Fast Human Detection in RGB-D Images with Progressive SVM-Classification

Domingo Iván Rodríguez González and Jean-Bernard Hayet

Centro de Investigación en Matemáticas (CIMAT)
Guanajuato, GTO., México
{drodriguez,jbhayet}@cimat.mx

Abstract. In this article, we propose a new, fast approach to detect human beings from RGB-D data, named Progressive Classification. The idea of this method is quite simple: As in several state-of-the-art algorithms, the classification is based on the evaluation of HOG-like descriptors within image test windows, which are divided into a set of blocks. In our method, the evaluation of the set of blocks is done progressively in a particular order, in such a way that the blocks that most contribute to the separability between the human and non-human classes are evaluated first. This permits to make an early decision about the human detection without necessarily reaching the evaluation of all the blocks, and therefore accelerating the detection process. We evaluate our method with different HOG-like descriptors and on a challenging dataset.

1 Introduction

The last decade has produced tremendous advances in the field of pedestrian detection. As an illustration of this evolution and of its impact on other computer vision areas, the availability of powerful detectors has modified radically the main paradigms in use for pedestrian target tracking. Traditional stochastic filtering approaches (Kalman filters, particle filters) coexist now with the so-called "detection-based" approaches that do not formulate anymore the tracking problem as an inference problem within Markov models, but instead as a problem of optimization, namely by association of positive detections given by reliable pedestrian detectors along sequences of frames [1]. Such progress has been made possible mainly by the appearance of a few robust, reliable human detectors such as the one proposed by Dalal and Triggs [2], based on an exhaustive classification of all possible test windows contained in the image as belonging to the pedestrian or non-pedestrian class. Now, one of the main bottlenecks with detection-based methods is specifically the computational times involved in the exhaustive testing of all possible windows, at several scales. These computational times can be dramatically reduced e.g. if the scene geometry is known, but it is not always possible to have this kind of prior knowledge. In this paper, we focus on the reduction of the computational times involved in the use of the acclaimed HOG-based detector [2], by taking advantage of the spatial and semantic structure contained in the HOG descriptor. This way, in most classification cases, we

can make an earlier decision, and produce a significant saving of computational resources. The main contributions of this work are (1) a progressive classification technique based on the HOG-SVM strategy allowing a much faster processing of images and (2) the proposal for a new depth/texture descriptor for RGB-D images. We compare our strategy for pedestrian detection on RGB-D images with several other HOG-like detectors from the literature on a challenging dataset.

This article is organized as follows: In Section 2, we review related works; In Section 3, we recall the principle of HOG descriptors, describe similar descriptors for RGB-D images, including a new one, and pinpoint the HOG-SVM score structure. In Section 4, we describe progressive classification, based on the exploitation of this score structure and on the disparities among the test window blocks in terms of class separability. In Section 5, we give results validating our approach on several pedestrian databases and in Section 6, we draw conclusions.

2   Related Work

One of the first pedestrian detection schemes that successfully combined visual features and classification techniques has been the one of Papageorgiou et al. [3], that used Haar-wavelets with Support Vector Machines (SVM). The next milestone has been the work of Viola and Jones [4], where Haar wavelets were used in combination with a cascade AdaBoost. This resulted in impressive results for face detection. In Sabzmeydani [5], the aforementioned cascade classifier was used with Shapelets as features. Dalal and Triggs [2] use Histograms of Oriented Gradients (HOG), together with Support Vector Machines. Their approach has been successful for pedestrian detection, as HOGs robustly encode the objects shape, and not only differences of contrast. Comparative studies on possible features and classifiers have shown that the combination HOG-SVM was the best one for pedestrian detection [6]. As the HOG-SVM process is quite heavy, proposals have been made to use integral histograms [7] or to replace trilinear interpolation in the histogram construction by spatial convolution [8].

Based on the aforementioned schemes, several combined, part-based classifiers have been proposed. In [9], the HOG-based detector is used to detect the whole body and its upper part. Both are combined with a face detector, a skin detector and a motion detector into a robust classifier. In [10], the pedestrian detection is made in stereo-vision, by using HOG-SVM classifiers for intensity, optical flow and depth images. In the work of [11], several features (HOG, Haar wavelets...) are concatenated into a large description vector used for classification (either with SVM or Adaboost). In [12], a pedestrian detector is obtained by combining several HOG-SVM classifiers trained for each part of the person, which allows robustness for highly deformed bodies.

In this paper, we focus on pedestrian detection methods in RGB-D images. Among the few works in which detectors specific to RGB-D images, [13] propose two descriptors, one (HOD) using only depth information, and a second one (Combo-HOD) combining the output of two different classifiers, one for the depth data, one for the texture. In [14], a descriptor is proposed that is based on the 3D orientations of the surface normals.

3   HOG-Based Descriptors in Multiple Channel Images

Since the pioneer work of Dalal and Triggs [2], HOG-based classification has been very popular for detecting instances of a particular class, such as pedestrians. We first quickly recall how HOG features are extracted and how they are used in classification problems, then we expose a few existing extensions of these features to RGB-D images, and we present a new descriptor, called HOGD.

3.1   HOG-Based Classifiers for Pedestrian Detection

Histograms of Gradient (HOGs) have been originally defined as a dense representation of the local shape and appearance of the image. They capture the local texture information on a cell basis, by generating a discrete distribution (histogram) of the gradient orientations inside this cell, similarly to what SIFT does for sparse interest points, i.e. for wide baseline image matching.

The original HOG features and their corresponding HOG descriptor are described in Section 3.2, together with its variants on RGB-D images. In the following, and in a generic way, we will refer to the extracted descriptor vector (from any of the mentioned methods) as $\mathbf{D}$. The SVM linear classifier response for detecting instances of a given class has the well-known linear form:

$$f_{\alpha,b}(\mathbf{D}) = \alpha \cdot \mathbf{D} + b, \tag{1}$$

where $\alpha \in \mathbb{R}^K$ (K being the dimension of $\mathbf{D}$) and $b \in \mathbb{R}$ are the linear SVM learning parameters and $\cdot$ is the usual dot product in $\mathbb{R}^K$. What is remarkable is that $\mathbf{D}$ has a clear spatial structure, resulting from the per-block concatenation. We intend to use this underlying structure to accelerate the decision process.

3.2   HOG Descriptors for RGB and RGB-D Images

In order to extend the approach of [2] to RGB-D images, some proposals have been done to provide image descriptors $\mathbf{D}$ that, in addition to the texture information (as in HOG), also include depth information. We describe briefly four of those descriptors that we use for evaluation purposes.

Histogram of Oriented Gradients (HOG): We implemented the original HOG version proposed by Dalal and Triggs for RGB images [2]. The image gradient is computed over a 64×128 detection window. The detection window is subdivided into a grid of cells of 4×4 pixels, and a 1D orientation histogram of 9 bins is associated to each cell to capture the local distribution of gradient orientations. Each pixel contributes, according to the gradient magnitude, to the histograms of the 4-connected cells by trilinear interpolation in x, y and orientation. Adjacent cells are grouped into blocks of 2×2 cells with an overlap of one cell in each direction, so that there are 105 blocks in the whole detection window. The histograms of the four cells within a block are locally normalized for contrast variations and then concatenated to form a vector of dimension 36.

The final HOG descriptor is a vector of dimension 3780 used to train a linear SVM classifier. Finally, a detection window is scrolled over an image at multiple scales, so that at each position and scale the HOG descriptor is computed and a classification decision is made by the linear SVM classifier.

Histogram of Oriented Depths (HOD): Designed for RGB-D images, it was first introduced by Spinello and Arras [13] based on HOG. It captures the local shape and appearance of a depth image through the local distribution of depth gradient orientations. The HOD implementation used for this paper follows the same procedure described above to compute HOG descriptors using depth images instead of RGB images. The resulting HOD descriptor is also a 3780 dimension vector. The detection process can be accelerated by taking advantage of depth properties: for each pixel, compatible scales are determined according to its depth information. A HOD descriptor is computed at a given position and scale only if there are enough pixels within the detection window that have a depth compatible with the current scale. This technique produces a significant improvement of the computational time to evaluate the descriptor and to perform the overall detection process. For more details, see [13].

Combo-HOD: It was also proposed by Spinello and Arras [13] to use both texture and depth information in RGB-D images, by combining the classification response of two classifiers based on HOG and HOD descriptors. Both descriptors are independently computed at each position and scale over the texture and depth images, and two linear SVM classifiers are also independently trained. In order to combine the classification responses, a posterior probability is calculated for each type of information as the probability of being a person ($y = 1$) given that the SVM response f is known $p(y = 1|f)$. A sigmoid function is fitted to map SVM responses to probabilities (See Platt et al. [15]). The final combined probability is the weighted average of both probabilities, so that each type of information contributes to the final decision depending on its detection confidence. Spinello and Arras [13] select the weights according to the false positive rate for each type of information at the equal error rate point of a validation set (i.e., at the point of the ROC curve specified by its precision value). Here, we modified a bit this policy by selecting the weights inversely proportional to the intersection of the distributions $p(f|y = 1)$ and $p(f|y = -1)$ estimated by Gaussian kernels on a validation set. This is equivalent to measuring the separability generated by each classifier. It is reported in [13] that the combination of texture and depth information results in a significant improvement of the detection rate.

HOGD: We propose a different descriptor that combines texture and depth information in a single descriptor. We concatenate the HOG and HOD descriptors into a single vector of dimension 7560. A single linear SVM classifier is trained with the HOGD descriptors from the training set. In spite of the increase in dimensionality, training time, and memory, this descriptor has the advantage of determining the contribution balance of texture and depth information by
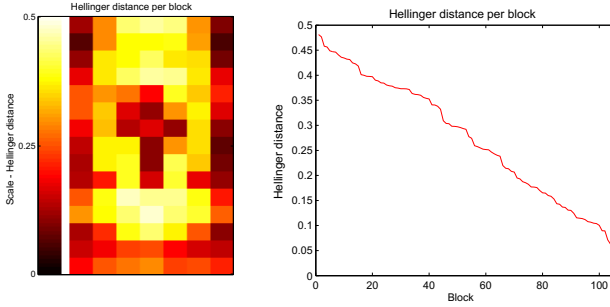
Fig. 1. Distribution of individual Hellinger distances (Eq. 5), for the HOG-SVM classifier, on a block basis. On the left, all blocks are depicted at their respective place in the test window (the brightest is the color, the highest is the distance). It is remarkable that blocks corresponding to the pedestrian silhouette can discriminate more. On the right, the same blocks distances distribution is depicted with decreasing values.

learning, so that this balance does not have to be determined by selecting an error rate point or evaluating a validation set as in [13]. It is also less expensive in time to classify this descriptor, since no posterior probability is required.

3.3 Block-Based Structure in the Descriptor Vector

In all HOG-related approaches, the descriptor vector $\mathbf{D}$ has a block structure, as it results from the concatenation of histograms evaluated inside the set of cells corresponding to the blocks. Hence, the classifier output can be rewritten as a sum of responses β_i by the N different blocks composing the test window,

$$
\begin{aligned}
f_{\alpha,b}(\mathbf{D}) &= \sum_{i=0}^{N} \beta_i + b, \\
\beta_i &= \sum_{k=0}^{M} \alpha_{iM+k} \mathbf{D}_{iM+k},
\end{aligned}
\tag{2}
$$

where M is the number of elements in the descriptors arising from each block. In HOG-like classification methods, the detection process does not consider a specific order for blocks i, as they are all evaluated and summed up in the classifier response. Here, we specify an order to the blocks, so as to evaluate the classifier at different stages of the incorporation of the blocks, and eventually to make an early decision on the binary output, to accelerate the detection process.

As an example, one could expect that in pedestrian detection, blocks corresponding to a pedestrian silhouette are more discriminative, whereas those holding background information or clothes are less discriminative. Hence, by evaluating the first ones as soon as possible, it may be possible to make an early decision about the presence or absence of a pedestrian inside a test window.

4 Progressive Classification

In the following, we define progressive classification as the classification process resulting from adding gradually (in a specific order!) to the global classifier score

the individual contribution β_i of each block $\mathbf{D}_i$. We also refer to the partial classifier output after evaluating i blocks as $f_{0:i}$, defined recursively as

$$\begin{cases} f_{0:0} = & b, \\ f_{0:i} = f_{0:i-1} + \beta_i \text{ for } i > 0. \end{cases} \tag{3}$$

4.1   Block Ordering

To evaluate the importance of any block in the classifier output, we use a criterion reflecting the class separability caused by each individual block i. For that purpose, let us define as f_i (with only one index) the output of the SVM classifier resulting from using only one block (block i):

$$f_i = \beta_i + b. \tag{4}$$

Once the learning has been done, we use our verification dataset to evaluate the discriminability of each atomic, block-based classifier i. With all the ground truth data compared to the application of the classifiers, we use Gaussian kernels to estimate the discrete distributions $P_i^+ = P(f_i|y = 1)$ and $P_i^- = P(f_i|y = -1)$, for each block. The more these distributions overlap, the less discriminative the block will be for the detection process. To evaluate this overlapping, we use the Hellinger distance (based on the Bhattacharyya coefficient) between P_i^+ and P_i^-:

$$d(P_i^+, P_i^-) = \sqrt{1 - \sum_{f^k} P_i^+(f^k) P_i^-(f^k)}, \tag{5}$$

where f^k covers all the discretized values of the score value f_i. This distance takes values between 0 and 1, where 1 occurs when the distributions are the same, and 0 when they have no common support. Hence, the closer to zero is $d(P_i^+, P_i^-)$, the more separated are the pedestrians/non-pedestrians classes. In Fig. 1, we depict these computed values for each block in the detection window.

The selected order of evaluation of the blocks is then the one corresponding to decreasing values of $d(P_i^+, P_i^-)$, in such a way that the blocks with the positive and negative classes most separated will be evaluated first.

4.2   Early Decision

Let us define $p_i = P(y = 1|f_{0:i})$ as the probability that a tested sample is positive, given that the partial output of the classifier (up to block i) is $f_{0:i}$

$$p_i = P(y = 1|f_{0:i}) = \frac{1}{1 + \exp(A_i f_{0:i} + B_i)}, \tag{6}$$

where A_i and B_i are the parameters of a sigmoid, adjusting the distributions $P_{0:i}^+ = P(f_{0:i}|y = 1)$ and $P_{0:i}^- = P(f_{0:i}|y = -1)$, obtained by evaluating the partial, block-ordered classifier from block 1 to i, as explained in [15].
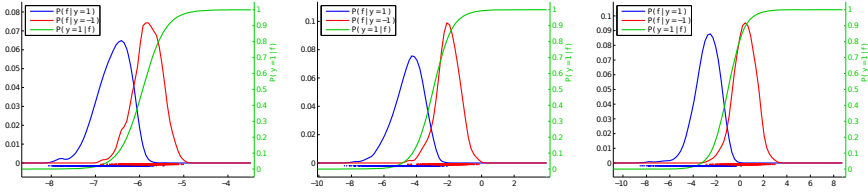
Fig. 2. Evolution of the $P_{0:i}^{+}$ (blue) and $P_{0:i}^{-}$ (red) distributions while incorporating (from left to right) 10, 50 and 105 ordered blocks. In green, we depict the corresponding p_i probability to be an instance of the class, given $f_{0:i}$.
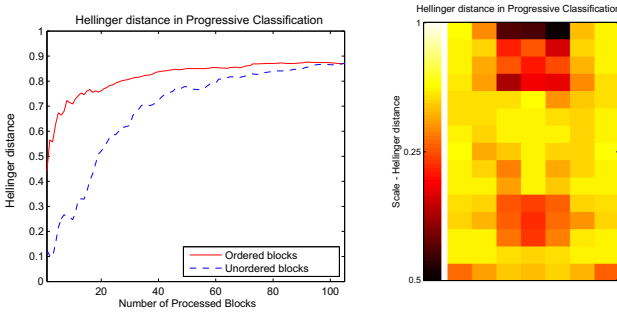
Fig. 3. Evolution of the Hellinger distance with the partial scores f_i (from Eq. 2). Left: The dashed curve is the evolution of the Hellinger distance when the number of blocks is increased, without ordering the blocks; the solid curve is the same with ordered blocks. Right: We display the separability of the partial classifier ending with each block. It can be noticed that the most discriminative blocks are evaluated first.

We depict (in blue and red, respectively) different partial distributions $P_{0:i}^{+}$, $P_{0:i}^{-}$ distributions in Fig. 2, for $i = 10, 50, 105$. Observe that the separability goes increasing with i. Moreover, we depict (in green) the p_i sigmoid curves for all these cases. Then, in Fig. 3, we show the evolution of the Hellinger distances $d(P_{0:i}^{+}, P_{0:i}^{-})$ with and without ordering of the blocks.

To make an early decision, we associate, to each ordered level i, a rejection threshold U_i, in such a way that if p_i (deduced from the evaluated $f_{0:i}$) is inferior to this threshold, the decision is made of an early rejection of the test window, without having to evaluate the rest of the blocks, expecting that the evaluation of the following blocks will not bring a large enough contribution. The decision at the step $i < N$ is made according to the following rule:

$$\begin{cases} y = -1 \text{ if } p_i < U_i \\ \text{No decision is made otherwise and } i = i + 1. \end{cases} \tag{7}$$

where p_i is the probability of a positive classification, derived from the partial output $f_{0:i}$ in Eq. 6. For the progressive classification to be effective, we heuristically define U_i as an increasing function of i. The idea is that at the final stage of

the evaluation ($i = N$), then we apply the threshold for the originally designed SVM, namely U_f; however, in the intermediary stages, we apply lower thresholds, since partial scores may be abnormally low because of some local phenomenon (illumination, occlusion...). Hence, the threshold at stage i is defined as

$$U_i = \frac{i}{N} U_f. \tag{8}$$

The defined threshold is a rejection threshold, and one could design a similar approach for an early acceptance threshold. However, in the case of object detection in a given image, the number of negative tests is by order of magnitudes superior to the number of positive tests. This suggests that an early rejection threshold will have much higher effect on the global computational time performance than an early acceptance threshold. Hence, if there is enough evidence in the first stages that the test window corresponds to a pedestrian, then the evaluating process is pursued, up to a point where the test window is definitely rejected or accepted (at the last stage $i = N$).

5   Experimental Results

We tested our method on our own dataset of RGB-D images captured with a Kinect sensor. The dataset was generated in a challenging environment with hard illumination conditions. It contains more than 2000 images, with an important variability of pedestrian appearances and poses, and the objects are located within a wide depth range (1 to 10 meters). We divided it into a learning set (1834 positive examples and 516 negative images), and a testing set (523 test images). We trained our linear SVM classifiers using the bootstrapping technique to improve the classification results as most works in this domain do. A few examples from this dataset are given in Fig 4. All the results in the following have been determined on these 523 test images. Close multiple detections from the classifier are merged by a simple criterion of overlapping. Detections are furthermore validated as positive when the detected window relative overlapping with the ground truth window reaches a threshold value.

Handling the Sensor Depth Resolution Variations with Range. It is important to note that the Kinect sensor has a limited depth range and that its resolution decreases as the distance to the sensor increases. The technical specification recommends to use it in a range between 0.8 to 4 meters, because depth information is significantly lost farther than this limit. Therefore, we can not perform detection with the same confidence for near and far depth ranges.

We observed that global results (on the whole range of depths) are degraded mainly because the relevant features are not the same at different depth ranges. Hence, we propose to train different classifiers for three different depth intervals.

We divided the set of training images into three subsets according to the distance of the sample to the sensor: 0 to 4 meters, 4 to 8 meters and farther than 8 meters. During detection, the correct linear SVM classifier is selected for

**Fig. 4.** A few examples of the dataset we used for evaluation purposes in this paper
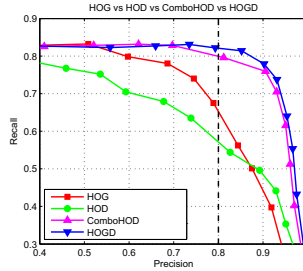
Fig. 5. Global performance comparison between the four HOG-based descriptors

the current image scale. This training procedure generates different contribution balances between texture and depth information for each depth range. For near samples, depth information has a greater contribution than texture, and for far samples, texture information is more discriminative. For ComboHOD, different weights are determined for each interval, and for HOGD, the independent learning processes generates three different detectors.

Since progressive classification orders the blocks according to the separability of each block alone, a different order is determined for each depth range. In the cases of HOGD and ComboHOD, a greater proportion of depth blocks are evaluated within the first blocks for the nearest depth range, whereas more texture blocks are computed first for the farthest range. Thus we can probabilistically determine which information is more discriminative at different depths.

Performance Comparison among HOG-Based Descriptors. Figure 5 shows a performance comparison between the four HOG-based descriptors on the same test set of RGB-D images. We observe that Combo-HOD and HOGD outperform HOG and HOD. This demonstrates that the combination of texture and depth information behaves better than a single type of information. Furthermore, we note that our new descriptor HOGD performs slightly better than ComboHOD on most of the operation domain.
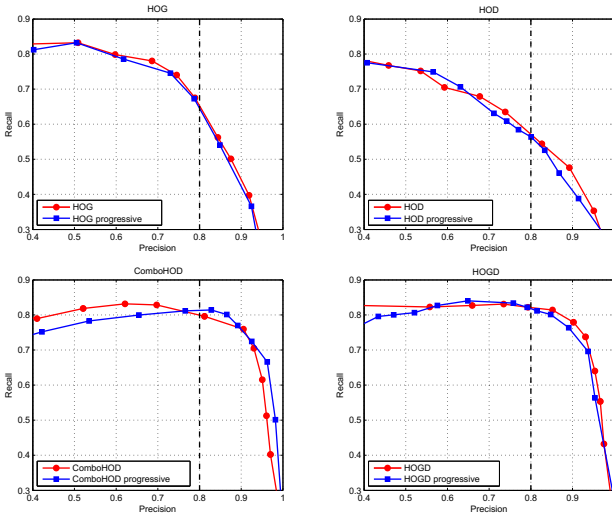
Fig. 6. ROC curves for detection performance with (blue) and without (red) progressive classification, for each HOG-based detector
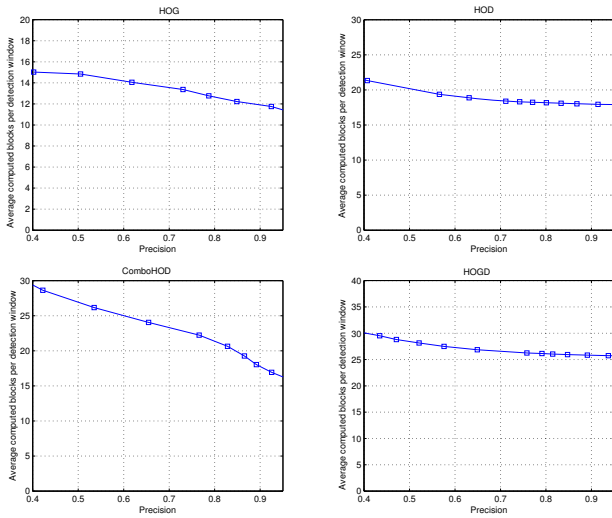
Fig. 7. Average computed blocks per image for each HOG-based descriptor using progressive classification. It must be remembered that HOG and HOD descriptors have a total of 105 blocks, whereas Combo-HOD and HOGD descriptors have 210 blocks.
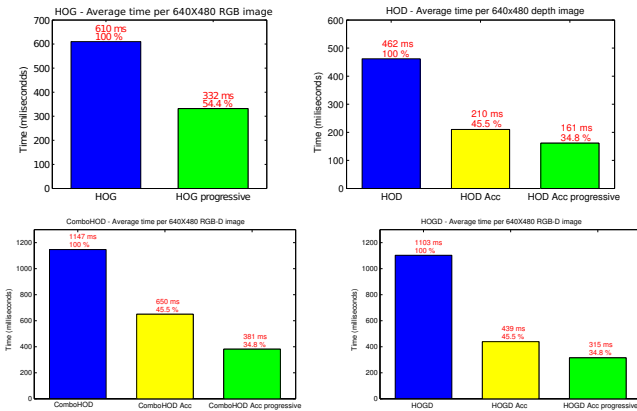
Fig. 8. Average time per image using progressive classification. Blue bars correspond to the original detection process. Yellow bars correspond to the detection process with acceleration using scale/depth information (when depth is available). Green bars show the average time with both acceleration using scale and progressive classification.

Computational Time Gains from Progressive Classification. In Fig. 6, we show the detection performance for each HOG-based detector with progressive classification. Even if this method does not compute the classification score on the whole testing window, it can be seen that the overall detection performance is not affected by making early decisions in progressive classification.

Fig. 7 shows the average number of computed blocks per detection window with progressive classification, as a function of the required precision rate (obtained for different values of the final U_f along ROC curves). Only a small percentage of the 105 blocks are finally computed, resulting in a significant reduction of computational processing. When the desired precision increases, then U_f also increases, and so does the number of early decisions. This is accentuated for the basic HOG-SVM classifier: The average number of evaluated blocks is around 15, to be compared with the 105 blocks included in the original version.

Finally, Fig. 8 depicts the average time reduction for each HOG-based descriptor. We can see, for example, that the average detection time is 45% smaller when using progressive classification for the original HOG (this time reduction is not linear with the number of processed blocks reduction, because of overheads in the algorithm). It must be remembered that for descriptors that rely on depth information, an additional acceleration technique is implemented, based on scale-depths correspondences. The average times are presented (1) without this scale-based acceleration method; (2) using this scale-technique alone (when possible, i.e. when depth is used) and (3) along with the progressive classification.

6   Conclusion

We have presented two promising contributions in the area of pedestrian detection from RGB-D images: The first one is a new descriptor combining texture

and depth elements in the form of a HOG-like descriptor, and for which we show that it exhibits better detection performances than other descriptors in the literature; the second one is a new strategy to accelerate the detection process by making early decision about the presence of a pedestrian in a test window, based on determining a specific order of evaluation of the blocks forming the HOG-like descriptors. For the descriptors we evaluated, this improvement allows significant gains in computational times, without significant losses in the performance.

References

1. Ben Shitrit, H., Berclaz, J., Fleuret, F., Fua, P.: Tracking multiple people under global appearance constraints. In: Proc. of IEEE Int. Conf. on Computer Vision, ICCV (2011)
2. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 886–893 (2005)
3. Papageorgiou, C., Oren, M., Poggio, T.: A general framework for object detection. In: Proc. of IEEE Int. Conf. on Computer Vision (ICCV), pp. 555–562 (1998)
4. Viola, P.A., Jones, M.J.: Rapid object detection using a boosted cascade of simple features. In: Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 511–518 (2001)
5. Sabzmeydani, P., Mori, G.: Detecting pedestrians by learning shapelet features. In: Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, CVPR (2007)
6. Enzweiler, M., Gavrila, D.M.: Monocular pedestrian detection: Survey and experiments. IEEE Trans. Pattern Anal. Mach. Intell. 31, 2179–2195 (2009)
7. Dollar, P., Belongie, S., Perona, P.: The Fastest Pedestrian Detector in the West. In: Proc. of British Machine Vision Conf., BMVC (2010)
8. Wang, X., Han, T.X., Yan, S.: An HOG-LBP Human Detector with Partial Occlusion Handling. In: Proc. of IEEE Int. Conf. on Computer Vision, ICCV (2009)
9. Choi, W., Savarese, S.: Multiple Target Tracking in World Coordinate with Single, Minimally Calibrated Camera. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 553–567. Springer, Heidelberg (2010)
10. Enzweiler, M., Eigenstetter, A., Schiele, B., Gavrila, D.M.: Multi-cue pedestrian classification with partial occlusion handling. In: Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 990–997 (2010)
11. Wojek, C., Schiele, B.: A performance evaluation of single and multi-feature people detection. In: Rigoll, G. (ed.) DAGM 2008. LNCS, vol. 5096, pp. 82–91. Springer, Heidelberg (2008)
12. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object Detection with Discriminatively Trained Part Based Models. IEEE Trans. on Pattern Analysis and Machine Intelligence 32 (2010)
13. Spinello, L., Arras, K.O.: People detection in rgb-d data. In: Proc. of IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, IROS (2011)
14. Tang, S., Wang, X., Lv, X., Han, T.X., Keller, J., He, Z., Skubic, M., Lao, S.: Histogram of oriented normal vectors for object recognition with a depth sensor. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) ACCV 2012, Part II. LNCS, vol. 7725, pp. 525–538. Springer, Heidelberg (2013)
15. Platt, J.C.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Advances in Large Margin Classifiers, pp. 61–74. MIT Press (1999)