# Occlusion Handling via Random Subspace Classifiers for Human Detection

Javier Marín, David Vázquez, Antonio M. López, Jaume Amores, Ludmila I. Kuncheva

***Abstract*—This paper describes a general method to address partial occlusions for human detection in still images. The Random Subspace Method (RSM) is chosen for building a classifier ensemble robust against partial occlusions. The component classifiers are chosen on the basis of their individual and combined performance. The main contribution of this work lies in our approach's capability to improve the detection rate when partial occlusions are present without compromising the detection performance on non occluded data. In contrast to many recent approaches, we propose a method which does not require manual labelling of body parts, defining any semantic spatial components, or using additional data coming from motion or stereo. Moreover, the method can be easily extended to other object classes. The experiments are performed on three large datasets: the INRIA person dataset, the Daimler Multicue dataset, and a new challenging dataset, called *PobleSec*, in which a considerable number of targets are partially occluded. The different approaches are evaluated at the classification and detection levels for both partially occluded and non-occluded data. The experimental results show that our detector outperforms state-of-the-art approaches in the presence of partial occlusions, while offering performance and reliability similar to those of the holistic approach on non-occluded data. The datasets used in our experiments have been made publicly available for benchmarking purposes.**

***Index Terms*—Human detection, partial occlusions, random subspace classifiers, ensemble.**

## I. Introduction

Vision-based human detection plays a relevant role in many applications related to robot sensing, surveillance, home automation and driver assistance. Detecting humans is a challenging task due to major difficulties coming from the wide variability of the target, such as the shape, clothing or pose; and the external factors, such as the scenario, illumination, and partial occlusions [1], [2], [3], [4].

Most promising methods of the state-of-the-art rely on discriminative learning paradigms. Along this line, researchers have been mostly working on two different issues: extracting features [5], [6], [7], [8], [9], and classification through machine learning algorithms [5], [6], [10], [11], [12], [13]. State-of-the-art approaches can be divided into two groups: holistic, which rely on detecting the target as a whole, and part-based, which combine the detection of different parts of the body (head, torso, arms, legs, etc.). Holistic methods offer robustness with respect to illumination, background and texture changes, whereas part-based methods have an advantage for different poses [3]. 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 [3].

Expectedly, detection in the presence of partial occlusions has sparked significant interest [14], [15], [7], [16], [17], [18]. For instance, an accident in which a vehicle hits a pedestrian is likely to occur when the pedestrian is not in full view to the driver, *e.g.*, when it appears from behind a parked car. Captured in a sequence of images, several frames prior to the accident will contain a partially occluded human figure. Therefore, accurate detection in the presence of partial occlusion is of paramount importance when building driver assistance systems.

Current methods for handling occlusion lack generalisation, 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 [15], [7], [16], [18]. Therefore, our aim is to introduce a general method for automatic, accurate and robust detection of human figures in the presence of partial occlusion.

Image windows framing partially occluded persons tend to be misclassified due to the fact that, given the descriptor of the whole window, the features corresponding to the occluded areas can be interpreted by the classifier as noise or background. Accordingly we argue that an appropriate solution for these situations is to apply classifiers trained on regions less likely to be occluded. More specifically, we propose to learn the different regions of the window by using random subspace classifiers [19], and subsequently find the optimal ensemble through a bespoke selection strategy.

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 [16]; 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.

Following our previous work [20], here we use a virtual-world based dataset with the occlusion labelling available by design. We also introduce a new real world dataset with occluded pedestrians for testing.

The remainder of this paper is organised as follows. Section II introduces the related work. Section III presents the method from a generic point of view. Section IV, presents a particular implementation for human detection. Section V, relates the design followed in our experiments. In Section VI we validate and discuss our method. Finally, Section VII draws the main conclusions and future work.

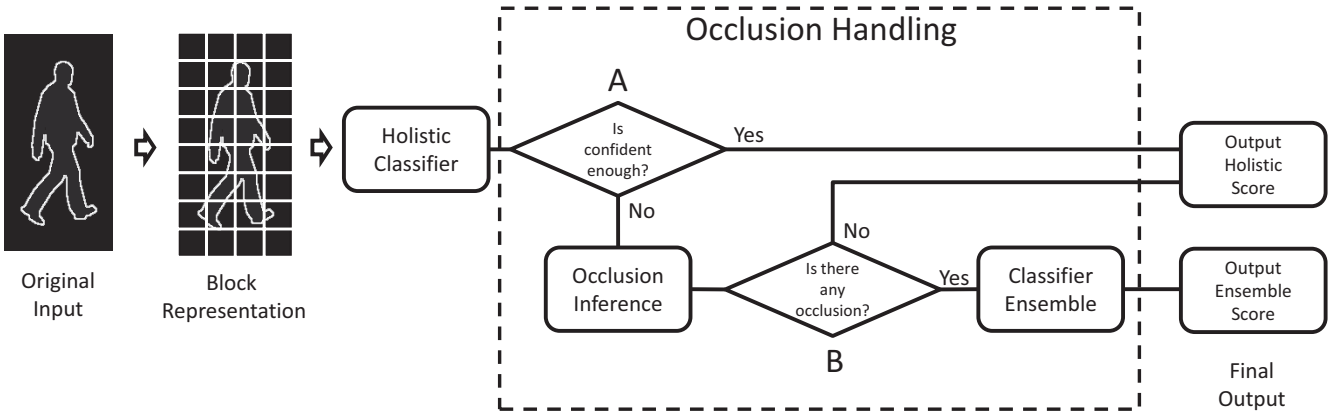



Fig. 1. Occlusion handling scheme. From left to right, the steps for classifying a window.

## II. Related Work

Dollar *et al.* [3] evaluated state-of-the-art detectors under occlusions, and demonstrated that both holistic and part-based methods have similar unsatisfactory performance. This is attributed to the fact that these methods are not specifically designed for handling occlusions.

Very few methods from the literature handle occlusions explicitly. In [14], Dai *et al.* 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 labelled in the training dataset, in particular, eight parts for face detection and seven parts for cars.

A general approach based on the response of different part detectors and a whole-object segmentation process is introduced in [15] by Wu *et al.* The method requires a hierarchical object-parts design with eleven components making up the head, the torso and the legs. The edge pixels of the object that positively contribute to the part detectors are extracted and used together with the part detector responses to obtain a joint likelihood of multiple objects. In this joint likelihood an occlusion reasoning is applied. In case of finding any inter-object occlusions, the occluded parts are ignored. The main drawback of this method is that it requires a manual spatial alignment of the objects, which has to be adapted to each object class. In addition, it requires a special camera set-up in which the camera has to look down on the ground-plane.

Wang *et al.* [7] propose a new scheme to handle occlusions. More concretely, the response at a local level of the Histograms of Oriented Gradients (HOG) [6] 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 characterises a pedestrian but not any other object class.

A mixture of experts for handling partial occlusion is presented in [16] by Enzweiler *et al.* The component layout the authors use is composed by three overlapped regions: head, torso and legs. Then, during the classification process, expert weights are computed to focus on the unoccluded region through a segmentation process applied to the depth and motion images. While the authors demonstrate the robustness of their method against partial occlusions, the drawback of this approach is that it requires both stereo vision and motion information, which limits its applicability if we do not have this additional information. Furthermore, the method is based on a pre-defined spatial layout that is characteristic of the pedestrian, which limits its applicability for other classes of objects.

Gao *et al.* [17] tackle occlusions by identifying and using in the training process cells of pixels which belong to the object in the bounding box. The method outputs not just the detection but also the inferred segmentation. However, the method requires the tedious task of manual labelling all the cells that belong to the object in the training set.

In [18], Girshick *et al.* propose an extension of the deformable part-based detector [11] with occlusion handling. Specifically, the method tries to place the different body parts over the window. Then, if some of the parts are not matched, the method tries to fit in their designated place occluding objects learned from the data. The obvious inconvenience of such an approach is the need of learning the objects that occlude the target. Besides, to extend the method to other classes a different occlusion reasoning has to be defined.

Here we propose a method for detecting human figures 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 [7] and [16], 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 generalises

the inference process presented in [7] by extending it to multiple descriptors.

## III. Occlusion Handling Method

### A. Proposal Outline

We present a general method for handling partial occlusions (see Fig. 1). In such a design, the window is described by a block-based feature vector. The resulting feature vector is evaluated by the holistic classifier. If the confidence given by the holistic classifier falls into an ambiguous range (Fig. 1-A), then an occlusion inference process is applied by using the block responses. Finally, if the inference process determines that there is a partial occlusion (Fig. 1-B), an ensemble classifies the window. Otherwise, the final output is given by the holistic classifier. Notice that, in order to obtain a more accurate decision, we apply the ensemble only when partial occlusion is suspected. In the following, we explain in detail the components shown in Fig. 1.

### B. Block Representation

Our detection system relies on using a block-based representation, one of the most successful descriptor types in use today [3]. A well-known example of such descriptor is the HOG of Dalal *et al.* [7], although there exist many other examples [21], [22]. In section 4 we explain our specific choice for this work. Fig. 2 illustrates the idea of this type of representation, where the window descriptor $\mathbf{x} \in \mathbf{R}^n$ is defined as the concatenation of the features extracted from every predefined block $\mathbf{B}_i$, $i \in \{1, \ldots, m\}$. A block is a fixed subregion of the window as shown in Fig. 2. Our method also allows the blocks to overlap. The descriptor is denoted as $\mathbf{x} = (\mathbf{B}_1, \ldots, \mathbf{B}_m)^T$.

The feature vector $\mathbf{x}$ is passed to a holistic classifier $H$:

$$\begin{array}{rccc} H: & \mathbf{R}^n & \longrightarrow & (-\infty, +\infty) \\ & \mathbf{x} & \longmapsto & H(\mathbf{x}) \end{array} \quad (1)$$

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 H the higher the confidence that there is a pedestrian in the given window. Note that the function $H$ can be any classifier that returns a continuous-valued output, for example, a hyperplane learnt with an SVM.

### C. Occlusion Inference and Posterior Reasoning

In order to detect if there is a partially occluded human figure in the image, we make use of a procedure similar to the one of Wang *et al.* [7]. First, we determine whether the score of the holistic classifier is ambiguous. For example, 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 $B_i$, $i \in \{1, \ldots, m\}$ we define a local classifier $h_i$:

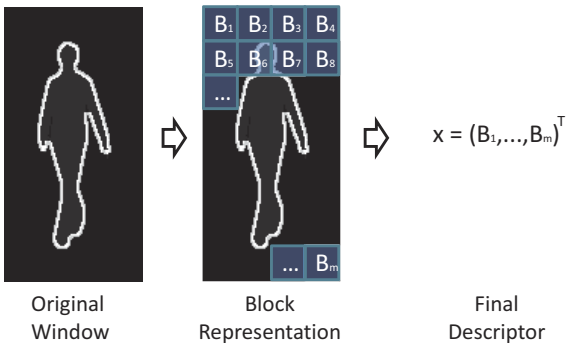



Fig. 2. Block-based representation. From left to right, the original input, then the division into blocks (note that Blocks can overlap), and finally, the feature descriptor.

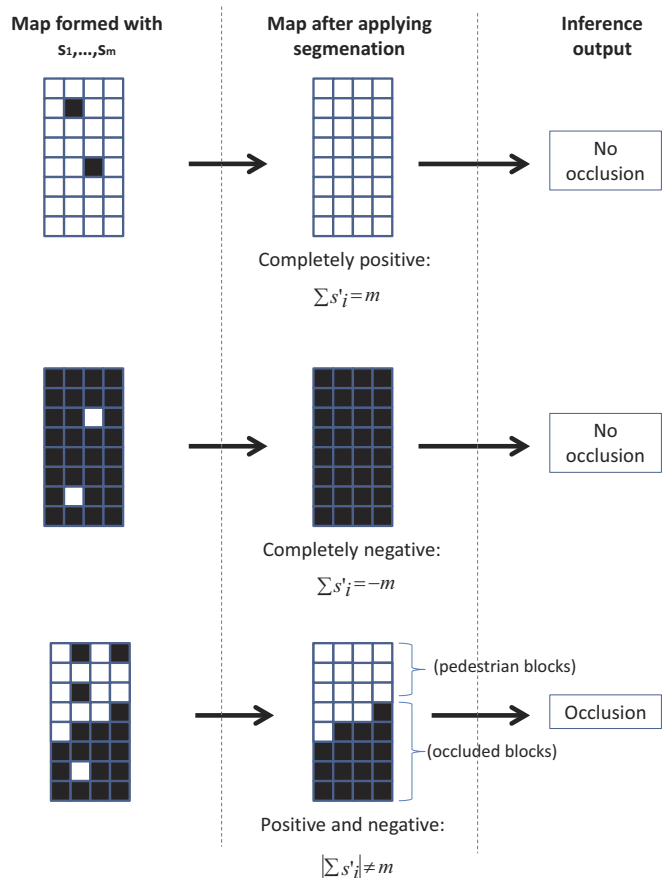



Fig. 3. Occlusion inference and posterior reasoning. From left to right, the initial map formed by the local responses $s_i$; in the middle, the output after segmentation, $s'_i$; at the right, the three inference outputs.

$$\begin{array}{rccc} h_i: & \mathbf{R}^q & \longrightarrow & (-\infty, +\infty) \\ & \mathbf{B}_i & \longmapsto & h(\mathbf{B}_i) \end{array} \quad (2)$$

where the classifier $h_i$ 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 part of the pedestrian or, otherwise, is part of an occluding object or background.

The algorithm for the occlusion inference and the posterior reasoning is described in Alg. 1. For each block $\mathbf{B}_i$ we obtain a

discrete label $s_i$ by thresholding the local response $h_i(\mathbf{B}_i)$ (see Alg. 1). The discrete label $s_i$ indicates whether the block $\mathbf{B}_i$ is part of the pedestrian ($s_i = 1$) or is part of an occluding object 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 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 $s'_i$ (see Fig. 3), and we can determine if there is actually an occlusion or not.

**Algorithm 1:** The occlusion inference and posterior reasoning (Fig. 1-B) pseudo-code.

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

Here $(s_1, \ldots, s_m)$ represents the binary image given by the sign of the local responses $(h_1(\mathbf{B}_1), \ldots, h_m(\mathbf{B}_m))$, being $s_i \in \{-1, 1\}$, $\forall i \in \{1, \ldots, m\}$. After obtaining the local responses $s_i$, the algorithm returns $(s'_1, \ldots, s'_m)$ 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 a pedestrian with no occlusions. Finally, if there are both, positive and negative values, we consider that there is a partial occlusion (see Fig. 3).

### *D. Ensemble of Local Classifiers*

In general, partial occlusions can vary considerably in terms of shape and size; hence a flexible model is needed. We propose an adapted Random Subspace Method (RSM) [19], [23] for this task. 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 below (see Fig. 4).

*1) Block-based Random Subspace Classifiers:* Given $I = \{1, \ldots, m\}$ 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 $J_1, \ldots, J_T$. The $k$-th subset $J_k$ contains $m_k$ indices, where this number can vary across different iterations.

Given the $k$-th subset $J_k = \{j_1^k, ..., j_{m_k}^k\}$, we define a subspace formed with the blocks indexed by $J_k : \{B_{j_1^k}, ..., B_{j_{m_k}^k}\}$. 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:

$$\begin{array}{ccccc} \mathbf{R}^{m \cdot q} & \xrightarrow{P_k} & \mathbf{R}^{m_k \cdot q} & \xrightarrow{g_k} & (-\infty, +\infty) \\ \mathbf{x} = \begin{pmatrix} \mathbf{B}_1 \\ \vdots \\ \mathbf{B}_m \end{pmatrix} & \longmapsto & \begin{pmatrix} \mathbf{B}_{j_1^k} \\ \vdots \\ \mathbf{B}_{j_{m_k}^k} \end{pmatrix} & \longmapsto & (g_k \circ P_k)(\mathbf{x}) \end{array} \tag{3}$$

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 below:

**Algorithm 2:** Our random subspace classifiers pseudo-code.

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

Here $D$ is the training set, $\mathbf{x}_j$ 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 $(r_1, \ldots, r_m)$, where $r_i = 1$ if the $i$-th block forms part of $J_k$, and $r_i = -1$ otherwise (see Fig. 5 left image). The segmentation is intended, again, as a means of obtaining spatial coherence in the selected blocks (see Fig. 5 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 (see Fig. 5 right image). Then, if the binary image $(r'_1, \ldots, r'_m)$ obtained after applying segmentation has all its

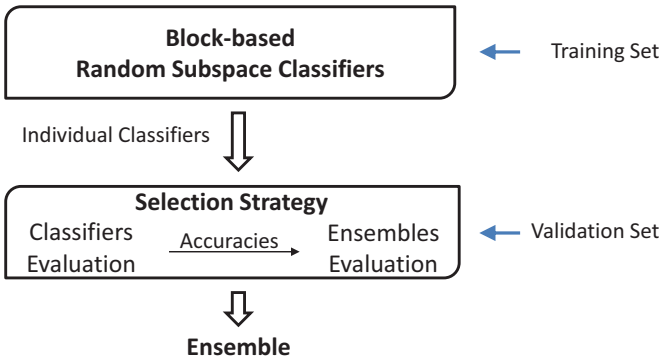



Fig. 4. Training of the adapted random subspace method for handling partial occlusion.

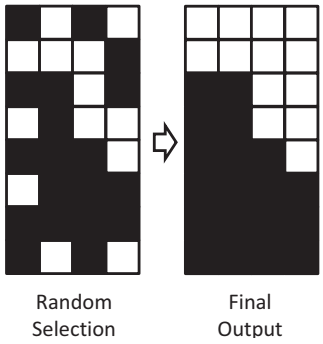



Fig. 5. Adapted random block selection. On the left, the initial randomly selected blocks (in white), and on the right the blocks selected after applying segmentation to obtain spatially coherent regions.

values set to 1 (the resulting classifier would be the holistic classifier), to -1 (no subspace can be defined) or $J'_k \in \mathcal{J}$ (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 characterised by the indices in $J'_k$.

Note that, in the original RSM a fixed number of features are randomly selected from the original space, *i.e.*, all the subspaces have the same dimension. In our case, the dimension $m_k$ may differ from one random subspace to the next as $m_k = |J'_k|$. This way, the classifiers are trained in areas with different sizes.

Algorithm 2 is used for generating $g_1, \ldots, g_T$ 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 $g_k$, $k \in \{1, \ldots, T\}$ 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 [24]. A validation set is used (see Section V-A) 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 (see Eq. 4). 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 $\mathcal{O}(T)$. Besides, during the evaluations it is not necessary to re-compute 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:

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

where $S$ is the set of the classifier indices that form the optimal ensemble, with $|S| \leq T$, and $\omega_k$ their corresponding weights. We derive $\omega_k$ using the validation set described in Section V-A.

Combining holistic and part classifier responses is a common technique used in part-based approaches [7], [11]. In our case, if the score given by the ensemble is not confident 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:

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

where $\alpha$ weights the scores of both classifiers. In Section V-D we describe how to obtain the best parameters for our method.

## IV. Human Detection with Occlusion Handling

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 [25], which has recently been proposed by Wang et al. [7] 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 $\mathbf{x} \in \mathbf{R}^n$, the decision function of the holistic classifier can be defined as:

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

where $\mathbf{w}$ is the weighting vector, and $\beta$ is the constant bias of the learnt hyperplane. Motivated by its success, we also

propose to use the linear SVM as the learning algorithm for the base classifiers described in Sec. III-D.

The HOG descriptor was proposed by Dalal *et al.* [6] for human detection. Since then, the descriptor has grown in popularity due to its success. These features are widely used now 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 Ojala *et al.* [25] has been successfully used in face recognition and human detection [26], [7], [27]. 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. Besides, this combination has been proven to outperform the original HOG descriptor [3]. 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 by Wang *et al.* [7], the constant bias $\beta$ can be distributed to each block $\mathbf{B}_i$ by using the training data (see Eq. 10 in [7]). 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 described in the previous Sect. III-C as:

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

where $\mathbf{w}_i$ and $\beta_i$ 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 work, instead of just using HOG features to infer whether there is a partial occlusion [7], 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 [7], the segmentation method used in our implementation is based on the mean shift algorithm [28], whose libraries are publicly available[1]. The mean shift weights are set to $w_i = |h_i(\mathbf{B}_i)|$.

[1]http://coewww.rutgers.edu/riul/research/code/EDISON/index.html

## V. Experimental Design

In this section, we outline the set-up followed in our experiments. We describe in detail the different datasets used, as well as the procedure conducted during the training and the testing phases. As explained in Section III-D2, as part of our training procedure we make use of a hold out validation set. In order to obtain this validation set we propose the use of virtual pedestrians, a sample of which is shown in Fig. 7. The Daimler multi-cue dataset, published recently [16], is proposed for evaluating the different approaches at the classification level. The INRIA person dataset [6], in which almost none of the pedestrians are occluded, is used to assess the detectors under no occlusions. To evaluate the detector under partially occluded data, we compiled a new dataset, called *PobleSec*, in which a significant number of partially occluded pedestrians are annotated.

### A. Validation dataset

For the validation stage, we need partially occluded data where only the bounding box of the entire object needs to be specified. Recently, the use of synthetic data in Computer Vision has grown in popularity [20], [29], [30], [31] due to their multiple advantages (no manual annotation is required, easy generation of more samples, the possibility of reproducing difficult scenarios, etc.). In this work, we generate a validation set of partially occluded pedestrians needed in the training process (see Fig. 4). In particular, using the same game engine as in our previous work [20], we built a scenario with 50 different human models (see Fig. 6), and created four different variations by introducing illumination, texture and object changes. Afterwards, we recorded 40 video sequences with a freely moving virtual camera, and extracted only positive examples in which humans were partially occluded (see Fig. 7). For validating the classifiers learnt in the INRIA dataset we extracted humans whose bounding boxes were at least 96 pixels tall (around 8000 positive samples in total), and for the classifiers learnt in the Daimler dataset, bounding boxes of height 72 pixels or more (over 12000 examples). Negative images (without humans) were extracted from the same scenario with its different variations. Note that real data with the corresponding label (partially/non-occluded) could also be used in the classifier selection. For the classifiers learnt in the INRIA and the Daimler datasets, we rescaled the extracted humans to the same sizes, *i.e.*, $64 \times 128$ and $48 \times 96$, respectively.

### B. Datasets

*1) INRIA person dataset:* This dataset was proposed by Dalal *et al.* [6], and it is still one of the most widely used datasets in human detection. The data is already divided into training and testing subsets. The annotations are provided for the original positive images (those containing pedestrians). The images come from a personal digital image collection, and pedestrians are shown in different poses against a variety of backgrounds (indoors, urban, rural) in which people are normally standing or walking. Examples and counterexamples

in the training set are normalised to $64 \times 128$ pixels, in which pedestrians are downscaled to a height of 96 pixels (a margin of 16 pixels is added around them). We use the INRIA training set for training the classifiers and the testing set to evaluate the detectors under no occlusions (see Table I for more detail).

*2) Daimler multi-cue dataset:* In 2010, Enzweiler *et al.* [16] published a new dataset, also divided into training and testing parts (see Table I). We used the same partition of the data in our experiment. Two different evaluations at the classification level are done, one assessing the classifiers against partially occluded pedestrians, and the other one only using non-occluded pedestrians. For each labelled pedestrian, Enzweiler *et al.* [16] generated additional samples by geometric jittering. The provided images were captured from a vehicle-mounted calibrated stereo camera rig (grayscale) in an urban environment. The authors also supply the stereo and flow images corresponding to each sample. Only cropped examples and counterexamples are provided, which have a resolution of $48 \times 96$ pixels and a margin of 12 pixels around each side. Non-pedestrian samples contain a bias towards more difficult patterns in terms of shape, which means that hard negative examples are also provided.

*3) PobleSec dataset:* In order to evaluate the different approaches under partial occlusions at per-image level, we have created a new challenging dataset, called *PobleSec*. We captured 327 positive images with a digital camera with a resolution of $640 \times 480$. The images have been taken in urban scenarios in Barcelona and both non-occluded and partially occluded pedestrians are annotated. *PobleSec* dataset has a similar number of labelled pedestrians to the Daimler Partially Occluded dataset. The details of the datasets used in the training and testing stages are shown in Table I.

TABLE I
COMPARISON OF THE DIFFERENT PEDESTRIAN DATASETS. THE NUMBER OF HUMANS SHOWN ARE THE TOTAL NUMBER OF LABELLED ONES.

| | Training | | | | Testing | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | # pedestrians | # non pedestrians | # pos. images | # neg. images | # non-occluded pedestrians | # partially occluded pedestrians | # non pedestrians | # pos. images | # neg. images |
| INRIA | 1208 | - | 614 | 1218 | 566 | - | - | 288 | 453 |
| Daimler | 6514 | 32465 | - | - | 3201 | 620 | 16235 | - | - |
| PobleSec | - | - | - | - | - | 577 | - | 327 | - |

### C. Implementation details

Following the same procedure as Dalal *et al.* [6], we train the holistic classifier by simply feeding the linear SVM with the positive samples and 10 random negative samples per negative image. Once the classifier is trained, we run the detector over the training negative images keeping all the false positive samples (also named hard negatives). Later, we retrain the classifier by using the initial and new hard negatives. For the upper/lower-body classifiers used in Wang's method and for the random subspace classifiers, the initial training is done by using the samples obtained at the first bootstrapping step in the holistic training. Next, we conduct an additional bootstrapping for each one of them (using only the corresponding dimensions). The holistic classifier is also retrained. This means that all the classifiers undergo a second bootstrapping phase.

Fig. 6. Virtual scenario.

The training with both INRIA and Daimler data is performed using only intensity images. For the different classifiers trained in the Daimler dataset, no additional bootstrapping is done, as positive and negative cropped samples are already provided. In our experiments we use the original size of the windows (in contrast to [16], where the windows were scaled to $36 \times 84$ pixels with 6 pixels of margin for their specific component layout). Observe that in this work we only focus on handling occlusion based on features extracted from intensity, so there is no need to follow their specific layout. We implemented Wang's method using both HOG and HOG-LBP descriptors following the same procedure as originally proposed [7].

In our implementation, the HOG descriptor of each window consists of $7 \times 15$ blocks with a spatial shift of 6 pixels for the Daimler data, and 8 pixels for the INRIA data. This leads to overlapping blocks for both data sets. Each block is divided into $2 \times 2$ cells of a fixed number of pixels. We applied $6 \times 6$ cells for the Daimler data and $8 \times 8$ cells for the INRIA data. The histogram of oriented gradients with 12 and 9 orientation bins were computed, respectively. The HOG feature vector is normalised using a L2_HYS norm. For the LBP descriptor, we compute cell structures using the same block HOG size with the same spatial shift. This means that both descriptors are computed in the same region. The L1-sqrt norm is applied for the normalization. In order to remove the aliasing effect when scaling the images (in the training procedure and the detection evaluation), we incorporate a bilinear interpolation.

### D. Training methodology

Different methodologies have been proposed in the literature to conduct the validation stage. Following [32], we use the hold-out protocol (H-method). It has low-computational cost and high reliability for large data sets, and is reproducible when training and testing data are specified. We divided the validation set into halves, one for estimating the individual

TABLE II
BEST PARAMETERS FOR WANG'S METHOD AND OUR METHOD.

| | $\alpha$ | $th$ | Ambiguous range |
|---|---|---|---|
| Wang *et al.* [7] | 0.7 | 1.5 | $[-2, 1]$ |
| Our method | 0.3 | 2 | $[-2, 1]$ |

performance of each base classifier, and the other for evaluating the $N$-best ensemble (see Sect. III-D). The human images were randomly split between the two halves.

In Table II we show the best parameters found by using our virtual dataset for both occlusion handling methods (Wang's approach and our approach). In particular, we found the best values for: the ambiguous range defined in Section III-A (see Fig. 1-A); the weights $w_k$, the classifier score threshold $th$, and the weight $\alpha$ defined in Section III-D3; the minimum and maximum random subspace dimensions used in our adapted RSM (15 and 90 blocks, respectively); and the MeanShift parameters.

### E. Performance Evaluation

We evaluate the classification rate (per window) and the detection rate (per image). A trade-off between missed detections and false positive detections is sought, per window (FPPW), and per image (FPPI), respectively. The curves plotting miss-detection rate versus false positive rate are a special case of ROC curves, in which the x-axis (false positives) is logarithmically scaled.

The classification system assigns a continuous-valued output to each input window related to the likelihood that the window contains a human. The detection system, on the other hand, employs a sliding window for different scales through a HOG/HOG-LBP features pyramid. The sliding window can be defined as a triple $(\Delta_x, \Delta_y, \Delta_s)$, in which the first two parameters denote the spatial stride, and the third parameter is the scale step. In our case, the triple was $(8, 8, 1.2)$. Thus, for each image a group of detections is returned with their respective confidences. Later, a verification refinement is conducted to prune several detections of the same pedestrian through a confidence based non-maximum suppression process. In our case, we follow the PASCAL VOC criterion [33] for object detection classes. Detections are considered as a true positive if they achieve an overlap ratio $\geq 0.5$ with the corresponding pedestrian bounding box, and only one detection per object is interpreted as such, the rest are considered as false positives.

Similarly to [3], instead of using a single point on the curve to compare the performances, we compute the log-average miss rate at nine points on the curve equally distributed over the logarithmic x-axis. Both evaluation methodologies (per window and per image) are frequently used comparing detection methods. In object detection, the per-image evaluation tends to be the standard evaluation methodology [34] because the main concern in real applications is the performance at the detection level.

For the experiments performed in the *PobleSec* dataset, we consider those labels mandatory in which the pedestrian are completely inside the frame, partially occluded and at least 96 pixels tall. Analogous to [3], we normalise all bounding boxes to have a width of $0.41$ times the height during the per-image evaluation. For each classifier $g_k$, $k \in \{1, \ldots, T\}$ described in Sec. III-D, its respective weight $w_k$ is set to be proportional to the log-average classification rate between $10^{-4}$ and $10^{-1}$ FPPW. The weights $w_k$ are normalised to sum to one.

## VI. RESULTS

In this section we describe and discuss the experimental results. Two state-of-the-art methods are compared with our approach, the holistic method and Wang's one with partial occlusion handling. To prove its viability, our approach should be tested for partially occluded as well as non-occluded data.

### A. Per Window

Figure 8 shows the results on the Daimler Non Occluded dataset at per-window level. As can be seen in Fig. 8 (a), the performances using HOG features between our approach and the holistic approach are similar (around 1 percentage point in log-average between performances). Wang's method, instead, shows a higher miss rate at low false positive per window. In Fig. 8 (b) we show the performances of the extended HOG-LBP methods. Again, the performances of our approach and the holistic approach are almost equivalent, which corroborates the HOG results. However, Wang's method, like when using HOG features alone, has a higher miss rate at low false positive per window.

In Fig. 9, we show the curves for the three different methods using HOG and HOG-LBP features on the Daimler Partially Occluded dataset. Fig. 9 (a) shows that, for HOG, Wang's approach is 2 percentage points better than the holistic approach, whereas our approach was 5 percentage points better. Fig. 9 (b) shows that both methods with explicit handling of occlusion outperform the baseline approach in the HOG-LBP feature space.

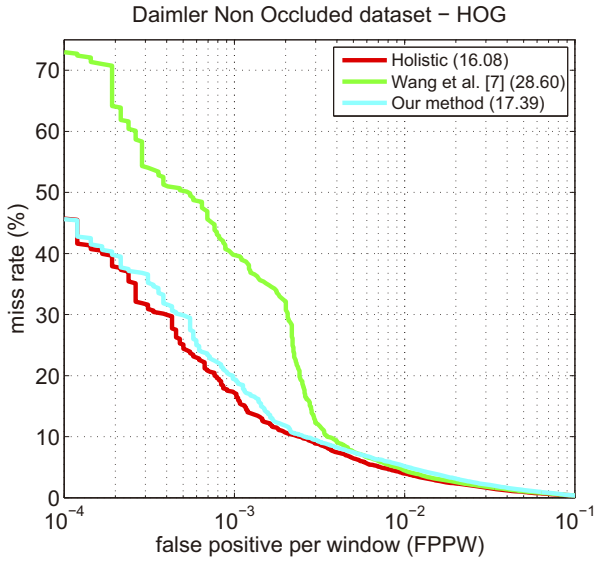
### B. Per Image

In Fig. 10 we show the per-image evaluation using HOG and HOG-LBP on the INRIA testing dataset. Both sub-figures indicate that the occlusion handling does not degrade the performance of the classifier for either Wang's or our method compared to the holistic approach.

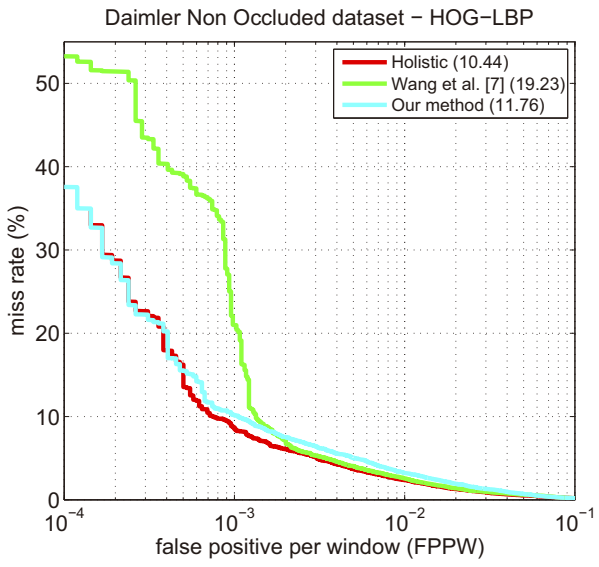
Figure 11 shows the detection curves on the *PobleSec* dataset using both HOG and HOG-LBP features. Only partially occluded humans were used in this evaluation as described earlier. The holistic method fails for both HOG and HOG-LBP features. The best performance is demonstrated by our method for both feature spaces. When using the HOG descriptor, our approach outperforms the holistic approach by 7 percentage points on average, and Wang's method by 4 percentage points. When using the HOG-LBP descriptor our approach outperforms the holistic method by 9 percentage points and Wang's method by 6 percentage points. In contrast to the other methods, our extended HOG-LBP based approach outperforms the HOG based one.

In Figures 13 and 14 we show a qualitative comparison between the different approaches at one FPPI using HOG

Fig. 7. Partially occluded humans under different types of occlusions included in the validation set.

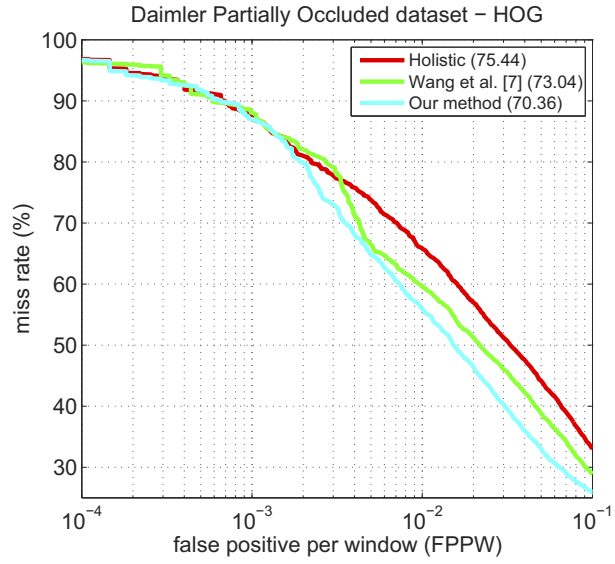



(a)

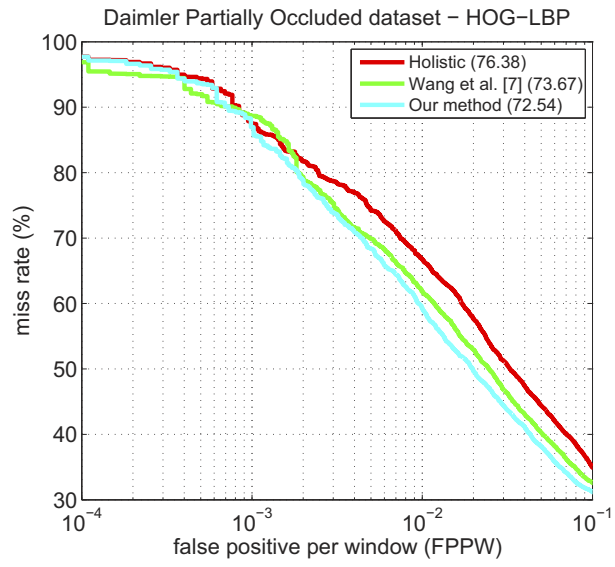



(b)

Fig. 8. Per-window evaluation on Daimler Non Occluded dataset of the three different methods. (a) Evaluation using HOG features. (b) Evaluation using HOG-LBP features. In parenthesis the log-average miss rate between $10^{-4}$ and $10^{-1}$.

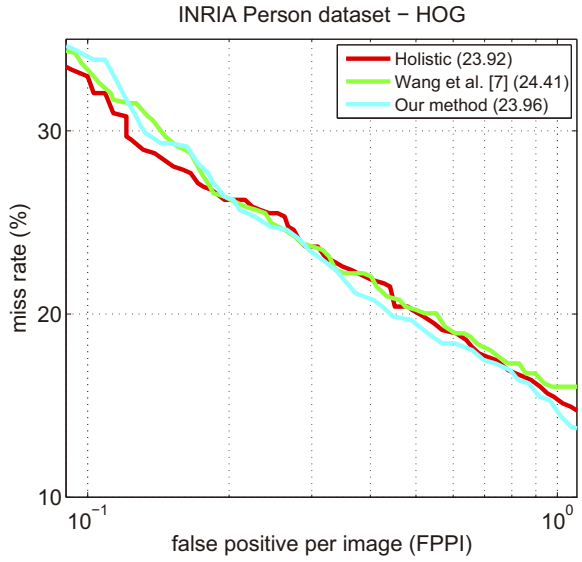



(a)

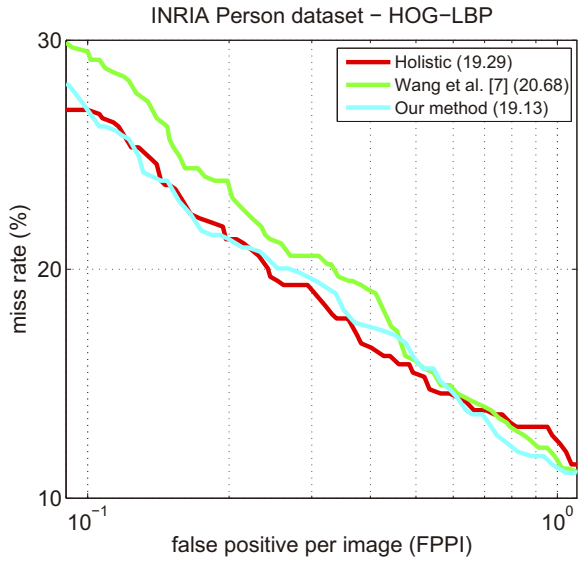



(b)

Fig. 9. Classification comparison on Daimler Partially Occluded dataset. (a) Evaluation of the different methods using HOG features. (b) Performance curves of the methods using HOG-LBP. In parenthesis the log-average miss rate between $10^{-4}$ and $10^{-1}$.

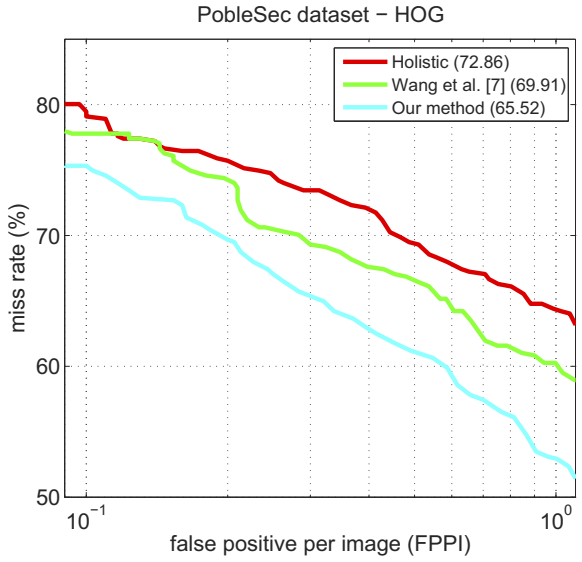



(a)

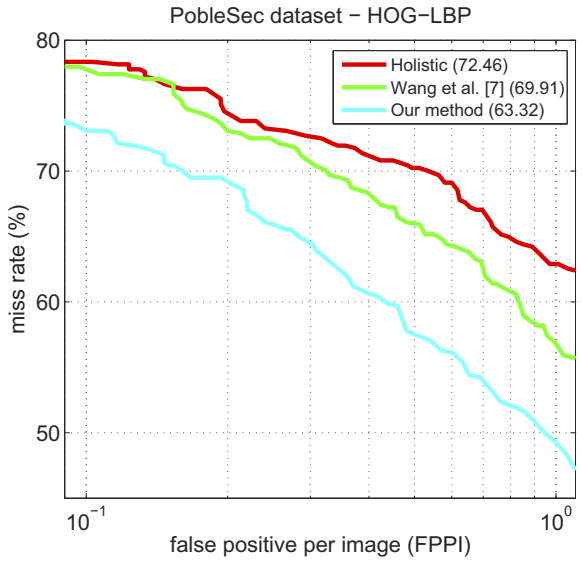



(b)

Fig. 10. Detection curves on the INRIA testing dataset. (a) Evaluation of the different methods on the test set using HOG features. (b) Performance curves of the approaches using HOG-LBP features. In parenthesis the log-average miss rate between $10^{-1}$ and $10^{0}$.




(a)




(b)

Fig. 11. Per-image curves generated on the *PobleSec* dataset. (a) Evaluation of the different methods on the test set using HOG features. (b) the three different curves using HOG-LBP features. In parenthesis the log-average miss rate between $10^{-1}$ and $10^{0}$.

and HOG-LBP descriptors. As can be seen, in both cases, the holistic approach is able to detect certain pedestrians which are partially occluded. However, it does not detect those with a higher level of occlusion. Both occlusion handling methods exhibit better performance by detecting cases missed by the holistic approach. Our approach manages to detect true positives where both other methods fail. This can be seen, for example, in the third and fifth columns of frames in both figures. When both methods have the same true positive detections, Wang's method tends to introduces more false detections, as seen in the second column of frames in Fig. 13.

### *C. Discussion*

After having presented and analyzed the results, we discuss here the points where the proposed framework shows a performance superior to both the holistic [6] and Wang's method [7].

As we have seen, both Wang's method and ours provide a significantly better performance than the holistic method when there are partial occlusions. This is due to the fact that the holistic method makes use of all the features in the window, including those ones that correspond to occluded parts. The latter features add noise to the classifier's decision, and significantly reduce the performance of the holistic method (see Fig. 11). In contrast, both Wang's method and our method focus only on the non occluded regions of the window. This fact makes these methods more robust when we have partial

occlusions, as shown in Fig. 11.

Now let us discuss the difference in performance between our method and Wang's method in the presence of partial occlusions, and explain the technical reasons why our method performs better in this case. Wang's method divides the window into two disjoint regions (upper/lower), therefore, destroying the relationship between features from the two parts. However, this relationship might be important for handling different types of partial occlusions. In contrast, our classifier model consists in an ensemble obtained through a selection process under which a large number of classifiers responsible for differently shaped parts of the window is used (see Fig. 12). Therefore, in our method the relationship between features from different parts of the window is maintained, in contrast with Wang's method. The model obtained with our method is more complete leading to a higher accuracy.

Based on the score of the classifier for each individual block, Wang's method selects the part of the window (upper or lower) that contains a lower number of occluded blocks. The drawback of this method is that, many times, the individual blocks are not very informative, and therefore the score obtained for these blocks is noisy. This leads to a poor part selection if we use Wang's method. In contrast, in our method the selection is based on performance statistics over a validation data set which contains only partially occluded samples. This drives our method to finding and using, collectively, regions in the window that are frequently non-occluded.

Finally, let us discuss the performance of the three methods (our method, Wang's method and the holistic one) in the situation where there are no occlusions. In this case, the three methods perform similarly (see Fig. 10). The conceptual reason why this happens is that both Wang's method and our method only handle the cases inferred as partial occluded targets. The rest of the windows are evaluated by the holistic method. This common design brings comparable performance to the holistic method for non-occluded targets and a significant improvement against partial occluded ones.

In Figure 12 we show four different heat-maps. Each one of them indicates which features (blocks) are actually used in each of our final ensembles (read figure's caption for more details). On one hand, the uneven shading in all the heat-maps shows that features from all parts of the window are present in the ensemble, be it only in a small number of classifiers. This fact demonstrates one of the advantages of our method described above, which consists of preserving and drawing upon relationships between features in the whole window. On the other hand, the large blue area in the bottom half of the window shows that the lower part is rarely useful (also supported by the study performed in [3]). These circumstances together with the results shown in this section highlight the benefit of relying on a supervised statistical learning of the type of occlusions that a given class typically undergoes, *i.e.*, in opposition to making a specific hard assumption about such occlusions (*e.g.*, upper/lower selection).

## VII. Conclusions and Future Work

In this work we present a general approach for human detection in still images with the presence of partial occlusion.

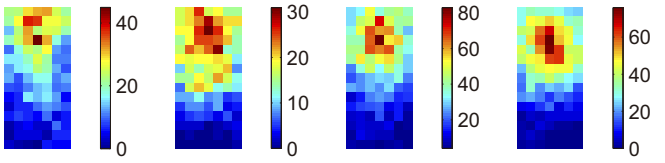



Fig. 12. Heat-maps of which features (blocks) are used in each of our final ensembles. For each block in the window, the figure shows a score (color) equal to the number of classifiers that use the block. From left to right, the heat-maps corresponding to the $48 \times 96$ classifiers using HOG and HOGLBP, and the $64 \times 128$ ones using HOG and HOGLBP, respectively.

The method is based on a modified random subspace classifier ensemble. The method can be easily extended to other objects, and allows to incorporate other block-based descriptors. Two of the most acclaimed descriptors in the literature of pedestrian detection have been implemented, HOG and HOG-LBP. The linear SVM was used as the base classifier. We evaluated our approach on two large datasets, INRIA and Daimler. The INRIA data is considered a standard benchmark for human detection. We designed and release for public use a new challenging dataset called *PobleSec*. The virtual-reality dataset for per-image detection is also released for public use. Both per-window and per-image evaluations have shown that the proposed approach works on a par with the holistic approach when no occlusions are present and outperforms both holistic and Wang's approaches for detection of partially occluded pedestrian images.

As future work, we plan on adding new descriptors, using new kernels (through embedding techniques), and applying our method to other objects.

## Acknowledgement

This work is supported by Spanish MICINN projects Consolider Ingenio 2010: MIPRCV (CSD200700018), TRA2011-29454-C03-01, TIN2011-29494-C03-02, the Ramón y Cajal fellowship RYC-2008-03789, and Javier Marín's FPI Grant BES-2008-007582.

## References

[1] M. Enzweiler and D. Gavrila, "Monocular pedestrian detection: Survey and experiments," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 31, no. 12, pp. 2179–2195, 2009.

[2] D. Gerónimo, A. M. López, A. D. Sappa, and T. Graf, "Survey of pedestrian detection for advanced driver assistance systems," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 32, no. 7, pp. 1239–1258, 2010.

[3] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 743–761, 2012.

[4] Z. Lin and L. Davis, "Shape-based human detection and segmentation via hierarchical part-template matching," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 32, no. 4, pp. 604–618, 2010.

[5] P. Viola and M. Jones, "Robust real-time face detection," in *Int. Journal on Computer Vision*, vol. 57, no. 2, 2004, pp. 137–154.

[6] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, San Diego, CA, USA, 2005, pp. 886–893.

[7] X. Wang, T. Han, and S. Yan, "An HOG–LBP human detector with partial occlusion handling," in *Proc. IEEE Int. Conf. on Computer Vision*, Kyoto, Japan, 2009, pp. 32–39.

[8] S. Walk, N. Majer, K. Schindler, and B. Schiele, "New features and insights for pedestrian detection," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, San Francisco, CA, USA, 2010, pp. 1030–1037.

[9] Y. Pang, H. Yan, Y. Yuan, and K. Wang, "Robust CoHOG feature extraction in human-centered image/video management system," *Proc. IEEE on Systems, Man, and Cybernetics. B, Cybernetics.*, vol. 42, no. 2, pp. 458–468, 2012.
[10] S. Maji, A. Berg, and J. Malik, "Classification using intersection kernel support vector machines is efficient," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Anchorage, Alaska, USA, 2008, pp. 1–8.
[11] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Anchorage, AK, USA, 2008, pp. 1–8.
[12] Y. Xu, X. Cao, and H. Qiao, "An efficient tree classifier ensemble-based approach for pedestrian detection," *Proc. IEEE on Systems, Man, and Cybernetics. B, Cybernetics.*, vol. 41, no. 1, pp. 107–117, 2011.
[13] Y. Xu, D. Xu, S. Lin, T. X. Han, X. Cao, and X. Li, "Detection of sudden pedestrian crossing for driving assistance systems," *Proc. IEEE on Systems, Man, and Cybernetics. B, Cybernetics.*, vol. PP, no. 99, pp. 1–11, 2012.
[14] S. Dai, M. Yang, Y. Wu, and A. Katsaggelos, "Detector ensemble," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Minneapolis, Minnesota, USA, 2007, pp. 1–8.
[15] B. Wu and R. Nevatia, "Detection and segmentation of multiple, partially occluded objects by grouping, merging, assigning part detector responses," in *Int. Journal on Computer Vision*, vol. 82, no. 2, 2009, pp. 185–204.
[16] B. S. M. Enzweiler, A. Eigenstetter and D. M. Gavrila, "Multi-cue pedestrian classification with partial occlusion handling," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, San Francisco, CA, USA, 2010, pp. 990–997.
[17] T. Gao, B. Packer, and D. Koller, "A segmentation-aware object detection model with occlusion handling," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Colorado Springs, CO, USA, 2011, pp. 1361–1368.
[18] R. B. Girshick, P. F. Felzenszwalb, and D. McAllester, "Object detection with grammar models," in *Neural Information Processing Systems*, Granada, Spain, 2011, pp. 442–450.
[19] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 832–844, 1998.
[20] J. Marín, D. Vázquez, D. Gerónimo, and A. M. López, "Learning appearance in virtual scenarios for pedestrian detection," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, San Francisco, CA, USA, 2010, pp. 137–144.
[21] O. Tuzel, F. Porikli, and P. Meer, "Pedestrian detection via classification on Riemannian Manifold," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 30, no. 10, pp. 1713–1727, 2008.
[22] W. R. Schwartz, A. Kembhavi, D. Harwood, and L. S. Davis, "Human detection using partial least squares analysis," in *Proc. IEEE Int. Conf. on Computer Vision*, 2009, pp. 24–31.
[23] L. I. Kuncheva, J. J. Rodriguez, C. O. Plumpton, D. E. J. Linden, and S. J. Johnston, "Random subspace ensembles for fMRI classification," *IEEE Trans. on Medical Imaging*, no. 2, pp. 531–542, 2010.
[24] D. Partridge and W. B. Yates, "Engineering multiversion neural-net systems," *Neural Computation*, vol. 8, pp. 869–893, 1995.
[25] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognition*, vol. 29, no. 1, pp. 51–59, 1996.
[26] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 2037–2041, 2006.
[27] Y. Yu, J. Zhang, Y. Huang, S. Zheng, W. Ren, K. Huang, and T. Tan, "Object detection by context and boosted HOG-LBP," in *PASCAL Visual Object Challenge Workshop, Proc. the European Conf. on Computer Vision*, Crete, Greece, 2010.
[28] D. Comaniciu and P. Meer, "Mean shift: a robust approach toward feature space analysis," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603–619, 2002.
[29] L. Pishchulin, A. Jain, C. Wojek, M. Andriluka, T. Thormaehlen, and B. Schiele, "Learning people detection models from few training samples," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2011, pp. 1473–1480.
[30] D. Vázquez, A. M. López, D. Ponsa, and J. Marín, "Virtual worlds and active learning for human detection," in *ACM International Conference on Multimodal Interaction*, Alicante, Spain, 2011, pp. 393–400.
[31] B. Kuneva, A. Torralba, and W. T. Freeman, "Evaluation of image features using a photorealistic virtual world," *Proc. IEEE Int. Conf. on Computer Vision*, pp. 2282–2289, 2011.
[32] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience, 2004.
[33] M. Everingham and al., "The 2005 PASCAL visual object classes challenge," in *Proceedings of the First PASCAL Challenges Workshop, LNAI, Springer-Verlag*, 2006.
[34] J. Ponce, T. L. Berg, M. Everingham, D. Forsyth, M. Hebert, S. Lazebnik, M. Marszałek, C. Schmid, C. Russell, A. Torralba, C. Williams, J. Zhang, and A. Zisserman, "Dataset issues in object recognition," in *Towards Category-Level Object Recognition*. Springer, 2006, pp. 29–48.

**Javier Marín** received the B.Sc. degree in Mathematics from the Universitat de les Illes Balears (UIB) in 2007. In 2008 he received the Postgraduate Certificate in Education from the UIB. He received his M.Sc. on Computer Vision (CV) and Artificial Intelligence (AI) at the Computer Vision Center (CVC) in 2009 from the Universitat Autònoma de Barcelona (UAB). The focus of his thesis was exploring the viability of using virtual data for training pedestrian detectors. During his Ph.D. he has done two internships, the first in 2011 at the Pattern group in the Bangor University, UK, and the second one, in 2012 at the Mobile Multimedia Processing group in the Aachen University, Germany. His research interest is pedestrian detection, virtual worlds, classifier ensembles, and specially random subspace methods. Currently, he is an active member of the Advanced Driver Assistance Systems (ADAS) group and an Assistant Professor teaching Software Engineer at the Computer Science Department in the UAB.

**David Vázquez** received the B.Sc. degree in Computer Science (Software Engineering) from Universidade da Coruña in 2006 (UDC) with stages at the Universidad Autónoma de Madrid (UAM) and the Universidad Rey Juan Carlos (URJC) where he performed his final project on Face Recognition. In 2008 he received the B.Sc. degree in Computer Science from the Universitat Autònoma de Barcelona (UAB) where he has done his final project on person and car detection in Intelligent Video Surveillance systems for the company Davantis. He received his M.Sc. on CV and AI at the Computer Vision Center (CVC) in 2009 doing his M.Sc. thesis on the study of the effect of the distance in pedestrian detection. Currently, he is working toward the Ph.d. on Informatics at the CVC. He has done an stage at Daimler A.G. His research interest is pedestrian detection, virtual worlds, domain adaptation and active learning. He is an active member of the Advanced Driver Assistance Systems (ADAS) and an Assistant Professor at the Computer Science Department in the UAB.

**Antonio M. López** received the B.Sc. degree in computer science from the Universitat Politècnica de Catalunya (UPC) in 1992, the M.Sc. degree in Image Processing and Artificial Intelligence from the Universitat Autònoma de Barcelona (UAB) in 1994, and the Ph.D. degree in 2000 from the UAB as well. Since 1992, he has been giving lectures in the Computer Science Department of the UAB, where he currently is an Associate Professor. In 1996, he participated in the foundation of the Computer Vision Center (CVC) at the UAB, where he has held different institutional responsibilities, presently being the head of the research group on Advanced Driver Assistance Systems (ADAS) by computer vision. He has been responsible for public and private projects, and is co-author of more than 100 papers, all in the field of computer vision.

**Jaume Amores** received the Ph.D. degree from the Universitat Autònoma de Barcelona (UAB), in 2006. He has held positions in the Computer Vision Center (CVC) and in the Institut National de Recherche en Informatique et en Automatique (INRIA). Currently, he holds a Ramon y Cajal fellowship as part of the Advanced Driver Assistance Systems (ADAS) group. His research interests include machine learning and pattern recognition, object recognition and detection, and medical imaging.

**Ludmila Kuncheva** received the M.Sc. degree from the Technical University of Sofia, Bulgaria, in 1982, and the Ph.D. degree from the Bulgarian Academy of Sciences in 1987. Until 1997 she worked at the Central Laboratory of Biomedical Engineering at the Bulgarian Academy of Sciences. Dr. Kuncheva is currently a Professor at the School of Computer Science, Bangor University, UK. Her interests include pattern recognition and classification, machine learning and classifier ensembles. She has published two books and above 200 scientific papers. Dr. Kuncheva is a Fellow of the International Association for Pattern Recognition (IAPR). She is also the recipient of two Best Paper Awards (IEEE Transactions on SMC, 2002, and IEEE Transactions on Fuzzy Systems, 2006).

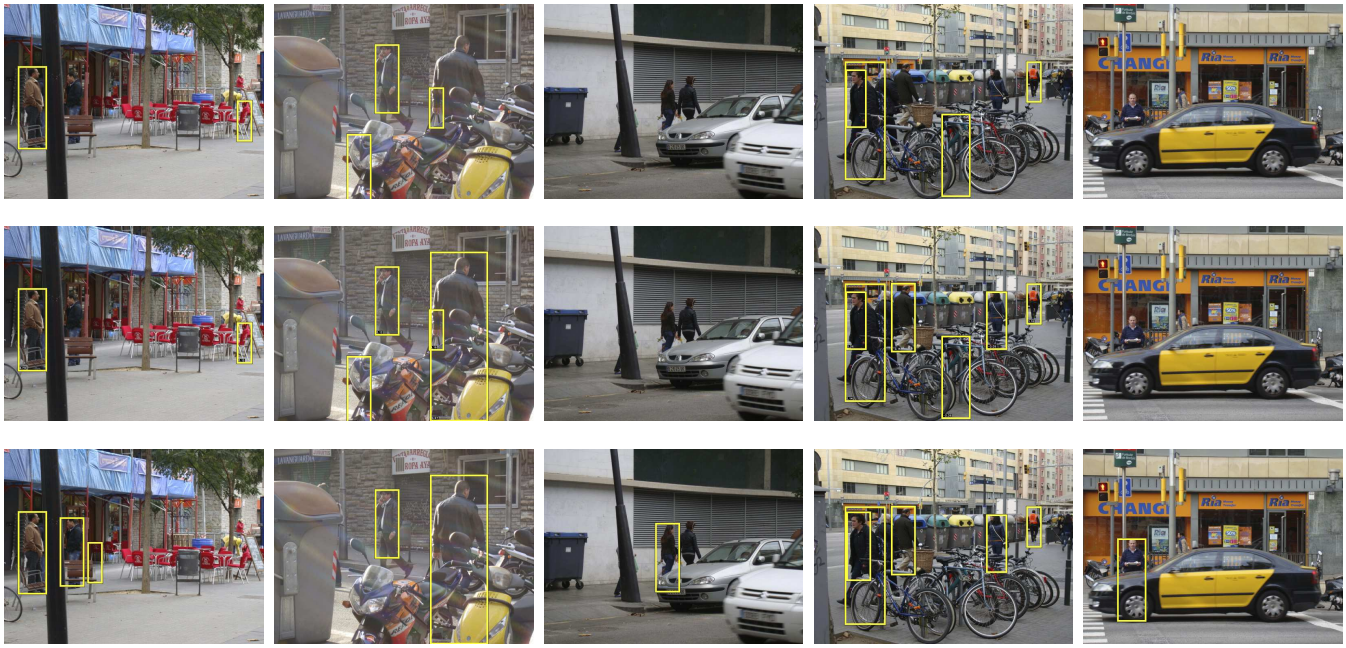


Fig. 13. Per-image results at one FPPI using HOG features. Top row, the detections using the holistic detector without occlusion handling. Middle row, the detections using Wang's detector. Bottom row, the detections using our method.

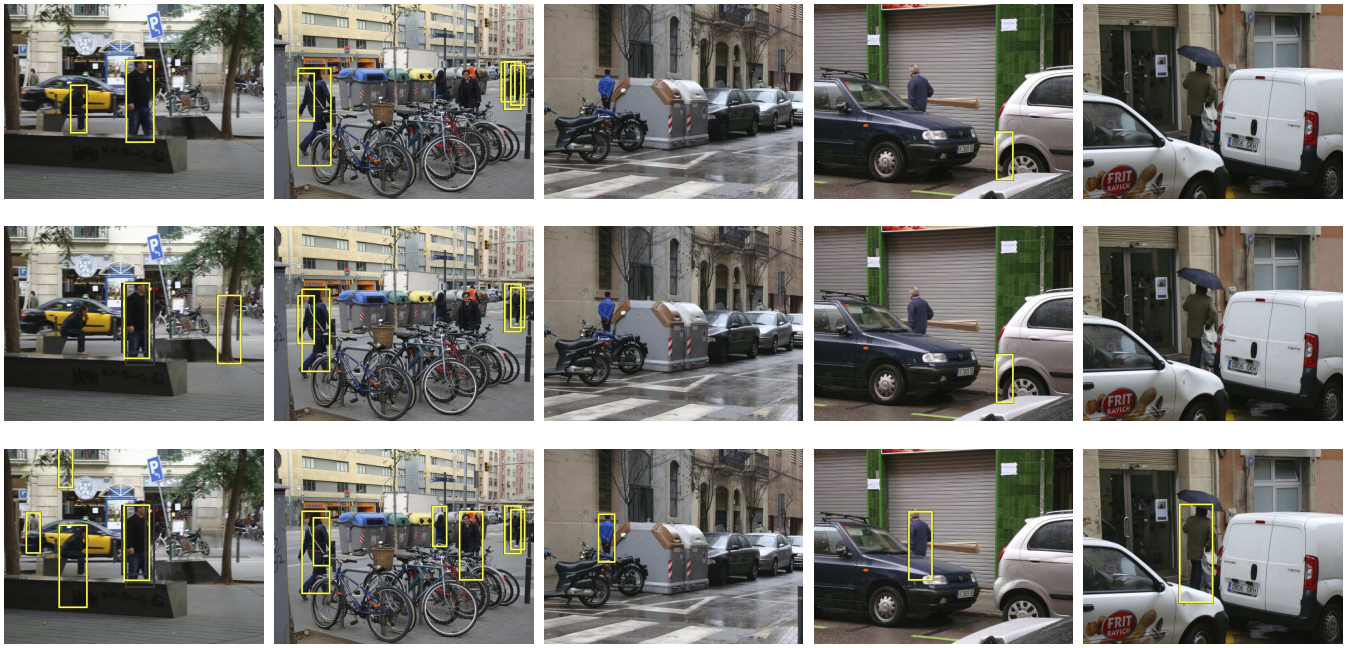


Fig. 14. Per-image results at one FPPI using HOGLBP features. Top row, the detections using the holistic detector without occlusion handling. Middle row, the detections using Wang's method. Bottom row, the detections using our method.