A Hybrid Real-Time Vision-Based Person Detection Method

J.Oliver[1], A.Albiol[2], S.Morillas[3], G.Peris-Fajarnés[1]
[1]Centro de Investigación en Tecnologías Gráficas
Universitat Politècnica de València
[2]Instituto de Telecomunicaciones y Aplicaciones Multimedia
Universitat Politècnica de València
[3]Instituto Universitario de Matemática Pura y Aplicada
Universitat Politècnica de València
Corresponding author: jaolmol@upvnet.upv.es

Abstract

In this paper, we introduce a hybrid real-time method for vision-based pedestrian detection made up by the sequential combination of two basic methods applied in a coarse to fine fashion. The proposed method aims to achieve an improved balance between detection accuracy and computational load by taking advantage of the strengths of these basic techniques. Haar-like features combined with Boosting techniques, which have been demonstrated to provide rapid but not accurate enough results in human detection, are used in the first stage to provide a preliminary candidate selection in the scene. Then, feature extraction and classification methods, which present high accuracy rates at expenses of a higher computational cost, are applied over boosting candidates providing the final prediction. Experimental results show that the proposed method performs effectively and efficiently, which supports its suitability for real applications.

Keywords: Boosting, Histograms of Oriented Gradients, Support Vector Machine, Feature Descriptor vector, Real-Time Imaging, Person Detection

1. Introduction

The CASBlip project [1] aims to develop a Cognitive Aid System to provide guidance for blind and visually impaired people navigation in changeable environments. The system, which is composed by a pair of stereo-cameras, a linear range sensor and an inertial sensor that are placed on the blind user, provides real-time simplified acoustic maps that describe the scene. The acoustic maps consist of a spatial representation of the main elements that appear in the scene by means of the use of localized sounds.

People, free paths, moving objects and other hazards are the main objects in the scene to be sonified. This work focuses on person detection.

Real-time person detection in video sequences is a challenging task since humans may appear in a great variability of appearances, poses and illumination conditions, and in variable scenarios such as urban, traffic and cluttered environments. Furthermore, whereas most part of artificial vision systems are fixed or work under controlled movement, the problem significantly increases when the acquisition system works onboard a person.

There exists extensive literature on vision-based object detection and in particular on human detection. The problem has been handled by authors combining different feature extraction methods with different classification machines for the retrieval process, which may be carried out either using single detection window[5] or a parts-based approach [6]. On the one hand, we can find some feature extraction methods based on edge information [7], linear features, symmetry and human templates [8] [9], Haar-like features [2] [3] [10] or gradient orientation information [5] [11]. On the other hand, different classifiers such as Nearest Neighbour [16], Neural Network [17, 18], Boosting-based classifiers [19] or Support Vector Machines (SVM) [20] have been proposed for the feature classification.

Haar-like features combined with Boosting classifiers have been commonly used for face detection by Viola and Jones [2] yielding excellent results. Oren et al. [12] use Haar-like features combined with SVM [22]. Viola et al. [13] extend Haar-Boosting based methods to handle space-time information for moving-human detection. Yu-Ting Chen and Chu-Song Chen [21] use meta stages in the Boosting classifier that enhances the detection rate. On the other hand, during the last years, feature extraction methods like Histograms of Oriented Gradients (HOG) in combination with SVM classifiers have been widely used by several authors [5] [6]. Other works enhance the HOG-SVM performance by exploiting also colour, infrared and multimodal stereo information in the preliminary stages of the process [15]. Other authors use HOG features combined with Boosting-based classifiers, where SVM are used as weak classifiers in each stage of the cascade [14]. Stereo-vision has been widely used in object detection simplifying the input information of the scene, allowing background extraction and 3D image segmentation, and so improving the detection accuracy [23] [24].

Recently, researchers have been focusing on other methods to speed up the process. Cao et al. [28] propose a generalized method to train SVM-based pedestrian classifiers by means of optimizing the feature set and the classifier kernel parameters. Paisitkriangkcrai et al. [29] propose a multiple layer Boosting-based classifier that combines Haar-like and covariance features. Geismann and Schneider propose a two-stage approach to pedestrian recognition using Haar-like features and HOG features [27]. On other hand, the problem of speeding up the detection process has also been handled from a hardware perspective, where several authors [30] have designed special hardware for accelerating the Boosting classification.

In this context, there is a willing to accelerate the process while preserving a good detection rate. Previous works on different classification problems have shown that Boosting-based methods are time-efficient methods [2] whereas, on the other hand, SVM-based techniques provide more accurate results at the expense of an increase in computation cost. For our application [1], we need an effective real-time processing method. Therefore, in order to obtain an improved trade-off between detection accuracy and computational efficiency we propose a hybrid pedestrian detection method which is made up by a coarse-to-fine combination of these two basic approaches.

The rest of the paper is organized as follows: Section 2 gives a description of the method. Section 3 shows the experimental evaluation. The main conclusions are summarized in Section 4.

2. System description

The CASBLIP system provides an image flow of 320x240 pixels at 27fps that is captured on board a person. The method that we propose, which we name Haar-Boosting-HOG-SVM detection method (HBHS), takes CASBLIP images as input and performs pedestrian identification task by dividing the processing chain into two stages (see Figure 1). In the first stage, a coarse selection of candidates in the scene is provided by a Haar-Boosting method, where Haar-like features are extracted from the input image and classified using a Boosting-based classifier. In the second stage, single-window HOG descriptors are computed on each preliminary candidate and then classified using SVM in order to refine the initial findings. Predictions yielded by SVM are the final labelling of the method.

2.1 Coarse Candidate Selection Module

A fast preliminary candidate selection is performed on the 320x240 pixels input images. This coarse selection is based on the extraction and classification of the Haar-like features. These features, which are based on differences of mean intensity among adjacent rectangular groups of 16x32 pixels, are extracted by convolving different masks with the input images. Figure 1 shows the several masks that have been used in this work.

These Haar-like features are classified by means of a Boosting based classifier. Boosting techniques for classification consist of several simpler classifiers (named stages) that are subsequently applied to a feature vector extracted from an image patch. Each stage classifier of the cascade is built out of basic classifiers, which are decision tree classifiers with at least 2 leaves. If a candidate passes all stages, it is accepted as a true candidate. Otherwise, the candidate is rejected. For this work, we use Gentle Adaboost technique in the classifier training, as described in [26].

2.2 Fine Labelling Module

The preliminary set of candidates provided by the coarse module is resized to 64x128 pixels. HOG descriptors are extracted on all

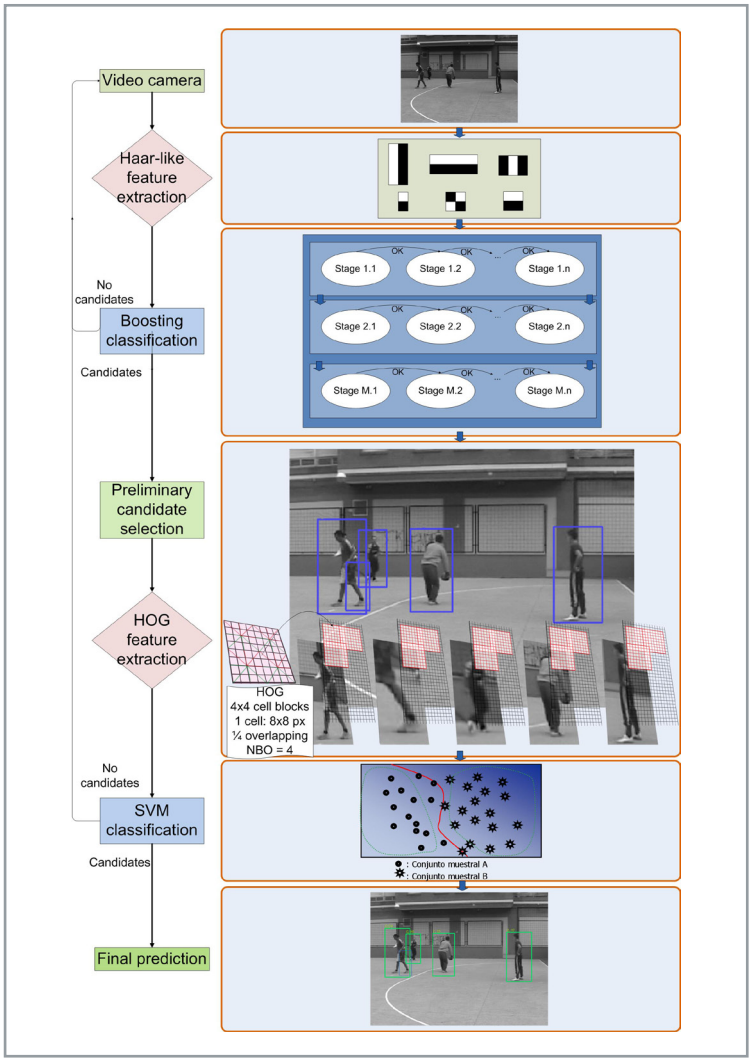

Figure 1. *Flowchart of the HBHS method*

candidates and they are classified using SVM, which generates the final labelling. SVM classification of HOG features provides an accurate identification but it is also much more computationally demanding. However, since this refined classification is performed only for the preliminary candidates, the increased computational load can be assumed within the required real-time processing.

HOG features, which are computed using a simplification of Lowe's SIFT algorithm [11], is a vector containing the accumulated histogram of the oriented gradients of the image.

Let $I(x,y)$ be the intensity of a pixel at location (x,y). Let $D_x(x,y)$ and $D_y(x,y)$ be the horizontal and vertical gradients computed on the pixel (x,y). The gradient on an arbitrary position (x,y) is described by its magnitude ρ and its orientation ϕ as follows:

$$(\rho,\phi) = (\sqrt{D^2{}_x + D^2{}_y}, \arctan(\frac{D_x}{D_y}))$$

$$(D_x, D_y) = (I_{x+1,y} - I_{x-1,y}, I_{x,y+1} - I_{x,y-1})$$

Gradients (magnitude and orientation) are computed over a dense cell grid, where cells are grouped into blocks of NBP (Number of Spatial Bins) cells. The gradient module is normalized by the highest value in the block and is spatially weighted by a Gaussian function. Blocks are overlapped in order to provide more precise description of the image patch. The gradient orientation information is discretized into NBO bins (Number of Orientation Bins). The accumulated histogram of gradients from all blocks shapes the image feature vector.

In order to improve detection speed and accuracy, we have not considered the sign of the gradient direction but the absolute value of the direction information, as claimed in [4]. This is due to the fact that the sign of the gradient does not provide further information in pedestrian detection since person clothes can be highly variable and do not yield fixed gradient prints. Therefore, the phase information would be:

$$\phi' = \phi - \pi, \quad when \phi = \pi$$

In the last step of the fine labelling module, HOG feature vectors are classified using SVM. SVM are a set of related supervised learning methods used for classification and regression. Given a training dataset with positive and negative samples in an N dimensional space, a classifier is trained by constructing a hyperplane or set of hyperplanes in a high dimensional space that separates both group of datapoints. A good training is achieved when a hyperplane has the largest distance to the nearest training datapoints of any class.

The original optimal hyperplane algorithm was proposed by Vapnik [31] in a linear approach. However, there are other recent non-linear approaches that outperform the original version. Whereas a linear classifier would separate the training set of samples with a N-1 dimensional hyperplane, the non-linear approach considers specific kernels that model the non-linear hyperplane.

In this work, several non-linear kernels such as Polynomial and Gaussian have been considered. The main parameters that characterize the kernel functions [20], such as the degree d of the Polynomial function and the exponential term gamma denoted as g for the Gaussian, have been analyzed.

3. Experimental Results

Section 3.1 discusses the best parameter setting for the Haar-Boosting module. A set of Boosting classifiers composed by 20 to 45 stages have been tested. Section 3.2 shows a comparison among a set of Polynomial and Gaussian SVM kernels for the fine module. Section 3.3 describes

the experimentation of the full-module HBHS system in real scenarios.

MIT-Pedestrian [3] database is a collection of a simple set of 509 training and 200 test images of people in different scenarios and different light conditions, but in similar poses. INRIA dataset [5] is a more challenging collection of 1239 positive and 1218 negative training and testing images that appear in different orientations and in a wide variety of backgrounds.

The coarse and the fine modules of the HBHS method have been evaluated separately. Both make use MIT-Pedestrian and INRIA datasets to train and test the classification machines individually. In both databases, left-right reflections of images have been considered in order to increase the imagery set.

The Boosting-based classification machine has been trained using 3365 positive and 2000 negative examples selected from the image databases.

The SVM classification machine has been trained using the two imagery datasets, obtaining a global dataset with 4991 positive and 16630 negative examples.

The classification machines performance is measured as the trade-off among *Accuracy, Precision, Recall,* given as $Accuracy = (FP+FN)/(TP+FN)$, $Precision = 100(1-TP/(TP+FP))$, $Recall = 100(FN/(TP+FN))$, and time *cost* parameters. The Accuracy is defined as the wrongly classified examples normalized to the number of positive testing images, being TP the true positives, FP the false positives and FN the false negatives. The Precision represents the strength of the method in detection, normalized to the number of true positives TP and false positives FP. The Recall shows the robustness of a classifier in not providing false candidates.

The full-module HBHS performance study is shown in section 3.3. We used our particular set of test sequences, which gather different scenarios, positions of the camera, backgrounds and light conditions, to test the performance. These sequences, which gather a total number of 1512 frames, sized 320x240 pixels, with a total number of 4146 pedestrians, have been acquired on board visually impaired and blind people. Sequence VLC001 covers a basketball match. Sequence VLC002 covers a football match. Sequences VLC003 and VLC004 are taken in a crowded sidewalk with several pedestrian walking and riding bikes along the street. These scenarios cover an area which is plenty of trees, street lights and other elements that make them challenging testing sequences. All sequences show people in foreground and background, and all have been grabbed at 27 fps, causing many times smooth blurring on the image that make these specific scenarios more challenging than others. Sequences are hand labelled and freely available throw the url [33].

The system performance is measured as a compromise between *FPPW (False Positives perWindow)* and *MR (Miss Rate),* where $FPPW = FP/(TotalNumbero\ fFrames)$ and $MR = 1-Recall = FN/(TP+FN)$.

This work has been implemented on an ACER single-core at 1.73 GHz, running under Ubuntu Linux Operating System, programmed in C using OpenCV [32] and svmlight [20]libraries.

3.1 Coarse module performance

A study of the Boosting classifier performance depending on the number of stages has been carried out using MIT and INRIA dataset collection. Table 1 shows that the results of boosting method by itself depends on the number of stages of the classifier. We find that for those classifiers with low number of stages the method provides high number of false alarms, whereas classifiers composed by more stages provide better results in terms of Precision. Regarding the Recall, we find that the highest value is obtained for 30 stages with a value of 80.60%. However, for 35 stages we obtain nearly the same Recall and Precision improves considerably. The last row in Table 1 shows that for all the different Boosting configurations, the computation cost is very similar and is less than 50 ms per frame. This behavior is due to the fact that the great amount of image patches belong to background and they are discarded in earlier stages of the cascade, where only a little percentage of these patches will contain a true candidate. Note that the mean computation time is mostly affected by the background patches classification, and the true candidates slightly affect this value. This is the main reason for the mean computation time to be similar to all configurations.

The main objective of this method is to provide good results in terms of Recall since it constitutes the first stage of the process, which is a coarse selection. The more accuracy on the prediction and the lower number of false candidates directly affect the SVM computation time, which considerably increases with higher number of test candidates. Therefore, the final study of the proposed HBHS method will determine the optimal combination.

3.2 Fine module performance

3.2.1 HOG parameters discussion

There are several parameters in the HOG computation, such as cell size (measured in pixels), block size (measured in cells), NBO and block overlapping, which determine the description capability of an image descriptor vector. An optimal relationship among these parameters that optimize the detection rate is given in at Dalal's work [4] and is found for blocks of 3x3 cells, cells of 6x6 pixels and 12 orientation bins. However, temporal optimization is not handled

The combination of Haar-Boosting and HOG-SVM methods exploits the stregths of each individual method

in it. Our study, though, aims not only to provide the best detection rate but also to an acceptable computational cost. This study has been carried out using 90% of imagery dataset for training a simple SVM polynomial kernel (d = 3) and 10% for testing.

First, we analyze the effect of considering boundary blocks in the image descriptor computation, setting by default: $NBP=4$, $NBO=4$, $\sigma = 8$ and block overlapping=1/2. Experimental results on the imagery dataset show that when boundary blocks are not taken into account, slightly worse accuracy results are obtained. $Accuracy$ only decreases 0.19% with respect to the integral case (from 99.20% to 99.01%). However, the number of features in the case of the boundary block reduction is significantly smaller than in the integral case, decreasing from 3200 to 2304 dimensions of the feature vector, which has direct consequences in the computational cost. This effect is due to the lower number of blocks needed to cover the whole image. The number of features of an image depends on the square of the block size, the number of orientation bins and the number of blocks to cover the image with ½ overlapping. Therefore, the boundary blocks reduction must be considered to speed up the process at expense of the decrease of 0.19 in the SVM accuracy. The improvement in computational cost is worth the insignificant decrease in accuracy.

Second, we analyze the effect of the block overlapping, considering ½, ¼ and ¾ overlapping, values which are proposed by Dalal. NBP, NBO and σ were set to 4, 4 and 8 respectively. Table 2 shows that the best trade-off between detection rate and computational cost were obtained for ¼ block overlapping, where detection rate was optimal and time cost per sample (including feature extraction and classification time) significantly drops. Achieving an optimal computation cost turns out to be essential for an optimal HBHS overall performance.

Furthermore, we analyze the effect of the block size. According to Dalal, a suboptimal relationship among block size and cell size that still yield acceptable $Recall$ is obtained for blocks of 2x2 and 4x4 cells where each cell comprises a group of 8x8 pixels. Our analysis was carried out comparing NBP = 4, 8 considering NBO=4, and showed that optimal values regarding detection rate and time cost were found for NBP=4, where accuracy on test raised from 99.20 to 99.32 and computation cost per sample dropped from 5.8 to 5.2ms.

	Nstg 20	Nstg 25	Nstg 30	Nstg 35	Nstg 40	Nstg 45
Prec (%)	10.53	15.29	22.86	34.28	50.07	69.14
Rec (%)	59.94	79.33	80.60	79.86	75.52	72.95
Tcost (ms)	48.8	48.4	49.0	47.3	49.2	48.2

Table 1. *Boosting method performance depending on the number of stages of the classifier. Optimal values in terms of computational cost and recall are found for 30 and 35 stages.*

Overlap	Accuracy (%)	Precision (%)	Recall (%)	Num SV	Feat Ext (ms)	Feat Clas (ms)
¾	98.80	90.26	98.52	4160	58.49	21.90
½	99.01	91.91	98.63	2304	17.87	12.90
¼	99.66	93.01	98.72	512	8.47	6.00

Table 2. *Analysis of the influence of block overlapping.*

d	Accuracy (%)	Precision (%)	Recall (%)	Num SV	Runtime (ms)
2	99.57	99.32	93.77	3842	12.3
3	99.57	99.61	93.57	4163	13.1
4	99.58	99.63	93.63	4506	14.1
5	99.57	99.64	93.53	4853	15.2
6	99.58	99.78	93.53	5202	16.4
7	99.57	99.96	93.26	5581	17.7
8	99.56	100	92.92	5973	18.8
9	99.55	100	92.79	6380	20.0
10	99.53	100	92.62	6821	21.7

Table 3. *Analysis of a linear SVM kernel performance for the HBHS method.*

3.2.2 SVN kernel selection

This section discusses the selection of the most appropriate SVM kernel for the HBHS system. We trained and tested two different kernels: Polynomial and Gaussian. Gaussian kernels are known to generally outperform the linear kernels in terms of Precision and Recall, although they are slower too. Many authors [5] [6] propose to use the linear ones for pedestrian recognition since they still achieve high detection rates and significantly reduce the computation time. However, in our particular application, where we use a preliminary rapid selection of candidates, a Gaussian function may perform better.

In order to make a prior decision on the optimal parameters of this classifier, we generated ten different scenarios from the imagery collection. In each scenario there is a dataset for training and other for testing. In order to make robust enough pedestrian classifiers using SVM, the authors have performed cross-validation with the ratio of training/testing images equal to 9/1, since using less training images is not enough due to the high dimension of HOG vectors. Note that people may appear in many different conditions, so it is necessary to build a classifier that integers so many different samples as possible in order to make the classifier strong enough to be successfully tested. To overcome

this and to avoid any bias in the test, we have created 10 different scenarios, all coming from the same image dataset, where each time a different set of images for training and testing was randomly selected. Thus, we have different training and testing packages each time, which gives us less dependency on the collection selected for the classification.

Table 3 shows the performance of a linear kernel varying the parameter d from 2 to 10. Precision increases with d, but *Recall* significantly drops too. Time cost also gets worse while increasing *d*. A good compromise between *Precision* and *Recall* are yield by this kernel for *d* values comprised between 4 and 7.

Table 4 shows the results on the Gaussian kernel study, sweeping g from 0.05 to 1. Precision reaches the highest value at $g=0.25$, still with acceptable *Recall* rates. *Recall* monotonically decreases when g increases. Focusing on time cost, middle rows show very low cost values at $g=0.175$ and $g=0.2$.

The best accuracy rates are obtained for kernels with $g=[0.15\text{-}0.3]$, where kernels with relative higher values of gamma achieve Precision values of 100% but Recall slightly drops and the computation cost takes more than 10 ms with respect to the optimal value. On other hand, lower values of gamma (g=0.05) yield the best Recall value, 93.90%, but Precision decreases to 99.25%. The best balance among these parameters is obtained for $g=0.2$, where $Precision=99.93\%$, $Recall=93.70\%$, $Accuracy=99.6\%$ and the runtime is 11ms. It is important to remark that we prefer higher Precision values rather than higher *Recall* values since SVM will work in the second stage of the process and will have to yield low number of false positive images.

All in all, *Gaussian* kernels can produce the highest accurate results with similar computation cost to the linear ones. Therefore, it is strongly motivated to use radial basis kernels for our particular application. Here on we will use a radial basis kernel with $g=0.175$.

g	Accuracy (%)	Precision (%)	Recall (%)	Num SV	Runtime (ms)
0.05	99.57	99.25	93.90	3967	12.9
0.1	99.58	99.46	93.83	4214	14.0
0.15	99.60	99.78	93.80	4457	15.8
0.175	99.60	99.82	93.73	4589	11.1
0.2	99.60	99.93	93.70	4717	11.4
0.25	99.60	100	93.63	4978	16.9
0.3	99.59	100	93.53	5233	18.1
0.5	99.55	100	92.26	6371	21.6
0.75	99.51	99.89	92.05	8079	28.9
1	99.43	99.63	90.40	10563	37.8

Table 4. *Analysis of a Gaussian SVM kernel performance for the HBHS method.*

3.3 HBHS performance

In this section we test the performance of the HBHS method. Several stages have been tested for the boosting classifier. HOG method parameters have been set to: NBP=4, NBO=4 0-180, block overlapping= 1/4, feature vector=512 components for a 64x128 window. A Radial Basis kernel with g=0.175 has been used for the SVM classifier. The optimal threshold in the classification has been studied. SVM candidates are considered as true positive when the centre of its bounding box has a maximum deviation of 15% of the horizontal and vertical size of the groundtruth, and the ratio between candidates and groundtruth labelling areas is lower than 2.

Figure 2 shows a comparison between HBHS and Dalal's method for different sequences. For all the scenarios, both Dalal and HBHS methods present similar rates of MR and FPPW. However, HBHS achieves better results in sequences VLC002, VLC003 and VLC004, excepting for Sequence VLC001, where Dalal method clearly outperforms the HBHS.

Focusing on the HBHS performance, the best coarse module configuration that yields the best HBHS results is found for Nstg=30, which generally provides good rates in all scenarios. The other coarse modules do no present stable rates and yield significant bad rates for several sequences.

We see that the Boosting stage involves a constraint

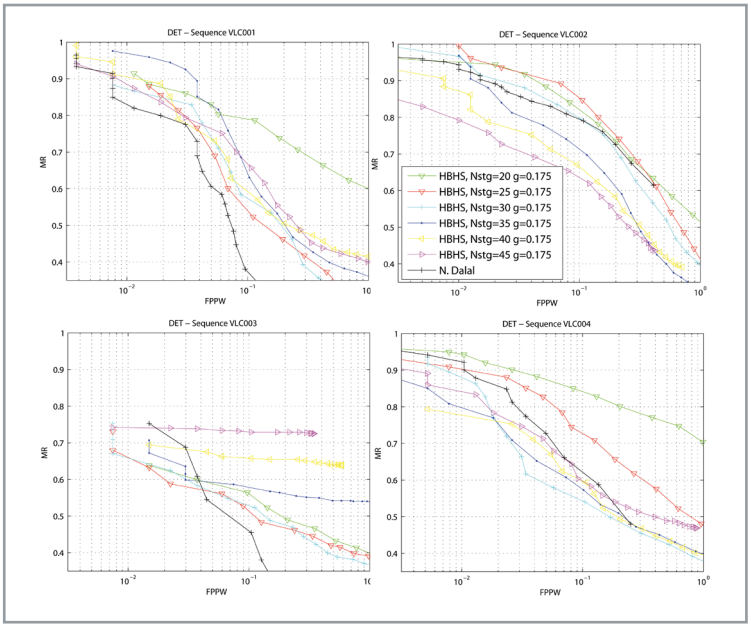

Figure 2. *Comparison among several HBHS classifiers varying the number of stages in the coarse module from 25 to 45, and Dalal's method. In colored lines, HBHS performance for the different coarse modules. In black, N. Dalal performance. Top-left figure shows results from a basketball match sequence. Top-right figure corresponds to a football match. Bottom-left and bottom-right figures correspond to crowded sidewalks.*

HOG-SVM module, wich works on the Boosting candidates, provides accurate labelling

	Nstg 20	Nstg 25	Nstg 30	Nstg 35	Nstg 40	Nstg 45
Boosting candidates	8.84	4.8	3.24	2.27	1.97	1.64
Boosting time (ms)	48.8	48.4	49.0	47.3	49.2	48.2
HOG time (ms)	141.6	78.1	52.8	36.4	31.9	25.8
SVM time (ms)	314.9	169.7	116.8	80.4	71.4	57.9
HBHS time (ms)	505.5	296.3	218.7	164.2	152.6	131.9

Table 5. *HBHS temporal analysis for different number of Boosting stages and a Gaussian kernel with g=0.175. Mean values per window.*

in the maximum performance achievable in the MR parameter since those true pedestrians that are missed in the first stage can not be retrieved by the following. On the other hand, SVM module significantly enhances the FPPW rates, although a good SVM classification strongly depends on a good framing of the Boosting classifier. A study on the computation cost of each of these methods is necessary to complete the study. Table 5 shows mean values of the HBHS time cost for the same scenario. Row 1 shows the number of candidates per image selected by the boosting method and Rows 2-4 show detailed averaged time information of each processing module during the scanning of the 320x240 pixels input images. It is easy to find that SVM method is the bottle neck of the process in terms of computational cost. Therefore, the lower number of candidates provided by Boosting positively affects the SVM computational cost. Note that for lower number of stages, more candidates are yielded by boosting and therefore higher computational cost is needed by SVM. However, for those situations with more stages in Boosting classifier, the necessary time for SVM significantly decreases. See that there is a significant step in the final temporal cost for a boosting classifier with 30 and 35 stages. The cost decreases in 50 ms for 35 stages, which is really significant. However, a classifier based on 40 stages does insignificantly reduce the cost in comparison to the 35 stages. According to the figure 2, the best results were found for 30, 35 and 40 stages. On the other hand, results on time study reveal that a classifier based on 35 stages is preferred rather than the 30 one. Although classifiers with 40 and 45 stages provide good temporal results, FPPW and MR are not good enough. Therefore, the best combination of Accuracy and Temporal Cost is reached for a boosting classifier with 35 stages.

Figure 3 shows some examples of the HBHS method performances over our collected dataset for different stages of boosting method. Boosting candidates are surrounded by blue boxes. SVM contribution and therefore the final prediction is represented in green colour. The visual analysis reveals that sometimes the positive boosting candidates are not well framed. Then, the feature extraction is carried out over an image where the candidate is not centred in the window. SVM is more sensitive to changes than boosting, and therefore a slight drift in the position implies a wrong classification. The problem increases for high stages of boosting classifiers. For that reason, there are several occasions where SVM fails in the labelling. However, the final labelling is accurate and fast enough for real-time processing of complex scenarios.

3.4 HBHS contribution

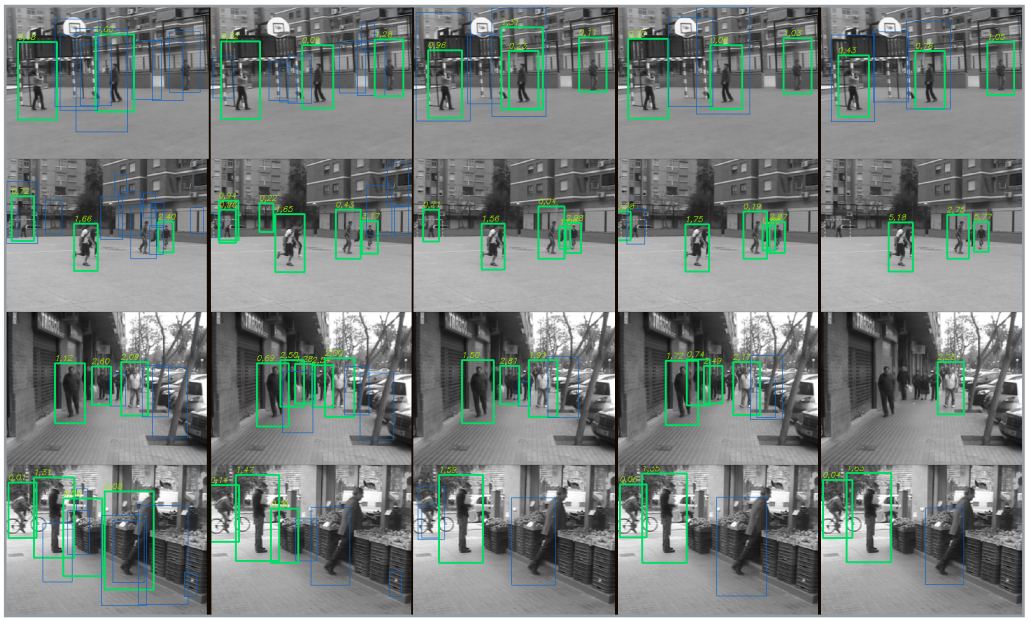

Figure 3. *HBHS performance over our particular imagery collection. In blue, coarse module candidates that have been rejected by the fine module. In green, the positive predictions of the fine module that represents the final prediction of the HBHS. Each column represents a HBHS method with different number of stages in the coarse module, varying from 25 (left column) to 45 (right column)*

The HBHS method has been compared to two well-known methods that have been widely tested by the great majority of similar works: an implementation of a Haar- Boosting method based on [26] and the HOG-SVM method implemented and provided by Dalal [4] [5]. The Haar-Boosting settings are the same than those used in the coarse stage of the HBHS method, which have been cited in section 2.1. Dalal's method uses dense rectangular HOG features with 9 orientation bins in 0º-180º, 16x16 pixel blocks of four 8x8 cells, a Gaussian spatial window with σ = 8pixel and block spacing stride of 8 pixels. A linear SVM kernel has been used for classification. HBHS is set with a 35-stage boosting classifier in the coarse module and dense rectangular HOG features with 4 orientation bins in 0o-180o, 4x4 pixel blocks of 32x32 pixel blocks of 4x4cells, a Gaussian spatial window with σ = 8pixel, a block overlapping of 1/4 and a Gaussian svm kernel with $g=0.175$.

Table 6 shows the comparison among these cited methods. The best rates of FPPW are achieved by HBHS. However, MR significantly increases in HBHS due to the accumulation of false negatives from both coarse and fine modules. Boosting method presents good MR values and the lowest computational cost, but FPPW is too high. On the other hand, Dalal's method yields good FPPW and good MR values, but temporal requirements are significant, too. HBHS, however, provides similar detection rates to Dalal and still presents low computational cost, 10 times lower than HOG-SVM requirements, achieving a cycle time that entirely fulfills the temporal requirements for the CASBLIP navigation assistance system. Therefore, the proposed system may considered to work in real-time.

	FPPW		MR		Time (ms)	
	mean	std dev	mean	std dev	mean	std dev
Boosting	2.25	0.91	0.38	0.06	48	3
Dalal's HOG-SVM	0.28	0.12	0.39	0.17	1990	5
HBHS	0.23	0.11	0.54	0.15	138	35

Table 6. *Comparison of HBHS among other methods. HBHS has been trained with 35 Boosting stages and Gaussian SVM kernel with g=0.175*

4. Conclusions

In this paper we present a hybrid real-time method to detect people in image sequences of diverse nature. This method has been devised with the aim of improving the balance detection accuracy and computational cost of the state-of-the-art methods. The proposed method comprised the sequential combination of two basic methods performing in a coarse-to-fine fashion. A fast coarse preliminary candidate selection is provided by boosting techniques. Then, SVM-based classification of HOG features extracted from the preliminary candidates refines the selection and provides the final detection. Extensive experimental results using real sequences have been carried out both for parameter adjustment and performance assessment. It can be seen that the proposed method obtains an improved trade-off between computational cost and detection accuracy.

Several modifications that may enhance the overall performance of the method might be the following: First, tracking techniques may be considered in order to improve the high miss rate and the slightly deviations of the candidate bounding box that Adaboost produces; Second, other implementations of SVM may be used to reduce the high computational cost of the fine classification. Third, stereo-cameras may be used in order to perform the preliminary candidate selection and intelligent scan considering the expected size of a person at each depth.

Acknowledgments

This work is supported by CASBLIP project 6-th FP \cite{RefCASBLIP}. The authors acknowledge the support of the Technological Institute of Optics, Colour and Imaging of Valencia - AIDO. Dr. Samuel Morillas acknowledges the support of Generalitat Valenciana under grant GVPRE/2008/257 and Universitat Politècnica de València under grant Primeros Proyetos de Investigación 13202. }

References

[1] http://www.casblip.upv.es FP6-2004-IST-4
[2] P. Viola and M. Jones: 'Rapid object detection using a boosted cascade of simple features', IEEE Computer Vision and Pattern Recognition. 25, 29. (2001)
[3] A. Mohan, C. Papageorgiou and T. Poggio: 'Example-based object detection in images by components', IEEE Trans. Pattern Anal. Mach. Intell., vol. 23, no. 4, pp. 349-361, Apr. 2001.
[4] N. Dalal, Finding People in Images and Videos: 'Thesis Doctoral', July 2006.
[5] N. Dalal and Bill Triggs: 'Histograms of Oriented Gradients for Human Detection', Proceedings of IEEE Conference Computer Vision and Pattern Recognition, San Diego, USA, pages 886-893, June 2005.
[6] I. Parra, D. Fernandez, M. A. Sotelo, L. M. Bergasa, P. Revenga, M. Oca˜na, M. A. Garcia: 'Combination of Feature Extraction Methods for SVM Pedestrian Detection', IEEE Trans. Intell. Trans. Sys., vol. 8, no. 2, June 2007.
[7] D. Gavrila and V. Philomin: 'Real-time object detection for "smart" vehicles', Proc IEEE Int. Conf. Comput. Vis, 1999, pp. 87-93.
[8] A. Broggi, M. Bertozzi, A. Fascioli and M. Sechi: 'Shape-based pedestrian detection', Proc. IEEE Intell. Veh. Symp. Dearborn, MI. Oct. 2000, pp.215-220.
[9] M. Bertozzi, A. Broggi, R. Chapuis, F. Chausse, A. Fascioli and A. Tibaldi: 'Shape-based

pedestrian detection and localization', Proc IEEE ITS Conf, Shanghai, China. Oct. 2003, pp. 328-333.
[10] C. Papageorgiou and T. Poggio: 'A trainable system for object detection', International Journal of Computer Vision, 38 (1): 15-33, 2000.
[11] D. Lowe: 'Distinctive image features from scale-invariant keypoints', International Journal of Computer Vision, 60 (2):91-110, 2004.
[12] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna and T. Poggio: 'Pedestrian Detection Usng Wavelet Templates', Computer Vision Pattern Recognition, 1997.
[13] P. Viola, M. Jones and D. Snow: 'Detecting pedestrians using patterns of motion and appearance', International Conference on Computer Vision, 2003.
[14] Qiang Zhu, Shai Avidan, Mei-Chen Yeh and Kwang-Ting Cheng: 'Fast Human Detection Using a Cascade of Histograms of Oriented Gradients', Proceedings of the IEEE Computer Vision and Pattern Recognition 2006.
[15] Stephen J. Krotosky and Mohan Manubhai Trivedi: 'On Color-, Infrared-, and Multimodal-Stereo Approaches to Pedestrian Detection', IEEE Transactions on Intel. Trans. Syst., vol. 8, no. 4, December 2007.
[16] T. Hastie, R. Tibshirani: 'Discriminant Adaptive Nearest Neighbor Classification', IEEE Transactions on Pattern Analysis and Machine Intelligence. Volume 16, Issue 6. Pages: 607-616. June 1996.
[17] G. Peter Zhang: 'Neural Networks for Classification: A Survey', IEEE Transactions on Systems, Man and Cybernetics-Part C: Applications and reviews, Volume 30, no. 4, November 2000
[18] Lawrence S., Giles C.L., Ah Chung Tsoi, Back A.D.: ' Face recognition: a convolutional neural-network approach', IEEE Transactions on Neural Networks, Volume 8, no. 1, pages 98-113, January 1997
[19] Y. Freund, R. E. Schapire: 'A decision-theoretic generalization of on-line learning and an application to boosting', Proceedings of the Second European Conference on Computational Learning Theory. Barcelona, March 1995
[20] T. Joachims: 'Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning, B. Sch¨olkopf and C. Burges and A. Smola', MIT-Press, 1999. http://svmlight.joachims.org/
[21] Yu-Ting Chen, Chu-Song Chen: 'Fast Human Detection Using a Novel Boosted Cascading Structure With Meta Stages', IEEE Transactions on Image Processing, Volume 17, no. 8, August 2008
[22] Zijian Yuan, Lei Yang, Yanyun Qu, Yuehu Liu, Xinchun Jia: 'A Boosting SVM Chain Learning for Visual Information Retrieval', Advances in Neural Networks - ISNN 2006. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, Volume 3971, May 2006.
[23] Fengliang Xu, Kikuo Fujimura: 'Human Detection Using Depth and Gray Images', Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance. July 2003
[24] R. Munoz-Salinas, E. Aguirre and M. Garcia-Silvente: 'People detection and tracking using stereo vision and color', Image and Vision Computing 25 (6), pp. 995-1007, June 2007.
[25] Y. Freund, R. E. Schapire: 'A Short Introduction to Boosting', Journal of Japanese Society for Artificial Intelligense. 14(5): 771-780, September 1999
[26] R. Lienhart, A. Kuranov, V. Pisrevsky: 'Empirical analysis of detection cascades of boosted classifiers for rapid object detection', MRL Technical Report, pp: 21, 28, 76, 2002
[27] P. Geismann, G. Schneider: 'A Two-staged Approach to Vision-based Pedestrian Recognition Using Haar and HOG Features', IEEE Intelligent Vehicles symposium, June 4-6, 2008
[28] XB Cao, YW Chu, D. Chen, H. Qiao.: 'Associated evolution of a support vector machine-based classifier for pedestrian detection', Information Sciences, 179 (8), pp. 1070-107, March 29, 2009.
[29] S. Paisitkriangkcrai, CH. Shen, J. Zhang.: 'Fast pedestrian detection using a cascade of boosted covariance features', IEEE Transactions on Circuits and Systems for Video Technology, 18 (8), pp. 1140-1151, August 2008.
[30] M. Hiromoto, H. Sugano, R. Miyamoto: 'Partially Parallel Architecture for AdaBoost-Based Detection With Haar-Like Features', IEEE Trans on Circuits and Systems for Video Technology. 19 (1), pp. 41-52, January 2009
[31] V. Vapnik: 'The Nature of Statistical Learning Theory', 314p. Springer 2000
[32] http://sourceforge.net/projects/opencvlibrary/
[33] http://gpiserver.dcom.upv.es/alalbiol/HBHS/HBHS.htm

Biographies

Javier Oliver was born in Valencia (Spain) in 1983. He received the M. Sc. Degree in Electrical and Electronic Engineering from Universidad Politécnica de Valencia (UPV) in 2006. He has worked for several Computer Vision companies. Currently he is a PhD candidate at UPV. His research interests include Computer Vision, pattern recognition, motion analysis, multiple view geometry, 3D articulated object reconstruction and tracking, design and development of tailored industrial artificial vision systems.

Alberto Albiol graduated in Telecommunications Engineering at the Universidad Politécnica de Valencia (UPV) in 1995. From 1999 until 2003 he stayed several times in Purdue University, IN (USA) pursuing his PhD on multimodal video indexing, which finally obtained from UPV in 2004. He started his teaching activities in 1995 in the University of Zaragoza and currently, he is Associate Professor of Signal and Image Processing at UPV. His research interests include video analysis, image face recognition, pattern recognition and computer vision applications. He is author of more than 26 refereed publications in journals and in international conferences, and actively participates on the review process of many related journal and conference proceedings.

Samuel Morillas was born in Granada (Spain) in 1979. He recieved the M.Sc. Degree in Computer Science from the Universidad de Granada in 2002, and the Ph.D. degree from the Universidad Politécnica de Valencia (Spain) in 2007. Currently he is an associated professor at the Department of Applied Mathematics of the Universidad Politécnica de Valencia and a member of the Instituto de Matemática Pura y Aplicada. His research interests include fuzzy metrics, fuzzy sets, nonlinear image and video processing and medical imaging.

Guillermo Peris-Fajarnés Born in Valencia (Spain) 1967. PhD. Full professor in the Graphic Engineering department at the Universitat Politécnica de Valencia. Head of the Centre of graphic Technologies. Industries and main areas: Tile Industry, Graphic Industry, Printing Industries, Internet, Security, e-learning, cognitive-cognition and accessibility. www.citg.es