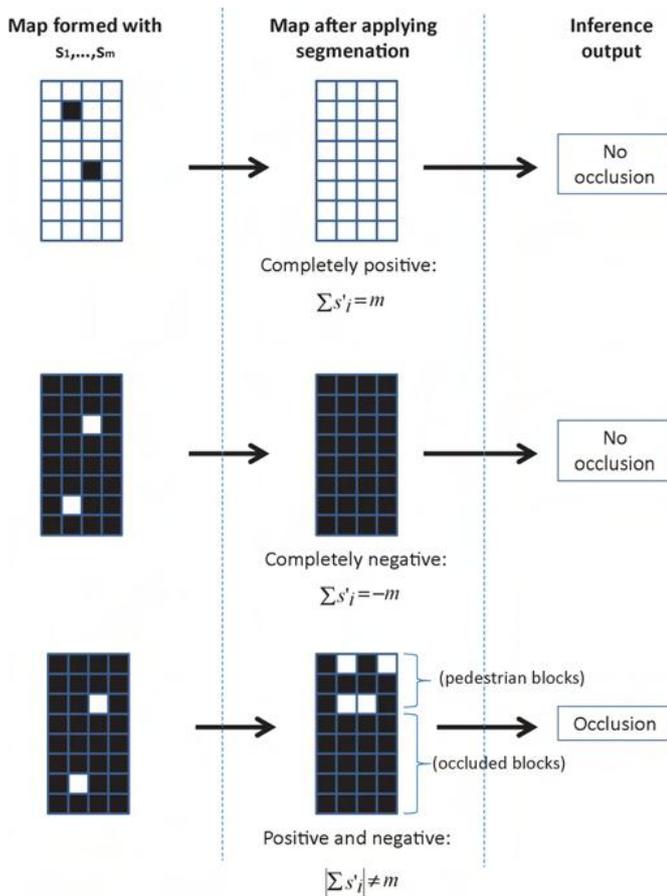


Fig. 3. Image mapping before and after segmentation.

1). *Block-based Random Subspace Classifiers*: Given $I = \{1, \dots, m\}$ the set of block indices, in the k -th iteration we generate a random subset J_k of indices, where $J_k \subset I$. This selection process is carried on until we obtain T different subsets of indices J_1, \dots, J_T . The k -th subset J_k contains m_k indices, where this number can vary across different iterations.

Given the k -th subset $J_k = \{j_1^k, \dots, j_{m_k}^k\}$, we define a subspace formed with the blocks indexed by J_k : $\{B_{j_1^k}, \dots, B_{j_{m_k}^k}\}$. For each subspace, we train an individual classifier g_k . Thus, the decision function of each base classifier of the ensemble can be expressed as a composition of functions

$$\mathbf{R}^{m \cdot q} \xrightarrow{P_k} \mathbf{R}^{m_k \cdot q} \xrightarrow{g_k} (-\infty, +\infty)$$

$$\mathbf{x} = \begin{pmatrix} \mathbf{B}_1 \\ \vdots \\ \mathbf{B}_m \end{pmatrix} \mapsto \begin{pmatrix} \mathbf{B}_{j_1^k} \\ \vdots \\ \mathbf{B}_{j_{m_k}^k} \end{pmatrix} \mapsto (g_k \circ P_k)(\mathbf{x}) \quad (3)$$

where P_k denotes the projection from the original space to the subspace defined by J_k , and g_k the corresponding classifier trained in such subspace. For simplicity of notation, from now on, we will use g_k instead of $(g_k \circ P_k)$.

The resulting algorithm for the random subspace classifiers generation is described in Alg. 2, where D is the training set, x_j denotes the j -th sample and l_j its respective label. Given the J_k indices we apply a segmentation algorithm to the binary image

(r_1, \dots, r_m) , where $r_i = 1$ if the i -th block forms part of J_k , and $r_i = -1$ otherwise (Fig. 4 left image). The segmentation is intended, again, as a means of obtaining spatial coherence in the selected blocks (Fig. 4 right image). As a result of this segmentation process we obtain a new binary image from which we construct a new set J'_k . In particular, let r'_i be the binary value of the i -th block after segmentation, then we define $J'_k = \{i: r'_i = 1\}$, i.e., the set of blocks that are positive in the segmented binary map (Fig. 4 right image).

Then, if the binary image (r'_1, \dots, r'_m) obtained after applying segmentation has all its values set to one (the resulting classifier would be the holistic classifier), to -1 (no subspace can be defined) or $J'_k \in J$ (which means that we have already trained a classifier in the subspace defined by J'_k) we discard this set. Otherwise, we train a classifier in the set D_k defined by the projection P'_k , which is characterized by the indices in J'_k .

Algorithm 2 is used for generating g_1, \dots, g_T trained on random blocks. Based on that, we obtain our final ensemble through the selection strategy described below.

2) *Classifier Selection (N-Best Strategy)*: The accuracy of g_k , $k \in \{1, \dots, T\}$ in our ensemble depends on the discriminative strength of the local region where this classifier is applied. In order to filter out the less accurate classifiers, our system uses the N-best algorithm. A validation set is used to select a subset of classifiers which work best when combined. For this purpose, the algorithm first sorts the classifiers by their individual performance on the validation set and evaluates how many best classifiers form the optimal ensemble. The single best classifier is considered first. Then an ensemble is formed by the first and the second classifiers and evaluated on the validation set. The third classifier is added, and the ensemble evaluated again, and so on. We apply a weighted average for calculating the final decision, in which weights are related to the individual performances. The ensemble with the highest accuracy is selected among the nested ensembles. One of the most important advantages of this strategy is its linear order of complexity regarding the number of evaluations. For an ensemble of T classifiers, we need T individual evaluations plus $T-1$ combined evaluations, giving complexity $O(T)$. Besides, during the evaluations it is not necessary to recompute the features.

3) *Final Ensemble*: Given \mathbf{x} and the classifiers g_k selected after the N-best strategy, the combined decision can be finally expressed as

$$E(\mathbf{x}) = \sum_{k \in S} \omega_k g_k(\mathbf{x}) \quad (4)$$

where S is the set of the classifier indices that form the optimal ensemble, with $|S| \leq T$, and ω_k their corresponding weights.

Combining holistic and part classifier responses is a common technique used in part-based approaches [3], [6]. In our case, if the score given by the ensemble is not confident

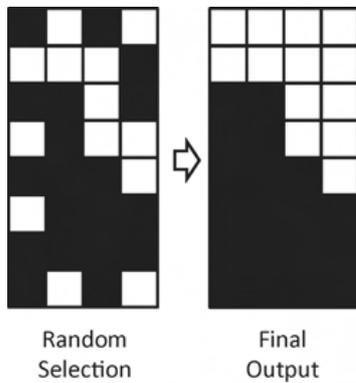
Algorithm 2: Random subspace classifiers pseudo code.**Input:** Training dataset $D = \{(\mathbf{x}_j, l_j) | 1 \leq j \leq n\}$, T **Output:** g_1, \dots, g_T **Procedure:** $I := \{1, \dots, m\};$ $J := \{\emptyset\};$ $k := 1;$ **while** $k \leq T$ **do** Randomly select a subset $J_k \subset I$ with $J_k \neq \emptyset$; Given J_k generate the according $(r_1, \dots, r_m);$ $(r'_1, \dots, r'_m) := \text{seg}(r_1, \dots, r_m);$ Obtain J'_k from $(r'_1, \dots, r'_m);$ **if** $|\sum r'_j \neq m \wedge J'_k \notin J$ **then** Train g_k in $D_k = \{(P'_k(\mathbf{x}_j), l_j) | 1 \neq j \neq n\};$ $J := J \cup \{J'_k\};$ $k := k + 1;$ **end**

Fig. 4. Adapted random block selection.

enough (i.e., the score is smaller than a fixed threshold th), we combine both scores. More precisely, we apply a linear combination between them

$$C(\mathbf{x}) = \alpha H(\mathbf{x}) + (1 - \alpha)E(\mathbf{x}) \quad (5)$$

where α weights the scores of both classifiers.

III. HUMAN DETECTION AS AN EXAMPLE

In the previous section, we presented a general method to handle partial occlusions for object detection. In order to illustrate and validate our approach, in this section we describe in detail a particular instantiation of our method for the class of humans. In order to apply our method to pedestrians, we make use of both linear SVMs and HOG descriptors, which have been proven to provide excellent results for this object class. In addition to HOG descriptor, we also test our system using the combination of the HOG and the local binary pattern (LBP) descriptor, which has recently been proposed in [3] for human detection. In the following we explain very briefly each of these components. Given a training dataset D , the linear SVM finds the optimal hyperplane that divides the space between

positive and negative samples. Thus, given a new input $x \in R_n$, the decision function of the holistic classifier can be defined as

$$H(\mathbf{x}) = \beta + \mathbf{w}^T \cdot \mathbf{x}$$

where w is the weighting vector, and β is the constant bias of the learnt hyperplane.

The HOG descriptor was proposed in [2] for human detection. Since then, the descriptor has grown in popularity due to its success. These features are now widely used in object recognition and detection. They describe the body shape through a dense extraction of local gradients in the window. Usually, each region of the window is divided into overlapping blocks where each block is composed of cells. A histogram of oriented gradients is computed for each cell. The final descriptor is the concatenation of all the blocks' features in the window. The LBP descriptor proposed first by in [17] has been successfully used in face recognition and human detection [3], [18]. These features encode texture information. In order to compute the cell-structured LBP descriptor, the window is divided into overlapping cells. Then, each pixel contained in a cell is labelled with the binary number obtained by thresholding its value to its neighbour pixel values. Later, for each cell a histogram is built using all the binary values obtained in the previous step. Finally, the cell-structured LBP is the result of concatenating all the histograms of binary patterns in such window. The HOG-LBP is the concatenation of both descriptors, HOG and LBP. These two descriptors complement each other, as they combine shape and texture information.

Note that in our case, we interpret every cell LBP as a block, thus a block HOG-LBP represents the concatenated block HOG and the cell LBP computed in the same region. Following the formulation proposed in [3], the constant bias β can be distributed to each block \mathbf{B}_i by using the training data [(5) in [3]]. This technique allows the possibility to rewrite the decision function of the whole linear SVM as a summation of classification results. Then, using this formulation we can define the local classifiers as

$$h_i(\mathbf{B}_i) = \beta_i + w_i^T \cdot \mathbf{B}_i$$

where w_i and β_i are the corresponding weights and distributed bias for each block \mathbf{B}_i , respectively. By defining the local classifiers this way, no additional training per block is required. Moreover, when computing the holistic classifier, the local classifiers are implicitly computed, which means that there is no extra cost.

In this paper, instead of just using HOG features to infer whether there is a partial occlusion [3], we extend the process to rely on both, HOG and LBP features. Thus, the response of each h_i is given by all the features computed in the same block i . As in [3], the segmentation method used in our implementation is based on the mean shift algorithm [19], whose libraries are publicly available. The mean shift weights are set to $w_i = |h_i(\mathbf{B}_i)|$.

IV. CONCLUSION

In this work, we presented a general approach for object detection in still images with the presence of partial occlusion. The method was based on a modified random subspace classifier ensemble. The method can be easily extended to any object, and allows to incorporate other block-based descriptors.

In order to illustrate and validate our approach, we describe in detail a particular instantiation of our method for the class of humans. In order to apply our method to pedestrians, we make use of both linear SVMs and HOG descriptors, which have been proven to provide excellent results for this object class. Two of the most acclaimed descriptors in the literature of the pedestrian detection—HOG and HOG-LBP— were implemented. The linear SVM was used as the base classifier.

REFERENCES

- [1] P. Viola and M. Jones, "Robust real-time face detection," *Int. J. Comput. Vision*, vol. 57, no. 2, pp. 137–154, Jul. 2004.
- [2] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. CVPR*, San Diego, CA, USA, 2005, pp. 886–893.
- [3] X. Wang, T. Han, and S. Yan, "An HOG-LBP human detector with partial occlusion handling," in *Proc. ICCV*, Kyoto, Japan, 2009, pp. 32–39.
- [4] S. Walk, N. Majer, K. Schindler, and B. Schiele, "New features and insights for pedestrian detection," in *Proc. CVPR*, San Francisco, CA, USA, 2010, pp. 1030–1037.
- [5] S. Maji, A. Berg, and J. Malik, "Classification using intersection kernel support vector machines is efficient," in *Proc. CVPR*, Anchorage, Alaska, USA, 2008, pp. 1–8.
- [6] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Proc. CVPR*, Anchorage, AK, USA, 2008, pp. 1–8.
- [7] C. Papageorgiou, T. Evgeniou, and T. Poggio, "A Trainable Pedestrian Detection System," *Proc. Symp. Intelligent Vehicles*, pp. 241–246, 1998.
- [8] O. Tuzel, F. Porikli, and P. Meer, "Human Detection via Classification on Riemannian Manifold," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1–8, 2007.
- [9] J. Shotton, A. Blake, and R. Cipolla, "Contour-Based Learning for Object Detection," *Proc. IEEE Int'l Conf. Computer Vision*, vol. 1, pp. 503–510, 2005.
- [10] A. Opelt, A. Pinz, and A. Zisserman, "A Boundary-Fragment- Model for Object Detection," *Proc. European Conf. Computer Vision*, vol. 2, pp. 575–588, 2006.
- [11] V. Ferrari, T. Tuytelaars, and L.V. Gool, "Object Detection by Contour Segment Networks," *Proc. European Conf. Computer Vision*, vol. 3, pp. 14–28, 2006.
- [12] V. Ferrari, L. Fevrier, F. Jurie, and C. Schmid, "Groups of Adjacent Contour Segments for Object Detection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 30, no. 1, pp. 36–51, Jan. 2008.
- [13] B. Wu and R. Nevatia, "Detection of Multiple, Partially Occluded Humans in a Single Image by Bayesian Combination of Edgelet Part Detectors," *Proc. IEEE Int'l Conf. Computer Vision*, pp. 90–97, 2005.
- [14] B. Wu and R. Nevatia, "Detection and segmentation of multiple, partially occluded objects by grouping, merging, assigning part detector responses," *Int. J. Comput. Vision*, vol. 82, no. 2, pp. 185–204, Apr. 2009.
- [15] B. S. M. Enzweiler, A. Eigenstetter and D. M. Gavrilu, "Multi-cue pedestrian classification with partial occlusion handling," in *Proc. CVPR*, San Francisco, CA, USA, 2010, pp. 990–997.
- [16] S. Dai, M. Yang, Y. Wu, and A. Katsaggelos, "Detector ensemble," in *Proc. CVPR*, Minneapolis, Minnesota, USA, 2007, pp. 1–8.
- [17] T. Ojala, M. Pietikainen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognit.*, vol. 29, no. 1, pp. 51–59, Jan. 1996.
- [18] Y. Yu, J. Zhang, Y. Huang, S. Zheng, W. Ren, K. Huang, and T. Tan, "Object detection by context and boosted HOG-LBP," in *Proc. PASCAL Visual Object Challenge Workshop*, *Proc. Eur. Conf. Comput. Vision*, Crete, Greece, 2011.
- [19] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 603–619, May. 2002.