Combination of Feature Extraction Methods for SVM Pedestrian Detection

Ignacio Parra Alonso, David Fernández Llorca, Miguel Ángel Sotelo, *Member, IEEE*, Luis M. Bergasa, *Associate Member, IEEE*, Pedro Revenga de Toro, Jesús Nuevo, Manuel Ocaña, and Miguel Ángel García Garrido

***Abstract*—This paper describes a comprehensive combination of feature extraction methods for vision-based pedestrian detection in Intelligent Transportation Systems. The basic components of pedestrians are first located in the image and then combined with a support-vector-machine-based classifier. This poses the problem of pedestrian detection in real cluttered road images. Candidate pedestrians are located using a subtractive clustering attention mechanism based on stereo vision. A components-based learning approach is proposed in order to better deal with pedestrian variability, illumination conditions, partial occlusions, and rotations. Extensive comparisons have been carried out using different feature extraction methods as a key to image understanding in real traffic conditions. A database containing thousands of pedestrian samples extracted from real traffic images has been created for learning purposes at either daytime or nighttime. The results achieved to date show interesting conclusions that suggest a combination of feature extraction methods as an essential clue for enhanced detection performance.**

***Index Terms*—Features combination, pedestrian detection, stereo vision, subtractive clustering, support vector machine (SVM) classifier.**

I. Introduction

THIS PAPER describes a comprehensive combination of feature extraction methods for vision-based pedestrian detection in Intelligent Transportation Systems (ITS). Vision-based pedestrian detection is a challenging problem in real traffic scenarios since pedestrian detection must perform robustly under variable illumination conditions, variable rotated positions and pose, and even if some of the pedestrian parts or limbs are partially occluded. An additional difficulty is given by the fact that the camera is installed on a fast-moving vehicle. As a consequence of this, the background is no longer static, and pedestrians significantly vary in scale. This makes the problem of pedestrian detection for ITS quite different from that of detecting and tracking people in the context of surveillance applications, where the cameras are fixed and the background is stationary.

To ease the pedestrian recognition task in vision-based systems, a candidate selection mechanism is normally applied. The selection of candidates can be implemented by performing an object segmentation in either a 3-D scene or a 2-D image plane. Not many authors have tackled the problem of monocular pedestrian recognition [1]–[3]. The advantages of the monocular solution are well known. It constitutes a cheap solution that makes mass production a viable option for car manufacturers. Monocular systems are less demanding from the computational point of view and ease the calibration maintenance process. On the contrary, the main problem with candidate selection mechanisms in monocular systems is that, on average, they are bound to yield a large amount of candidates per frame in order to ensure a low false negative ratio (i.e., the number of pedestrians that are not selected by the attention mechanism). Another problem in monocular systems is the fact that depth cues are lost unless some constraints are applied, such as the flat terrain assumption, which is not always applicable. These problems can be easily overcome by using stereo vision systems, although other problems arise such as the need to maintain calibration and the high computational cost required to implement dense algorithms.

In this paper, we present a full solution for pedestrian detection at daytime, which is also applicable, although constrained, to nighttime driving. Other systems already exist for pedestrian detection using infrared images [4]–[6] and infrared stereo [7]. Nighttime detection is usually carried out using infrared cameras as long as they provide better visibility at night and under adverse weather conditions. However, the use of infrared cameras is quite an expensive option that makes mass production an untraceable problem nowadays, especially for the case of stereo vision systems where two cameras are needed. They provide images that strongly depend on both weather conditions and the season of the year. Additionally, infrared cameras (considered as a monocular system) do not provide depth information and need periodic recalibration (normally once a year). In principle, the algorithm described in this paper has been tested using cameras in the visible spectrum. Nonetheless, as soon as the technology for night-vision camera production becomes

Manuscript received February 20, 2006; revised July 12, 2006, October 3, 2006, and December 11, 2006. This work was supported in part by the Spanish Ministry of Education and Science under Grants DPI2002-04064-C05-04 and DPI2005-07980-C03-02 and in part by the Spanish Ministry of Public Works under Grant FOM2002-002. The Associate Editor for this paper was N. Papanikolopoulos.

The authors are with the Department of Electronics, Escuela Politécnica Superior, University of Alcalá, Madrid 28801, Spain (e-mail: parra@depeca.uah.es; llorca@depeca.uah.es; miguel.sotelo@uah.es; bergasa@depeca.uah.es; revenga@depeca.uah.es; jnuevo@depeca.uah.es; mocana@depeca.uah.es; garrido@depeca.uah.es).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITS.2007.894194

cheaper, the results could easily be extended to a stereo night-vision system.

Concerning the various approaches proposed in the literature, most of them are based on shape analysis. Some authors use feature-based techniques, such as recognition by vertical linear features, symmetry, and human templates [2], [8], Haar wavelet representation [9], [10], hierarchical shape templates on Chamfer distance [3], [11], correlation with probabilistic human templates [12], sparse Gabor filters and support vector machines (SVMs) [13], graph kernels [14], motion analysis [15], [16], and principal component analysis [17]. Neural-network-based classifiers [18] and convolutional neural networks [19] are also considered by some authors. In [4], an interesting discussion is presented about the use of binary or gray-level images as well as the use of the so-called hotspots in infrared images versus the use of the whole candidate region containing both the human body and the road. Using single or multiple classifiers is another topic of study. As experimentally demonstrated in this paper and supported by other authors [1], [4], [20], the option of multiple classifiers is definitely needed. Another crucial factor, which is not well documented in the literature, is the effect of pedestrian bounding box accuracy. Candidate selection mechanisms tend to produce pedestrian candidates that are not exactly similar to the pedestrian examples that were used for training in the sense that online candidates extracted by the attention mechanism may contain some part of the ground or may cut the pedestrians' feet, arms, or heads. This results in significant differences between candidates and examples. As a consequence, a decrease in Detection Rate (DR) takes place. The use of multiple classifiers can also provide a means to cope with day and nighttime scenes, variable pose, and nonentire pedestrians (when they are very close to the cameras). In sum, a single classifier cannot be expected to robustly deal with the whole classification problem.

In the last years, SVMs have been widely used by many researchers [1], [9], [10], [20], [21] as they provide a supervised learning approach for object recognition as well as a separation between two classes of objects. This is particularly useful for the case of pedestrian recognition. Combinations of shape and motion are used as an alternative to improve the classifier robustness [1], [22]. Some authors have demonstrated that the recognition of pedestrians by components is more effective than the recognition of the entire body [10], [21]. In our approach, the basic components of pedestrians are first located in the image and then combined with an SVM-based classifier. The pedestrian searching space is reduced in an intelligent manner to increase the performance of the detection module. Accordingly, road lane markings are detected and used as the main guidelines that drive the pedestrian searching process. The area contained by the limits of the lanes determines the zone of the real 3-D scene from which pedestrians are searched. In the case where no lane markings are detected, a basic area of interest is used instead of covering the front part ahead of the ego-vehicle. A description of the lane marking detection system is provided in [23]. The authors have also developed lane tracking systems for unmarked roads [24], [25] in the past. Nonetheless, a key problem is to find out the most discriminating features in order to significantly represent pedestrians. For this purpose, several feature extraction methods have been implemented, compared, and combined. While a large amount of effort in the literature is dedicated to developing more powerful learning machines, the choice of the most appropriate features for pedestrian characterization remains a challenging problem nowadays to such an extent that it is still uncertain how the human brain performs pedestrian recognition using visual information. An extensive study of feature extraction methods is therefore a worthwhile topic for a more comprehensive approach to image understanding.

The rest of the paper is organized as follows: Section II provides a description of the candidate selection mechanism. Section III describes the component-based approach and the optimal combination of feature extraction methods. In Section IV, the SVM-based pedestrian classification system is presented. In Section V, the multiframe validation and tracking system is described. The implementation and comparative results achieved to date are presented and discussed in Section VI. Finally, Section VII summarizes the conclusions and future work.

II. Candidate Selection

An efficient candidate selection mechanism is a crucial factor in the global performance of the pedestrian detection system. The candidate selection method must assure that no misdetection occurs. Candidates, which are usually described by a bounding box in the image plane, must be detected as precisely as possible since the detection accuracy has a remarkable effect on the performance of the recognition stage, as demonstrated in Section VI. In order to extract information from the 3-D scene, most authors use disparity map techniques [18] as well as segmentation based on v-disparity [20], [26]. The use of disparity-based techniques is likely to yield useful results in open roadways. However, depth disparity clues are unlikely to be useful for segmenting out pedestrians in city traffic due to the heavy disparity clutter. We disregarded this option because of the disadvantages associated with disparity computation algorithms, since the image pair has to be rectified prior to the disparity map generation to ensure good correspondence matching. In addition, the computation of accurate disparity maps requires fine grain texture images in order to avoid noise generation. Otherwise, disparity-based methods are prone to produce many outliers that affect the segmentation process. Concerning the v-disparity image, the information for performing generic obstacles detection is defined with a vertical line. This implies managing very little information to detect obstacles, which may work well for big object detection, such as vehicles [26], but might not be enough for small thin object detection, such as pedestrians. Conversely, we propose a candidate selection method based on the direct computation of the 3-D coordinates of relevant points in the scene. Accordingly, a nondense 3-D geometrical representation is created and used for candidate segmentation purposes. This kind of representation allows for robust object segmentation whenever the number of relevant points in the image is high enough. A major advantage is that outliers can be easily filtered out in 3-D space, which makes the method less sensitive to noise.

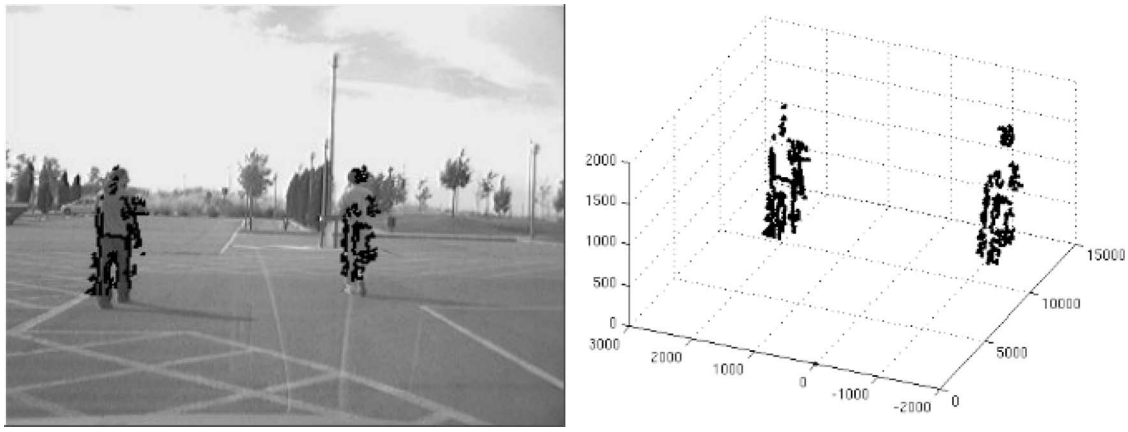


Fig. 1. (Left) Two-dimensional points overlayed on left image. (Right) Three-dimensional coordinates of detected pixels.

A. Three-Dimensional Computation of Relevant Points

The 3-D representation of relevant points in the scene is computed in two stages. In the first stage, the intensities of the left and right images are normalized, and the radial and tangential distortions are compensated for. Relevant points in the image are extracted using a well-known Canny algorithm with adaptive thresholds. Features such as heads, arms, and legs are distinguishable, when visible, and are not heavily affected by different colors or clothes. In the second stage, a 3-D map is created after solving the correspondence problem. The matching computational cost is further reduced in two ways. First, the matching searching area is greatly decreased by using the parameters of the fundamental matrix. Second, pixels in the right image are considered for matching only if they are also relevant points. Otherwise, they are discarded, and correlations are not computed for that pixel. Computation time is abruptly decreased while maintaining similar detection results. Among the wide spectrum of matching techniques that can be used to solve the correspondence problem, we implemented the *Zero Mean Normalized Cross Correlation* [27] because of its robustness. The Normalized Cross Correlation between two image windows can be computed as follows:

$$\text{ZMNCC}(p,p') = \frac{\sum_{i=-n}^{n}\sum_{j=-n}^{n} A \cdot B}{\sqrt{\sum_{i=-n}^{n}\sum_{j=-n}^{n} A^2 \sum_{i=-n}^{n}\sum_{j=-n}^{n} B^2}} \quad (1)$$

where A and B are defined by

$$A = \left(I(x+i, y+j) - \overline{I(x,y)}\right) \quad (2)$$

$$B = \left(I'(x'+i, y'+j) - \overline{I'(x',y')}\right) \quad (3)$$

where $I(x,y)$ is the intensity level of pixel with coordinates (x,y), and $\overline{I(x,y)}$ is the average intensity of a $(2n+1) \times (2n+1)$ window centered around that point. As the window size decreases, the discriminatory power of the area-based criterion is decreased, and some local maxima appear in the searching regions. An increase in the window size causes the performance to degrade due to occlusion regions and smoothing of disparity values across boundaries. According to the previous statements, a filtering criterion is needed in order to provide outlier rejection. First, a selection of 3-D points within the pedestrian searching area is carried out. Second, road surface points as well as high points (points with a Y coordinate above 2 m) are removed. Finally, an XZ map (bird's eye view of the 3-D scene) is filtered following a neighborhood criterion. As depicted in Fig. 1, the appearance of pedestrians in 3-D space is represented by a uniformly distributed set of points.

B. Subtractive Clustering

Data clustering techniques are related to the partitioning of a data set into several groups in such a way that the similarity within a group is larger than that among groups. Normally, the number of clusters is known beforehand. This is the case of *K-means*-based algorithms. In this paper, the number of clusters is considered unknown since no *a priori* estimate about the number of pedestrians in scene can be reasonably made. The effects of outliers have to be reduced or completely removed, being necessary to define specific space characteristics in order to group different pedestrians in the scene. For these reasons, a *Subtractive Clustering* method [28] is proposed, which is a well-known approach in the field of *Fuzzy Model Identification Systems*. Clustering is carried out in 3-D space based on a density measure of data points. The idea is to find high-density regions in 3-D space. Objects in the 3-D space are roughly modeled by means of Gaussian functions. It implies that, in principle, each Gaussian distribution represents a single object in 3-D space. Nonetheless, objects that get too close to each other can be modeled by the system as a single one and, thus, represented by a single Gaussian distribution. The complete representation is the addition of all Gaussian distributions found in the 3-D reconstructed scene. Accordingly, the parameters of the Gaussian functions are adapted by the clustering algorithm to best represent the 3-D coordinates of the detected pixels. The 3-D coordinates of all detected pixels are then considered as candidate cluster centers. Thus, each point p_i with coordinates

(x_i, y_i, z_i) is potentially a cluster center whose 3-D spatial distribution D_i is given by the following equation:

$$D_i = \sum_{j=1}^{N} \exp\left(-\frac{(x_i - x_j)^2}{\left(\frac{r_{ax}}{2}\right)^2} - \frac{(y_i - y_j)^2}{\left(\frac{r_{ay}}{2}\right)^2} - \frac{(z_i - z_j)^2}{\left(\frac{r_{az}}{2}\right)^2} \right) \tag{4}$$

where N represents the number of 3-D points contained in a neighborhood defined by radii r_{ax}, r_{ay}, and r_{az}. Cluster shape can then be tuned by properly selecting the parameters r_{ax}, r_{ay}, and r_{az}. As can be observed, candidates p_i surrounded by a large number of points within the defined neighborhood will exhibit a high value of D_i. Points located at a distance well above the radius defined by $(r_{ax}, r_{ay} \cdot r_{az})$ will have almost no influence over the value of D_i. Equation (4) is computed for all 3-D points measured by the stereovision algorithm. Let $p_{cl} = (x_{cl}, y_{cl}, z_{cl})$ represent the point exhibiting the maximum density denoted by D_{cl}. This point is selected as the cluster center at the current iteration of the algorithm. The densities of all points D_i are corrected based on p_{cl} and D_{cl}. For this purpose, the subtraction represented as

$$D_i = D_i - D_{cl} \exp\left(-\frac{(x_i - z_j)^2}{\left(\frac{r_{bx}}{2}\right)^2} - \frac{(y_i - y_j)^2}{\left(\frac{r_{by}}{2}\right)^2} - \frac{(z_i - z_j)^2}{\left(\frac{r_{bz}}{2}\right)^2} \right) \tag{5}$$

is computed for all points, where the parameters (r_{bx}, r_{by}, r_{bz}) define the neighborhood where the correction of point densities will have the largest influence. Normally, the parameters (r_{bx}, r_{by}, r_{bz}) are larger than (r_{ax}, r_{ay}, r_{az}) in order to prevent closely spaced cluster centers. Typically, $r_{bx} = 1.5\ r_{ax}$, $r_{by} = 1.5\ r_{ay}$, and $r_{bz} = 1.5\ r_{az}$. In this paper, these parameters have been set to $r_{ax} = r_{az} = 1$ m, $r_{ay} = 1.5$ m, $r_{bx} = r_{bz} = 1.5$ m, and $r_{by} = 2.25$ m. After the subtraction process, the density corresponding to the cluster center p_{cl} gets strongly decreased. Similarly, densities corresponding to points in the neighborhood of p_{cl} also get decreased by an amount that is a function of the distance to p_{cl}. All these points are associated with the first cluster computed by the algorithm, which is represented by its center p_{cl}, and will have almost no effect in the next step of the subtractive clustering. After the correction of densities, a new cluster center $p_{cl,\mathrm{new}}$ is selected, which corresponds to the new density maximum $D_{cl,\mathrm{new}}$, and the process is repeated whenever the condition expressed as

$$\text{if } U_{\mathrm{rel}} > \frac{D_{cl}}{D_{cl,\mathrm{new}}} \quad D_{cl,\mathrm{new}} > U_{\min} \quad \Rightarrow \quad \text{new cluster} \tag{6}$$

is met, where U_{rel} and $U_{\min}$ are experimentally tuned parameters that permit the establishment of a termination condition based on the relation between the previous cluster density and the new one, as well as a minimum value of the density function. In this paper, this parameter has been set to $U_{\min} = 40$. The process is repeated until the termination condition given by (6) is not met. After applying subtractive clustering to a set of input data, each cluster finally represents a candidate. The algorithm can be summarized as follows.

1) The parameters (r_{ax}, r_{ay}, r_{az}) and (r_{bx}, r_{by}, r_{bz}) are initialized.
2) The densities of all points are computed using (4).
3) The point p_{cl} that exhibits the highest density value D_{cl} is selected as a cluster center.
4) Densities are corrected according to (5).
5) A new maximum density $D_{cl,\mathrm{new}}$ is computed.
6) If the condition given by (6) is met, a new cluster is considered, which is represented by its center $p_{cl,\mathrm{new}}$, and the algorithm is resumed from Point 4. Otherwise, the algorithm is stopped.

Pedestrian candidates are then considered as the 2-D region of interest (ROI) defined by the projection in the image plane of the 3-D candidate regions. The number of candidates is bound to change depending on traffic conditions, since some cars can be considered as candidates by the subtractive clustering algorithm.

C. Multicandidate (MC) Generation

In practice, a multiple candidate selection strategy has been implemented. The purpose is to produce several candidates around each selected cluster in an attempt to compensate for the effect of the candidate bounding box accuracy in the recognition step. Accordingly, several candidates are generated for each candidate cluster by slightly shifting the original candidate bounding box in the u and v axes in the image plane. The candidate selection method yields generic obstacles with a 3-D shape that is similar to that of pedestrians. The 2-D candidates are then produced by projecting the 3-D points over the left image and computing their bounding box. Two bounding box limits are defined, i.e., for the maximum and minimum values of width and height, respectively, taking into account people taller than 2 m or shorter than 1 m. The 3-D candidate position is given by the stereo-based candidate selection approach (subtractive clustering), which provides the 3-D cluster center coordinates. Nonetheless, the 2-D bounding box corresponding to a 3-D candidate might not perfectly match the candidate appearance in the image plane due to several effects: body parts that are partially occluded or camouflaged with the background, 3-D objects that have been subtracted together with a pedestrian (for example, pedestrians beside traffic signals, trees, cars, etc.), low contrast pedestrians represented by a low number of 3-D points, etc. These badly bounded pedestrians will be classified as nonpedestrians if the positive samples used to train the classifier are well fitted. Let us note that this problem also appears with 2-D candidate selection mechanisms [1] with the additional drawback of losing the actual pedestrian depth.

Two strategies are proposed to solve the "bounding accuracy effect." The first one consists of training the classifier with additional badly fitted pedestrians in an attempt to absorb either the extra information due to large bounding boxes containing part of the background or the loss of information due to small bounding boxes in which part of the pedestrian is not visible. In other words, the positive samples yielded by the candidate selection method are included in the training set. For that purpose,

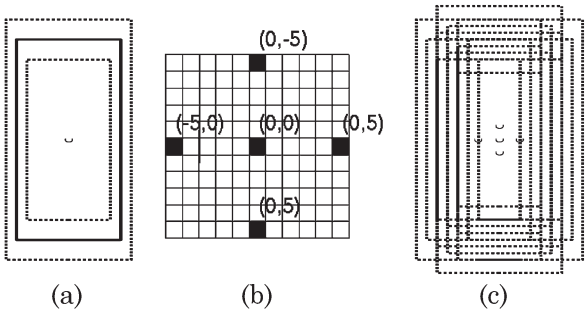


Fig. 2. MC generation approach. (a) Oversized and downsized windows. (b) Spatial centers for each window. (c) Fifteen candidates are generated.

it is necessary to execute the candidate selection process with offline validation to distinguish pedestrians from nonpedestrians. In [1] and [10], the same procedure is only applied to nonpedestrian samples. The second strategy consists of performing an MC generation for every extracted candidate, trying to hit the target and add redundancy. Three window sizes are defined: 1) the window size generated by the candidate selection method; 2) a 20% oversized window; and 3) a -20% downsized one. These three windows are shifted five pixels in each direction: top, down, left, and right. Thus, a total of 15 MCs are generated for each original candidate, as depicted in Fig. 2.

A majority criterion is followed in order to validate a pedestrian. Thus, the MC strategy yields a pedestrian if more than five candidates are as pedestrians. This number has been defined after extensive experiments. In average, the candidate selection mechanism generates six windows per frame, which yields a total of 90 candidates per frame after the MC process. In case the number of candidates generated by the attention mechanism increases abruptly, the MC approach might become impractical. A major benefit derived from the MC approach is the fact that the classification performance of pedestrians at long distance increases. Fig. 3 depicts typical images from our test sequences. The number below the bounding box represents range. The rightmost image shows a motorcyclist that is detected as a pedestrian (false positive). In the leftmost image, two kids are properly detected, and their range is correctly measured.

III. Feature Extraction

The optimal selection of discriminant features is an issue of the greatest importance in a pedestrian detection system considering the large variability problem that has to be solved in real scenarios. A set of features must be extracted and fed to a pedestrian recognition system.

A. Component-Based Approach

There are some important aspects that need to be addressed when constructing a classifier, such as the global classification structure and the use of single or multiple cascaded classifiers. These issues are strongly connected to the way features are extracted. The first decision to make implies the development of a holistic classifier against a component-based approach. In the first option, features are extracted from the complete candidate described by a bounding box in the image plane. The component-based approach suggests the division of the candidate body into several parts over which features are computed. Each pedestrian body part is then independently learned by a specialized classifier in the first learning stage. The outputs provided by individual classifiers, which correspond to individual body parts, can be integrated in a second stage that provides the final classification output. In Section IV, two possible methods for developing a second-stage classifier are described. As long as a sufficient number of body parts or limbs are visible in the image, the component-based approach can still manage to provide correct classification results. This allows for the detection of partially occluded pedestrians whenever the contributions of the pedestrian visible parts are reliable enough to compensate for the missing ones.

After extensive trials, we propose a total of six different subregions for each candidate ROI, which has been rescaled to a size of 24×72 pixels. This solution constitutes a tradeoff between exhaustive subregion decomposition and the holistic approach. The optimal location of the six subregions, which are empirically achieved after hundreds of trials, has been chosen in an attempt to detect coherent pedestrian features, as depicted in Fig. 4. Thus, the first subregion is located in the zone where the head would be. The arms and legs are covered by the second, third, fourth, and fifth regions, respectively. An additional region is defined between the legs, which covers an area that provides relevant information about the pedestrian pose. This subregion is particularly useful to recognize stationary pedestrians.

B. Combination of Feature Extraction Methods

The choice of the most appropriate features for pedestrian characterization remains a challenging problem nowadays since recognition performance depends crucially on the features that are used to represent pedestrians. In the first intuitive approach, some features seem to be more suitable than others for representing certain parts of human body. Thus, legs and arms are long elements that tend to produce straight lines in the image, while the torso and head are completely different parts, which are not so easy to recognize. This statement, although based on intuition, suggests the combination of several feature extraction methods for the different subregions into which a candidate is divided. Accordingly, we have tested a set of seven different feature extraction methods. The selection of features was made based on intuition, previous work carried out by other authors, and our own previous work on other applications. The proposed features are briefly described in the following lines.

- Canny image: The Canny edge detector [29] computes image gradient, i.e., highlighting regions with high spatial derivatives. The computations of edges significantly reduce the amount of data that needs to be managed and filter out useless information while preserving shape properties in the image. The result obtained after applying a Canny filter to the ROI is directly applied to the input of the classifier. The Canny-based feature vector is the same size as the candidate image, i.e., 24×72.
- Haar wavelets, which were originally proposed for pedestrian recognition in [9]: In this paper, only the vertical features have been considered. This yields a feature

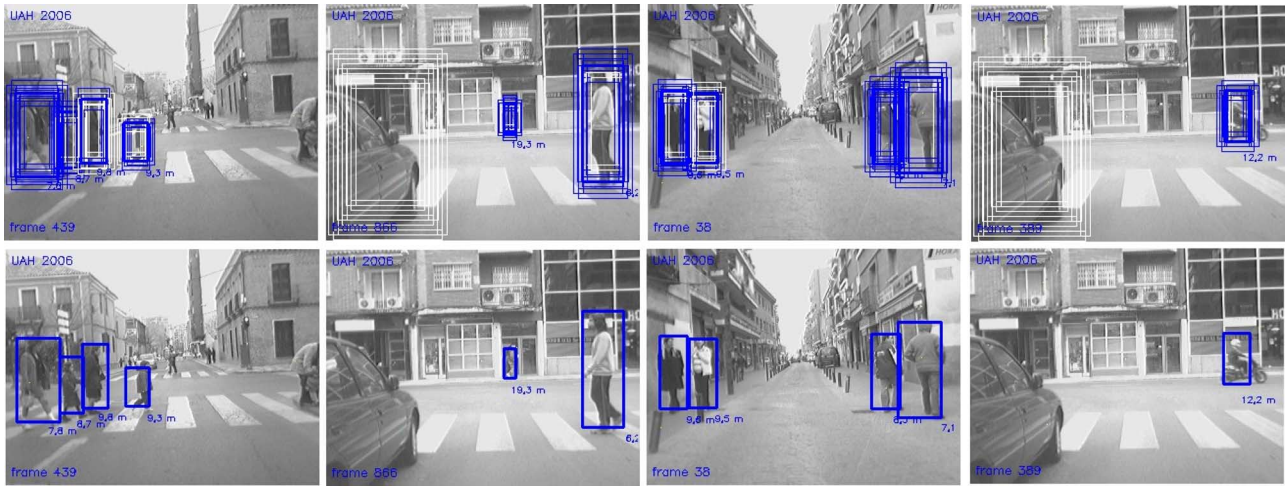


Fig. 3. (Upper row) MC generation. (Bottom row) Results after classifying the 15 candidates.

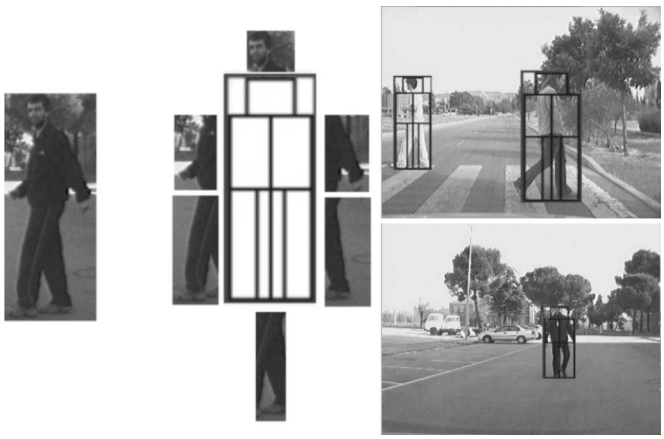


Fig. 4. Decomposition of a candidate ROI into six subregions.

vector of 432 elements, i.e., the candidate size (24×72) divided by 4.

- Gradient magnitude and orientation: The magnitude of the spatial derivatives g_x and g_y are computed for all pixels in the image plane. After that, orientation is calculated as $\theta = \arctan(g_x, g_y)$. The resulting feature vector has twice the size of the candidate image, i.e., the vector has $2 \times 24 \times 72$ elements.
- Cooccurrence matrix [30]: Cooccurrence is specified as a matrix of relative frequencies $P_{i,j}$ with which two neighboring pixels, which are separated by distance d at orientation θ, cooccur in the image: one with gray level i and the other with gray level j. The Cooccurrence matrix can be computed over the gray-level image or over the Canny image. The resulting matrices are symmetric and can be normalized by dividing each entry in a matrix by the number of neighboring pixels used in the matrix computation. In our approach, we propose a distance of one pixel and four different cooccurrence matrices for the following orientations (bins): (0°, 45°, 90°, 135°). The resulting size of the feature vector depends on whether the cooccurrence matrix is computed over the original gray-level image or over a binary one (after applying the Canny operator). The cooccurrence matrix over the Canny image yields a feature vector of $4 \times 2 \times 2$ elements.
- Histogram of intensity differences: The relative frequencies of intensity differences are computed between neighboring pixels along four orientations over a normalized image of 128 gray levels. This generates a features vector of 4×128 elements.
- Histogram of normalized gradients [31] (HON): The Gradient image is considered. Orientation is discretized to 20 bins (corresponding to an accuracy of 18°). Only pixels in the Gradient image exhibiting a magnitude greater than some threshold (10) are considered. For those pixels, the values of gradient are accumulated in a 20-bin histogram. Thus, the resulting features vector has 20 elements.
- Number of texture unit (NTU) [32]: The local texture information for a pixel can be extracted from a neighborhood of 3×3 pixels, which represents the smallest complete unit of texture. The corresponding texture unit is computed by comparing the pixel under study with its eight neighboring pixels. The NTU process generates a feature vector with the same size as the candidate image, i.e., a feature vector of 24×72 elements.

An appropriate selection of discriminant features is then carried out in order to determine the best performing features for pedestrian detection. In the first approach, performance comparison was made by following the next steps. First, each feature extractor is applied over the six candidate subregions. This yields a set of six feature vectors for each candidate. Then, the obtained feature vectors are applied to the input of a classifier system that provides a single output, which represents whether the candidate is classified as pedestrian or nonpedestrian. Performance comparison can then be easily done by analyzing the DR and False-Positive Rate (FPR) achieved by the classifier for the seven different feature vectors under test, as described in Section VI.

In theory, the best performing feature extractor method should be selected in order to implement the final detection system. However, a detailed observation of partial results reveals that some feature extraction methods prove to be more discriminant than others for certain subregions, as depicted in

Fig. 6. Thus, NTU and Histogram perform the best for head and arms, while HON, Canny, and Histogram seem to perform the best for legs. Similarly, the area between-the-legs is best recognized by NTU. There seems then to be an optimal feature extraction method for each candidate subregion. Thus, each candidate subregion will be learned separately by an independent classifier. The input to the classifier associated to a given subregion will be the features vector corresponding to the best performing method for such a subregion. The fine-grain selection of optimal feature extraction methods has been carried out, as described next. First, the three best performing methods have been selected for each subregion. Then, the performance difference among the three selected feature extraction methods has been evaluated. If there is a method that clearly outperforms the rest of the methods, it is selected as the optimal method for the subregion under consideration. Otherwise, a decision is made considering other aspects such as the feature vector size. In such a case, the two feature extraction methods yielding a smaller vector size are chosen among the three best performing ones. According to these parameters, a preselection of features is made, with the following result: head—NTU; arms—NTU and Histogram; legs—HON and Canny; between-the-legs—NTU. An iterative process is started to test the four possible combinations using the previously mentioned preselected feature extraction methods. The comparison among the results achieved in the four experiments yields the final combination of features used in this paper: head—NTU; arms—Histogram; legs—HON; between-the-legs—NTU. The increase in performance due to the use of the proposed optimal combination of feature extraction methods is illustrated in Section VI. The optimal combination of feature extraction methods eases the learning stage, which makes the classifier less sensitive, in particular, to clothing.

IV. Pedestrian Detection Using SVM

Pedestrian detection is done using SVMs. Two aspects are essential in the deployment of SVM classifiers: the training strategy and the classifier structure.

A. Training Strategy

The first step in the design of the training strategy is to create representative databases for learning and testing. The training and test sets were manually constructed using the TSetBuilder tool [33] developed in our lab. The following considerations must be taken into account when creating the training and test sets.

- The ratio between positive and negative samples has to be set to an appropriate value. A very large number of positive samples in the training set may lead to a high percentage of false-positive detections during online classification. On the contrary, a very large number of negative samples produce mislearning. A tradeoff of one positive sample for every two negative samples was initially chosen in our application and compared to the 1/1 option, as described in Section VI.
- The size of the database is a crucial factor to take care of. As long as the training data represent the problem well, the larger the size of the training set, the better it is for generalization purposes. Nonetheless, the value of the regularization coefficient C [34] is important since this parameter controls the degree of overlearning. Thus, a small value of C allows a large separation margin between classes, which reduces overlearning and improves generalization. In this paper, a value of $C = 1.0$ has been used after extensive trials. This value can be considered as a small one. The dimension of the database has been designed in order to achieve real generalization, as demonstrated in practical experiments.
- The quality of negative samples has a strong effect in the DR. Negative samples have to be properly selected to account for ambiguous objects, such as poles, trees, advertisements, and the like. Only by following this strategy when creating the training sets can a really powerful classifier be achieved in practice.
- A sufficiently representative test set must be created for verification. The content of the test set has similar characteristics to those of the training sets in terms of variability, ratio of positive/negative samples, and quality of negative samples.

A detailed observation of the classifier operation in practice suggests the subdivision of the classification task into several more tractable learning sets according to different practical considerations. A major issue is the effect of illumination conditions. It is clear that daytime and nighttime samples must be compulsorily separated in order to create multiple specialized classifiers. The nighttime classifier can be reasonably expected to operate correctly only in very short distances (below 6–8 m) for nonilluminated areas, where pedestrians can be appropriately illuminated by the car beams (infrared images would be needed in order to achieve long-range detection, as mentioned in Section I). Nonetheless, nighttime pedestrian detection can be done up to 15–20 m in illuminated areas. The separation between day and night specialized classifiers may not be enough to cover the most significant cases of pedestrian variability. In fact, as observed in practice, the effect of depth is determinant. Shapes and edges are not so neatly distinguished when pedestrians are beyond 12–15 m from the cameras. Accordingly, the effect of depth suggests the development of specialized SVM classifiers at daytime. Albeit several subdivisions could be done for very short, short, medium, long, and very long range, two specialized classifiers for short- and long-range detections have been considered to be enough in practice. The threshold between short and long range has been empirically set to 12 m.

The effect of pose must also be taken into account as a significant source of variability in pedestrian appearance. Differences between walking and stationary pedestrians are clear. There are even some remarkable differences between pedestrians moving laterally, with regard to vehicle trajectory, and those moving longitudinally. Pedestrians intersecting the vehicle trajectory from the sides are usually easier to recognize since their legs are clearly visible and distinguishable. In fact, some authors have proposed two separate SVM classifiers according to this

statement [20]. A more complicated case occurs when a pedestrian crouches or bends down. Changes due to different clothing also contribute to further complexity in the variability problem. Thus, large skirts and coats make pedestrians look very different from those in trousers and suits. Likewise, pedestrians bringing trolleys or bags make the recognition problem even more difficult. Had it not been enough, the pedestrians' legs are not always visible in the image, especially when pedestrians are very close to the vehicle. This is a critical case of great importance for precrash protection systems.

In order to handle all these variability cases, we have created separate training sets intended to perform pedestrian learning in short and long range at daytime and nighttime, respectively. Four training sets were built for this purpose, which contain a number of negative samples that double the number of positive ones: a training set of 9000 daytime long-range samples (denoted by DL), a training set with 15 000 daytime short-range samples (denoted by DS), a third training set containing 6000 nighttime samples (denoted by N), and a global training set containing the concatenation of all samples in DL and DS (denoted by G, 24 000 samples). Similarly, four test sets were created and denoted by test set for daytime short range (TDS, 5505 samples), test set for daytime long range (TDL, 4320 samples), test set for nighttime (TN, 3225 samples), and global test set composed by the concatenation of TDS and TDL (TG, 9825 samples), respectively. In order to test the effect of the positive/negative ratio in the training process, the original training sets were modified to contain the same number of positive and negative samples. Accordingly, the modified sets have a size that is two thirds the size of the original sets, as long as half of the negative samples were removed while the positive ones remained untouched. Variability due to pose, clothing, and other artifacts is handled by creating adequate training databases containing as many representative cases as possible. In this stage, pedestrians in different poses (standing, walking, ducked, etc.) and clothing (coats, skirts, etc.) are included in the database as well as pedestrians with handbags and other artifacts. In total, the training sets contain up to 30 000 samples, while the test sets amount up to 13 050 samples.

B. Classifier Structure

A two-stage classifier is proposed in order to cope with the components-based approach. In the first stage of the classifier, features computed over each individual fixed subregion are fed to the input of individual SVM classifiers. Thus, there are six individual SVM classifiers corresponding to the six candidate subregions. These individual classifiers are specialized in recognizing separate body parts corresponding to the prespecified candidate subregions. It must be clearly stated that no matching of parts is carried out. Instead, each individual SVM is fed with features computed over its corresponding candidate subregion and provides an output that indicates whether the analyzed subregion corresponds to a pedestrian part ($+1$, in theory) or not (-1, in theory). In the second stage of the classifier, the outputs provided by the six individual SVMs are combined. Two different methods have been tested to carry out this operation. The first method implements what we denote as simple-distance criterion. A simple addition is computed as

$$S_{\text{distance-based}} = \sum_{i=1}^{6} S_i \tag{7}$$

where S_i represents the real output of the SVM classifier (not strictly contained in the ideal range $[-1, +1]$) that corresponds to subregion i. In theory, subregions corresponding to nonpedestrians or missing parts should contribute with negative values to $S_{\text{distance-based}}$. Likewise, subregions corresponding to pedestrian parts should contribute with positive values to the final sum. A threshold value T is then established in order to perform candidate classification. This threshold is parameterized for producing the *Receiver Operating Characteristic (ROC)*. The difference between pedestrians and nonpedestrians is set depending on the distance between T and $S_{\text{distance-based}}$. Thus, if $S_{\text{distance-based}}$ is greater than T, the candidate is considered to be pedestrian. Otherwise, it is regarded as nonpedestrian. This simple mechanism is what we denote as distance-based criterion.

The second method that has been tested to implement the second stage of the classifier relies on the use of another SVM classifier. A second-stage SVM merges the outputs of the six individual first-stage SVM classifiers and provides a single output representing the candidate classification result. The resulting global structure is denoted as two-stage SVM classifier. Obviously, the second-stage SVM classifier has to be trained with supervised data. The training set for the second-stage SVM classifier has been built as follows. First, the six individual first-stage SVM classifiers are properly trained using training set DS (which contains 15 000 samples) in which the desired outputs (pedestrian or nonpedestrian) are set in a supervised way. Then, a new training set is created by taking as inputs the outputs produced by the six already trained first-stage SVM classifiers (in theory, between -1 and $+1$) after applying the 15 000 samples contained in DS and taking as outputs the supervised outputs of DS. The test set for the second-stage SVM classifier is created in a similar way using test set TDS (containing 5505 samples).

Additionally, an optimal kernel selection for the SVM classifiers has been performed. For this purpose, we used a small training set of 2000 samples for which the well-known Gaussian (Radial Basis Function), sigmoid, polynomial, and linear kernels [34] were tested. The Gaussian kernel was finally chosen as the optimal one after the trials.

V. Multiframe Validation and Tracking

Once candidates are validated by the SVM classifier, a tracking stage takes place. Pedestrian tracking is needed to filter detection results and minimize the effect of both false-positive and false-negative detections. For this purpose, detection results are temporally accumulated. The multiframe validation and tracking algorithm relies on Kalman filter theory to provide spatial estimates of detected pedestrians and Bayesian probability to provide an estimate of pedestrian detection certainty over

time. Spatial estimates of the detected pedestrians are given by a linear Kalman filter. Tracking is done in 3-D space. The state vector is composed of five elements, which contain the 3-D pedestrian position (X, Y, Z) (indeed, the position of the mass center) and the pedestrian width W and height H. Thus, $x = (X, Y, Z, W, H)^T$. The 3-D relative velocity between the car and the target pedestrian $v_R = (v_{Rx}, v_{Ry}, v_{Rz})$ is considered in the state transition matrix A together with the sampling rate of the complete algorithm Δ_t for predicting x_k^-. Relative velocity is computed as a function of the 3-D relative distance $(\Delta_x, \Delta_y, \Delta_z)$ between the ego-vehicle and the target pedestrian in two consecutive frames. Each newly detected pedestrian is tracked by an individual Kalman filter. Multiframe validation is needed to endow the tracking system with robustness. The use of Bayesian probability is proposed to provide estimates of pedestrian detection certainty over time. In other words, a sort of low-pass filter has been designed based on Bayesian probability. The process is divided in two stages: pretracking and tracking. Newly detected pedestrians enter the pretracking stage. Only after consolidation in the pretracking stage do they start to be tracked by the system. The process followed in the pretracking stage after detecting a pedestrian candidate is described next.

1) The 3-D position of the newly detected pedestrian is compared to the 3-D position of all pedestrians that are being tracked by the system at time k. The system maintains a list of tracked pedestrians. The candidate pedestrian is validated using Probabilistic Data Association. The idea is to provide matching between newly detected candidates and already existing pedestrians under tracking. For that purpose, the detected pedestrian is associated with the closest already existing pedestrian following the Euclidean distance criterion. Association with the closest pedestrian is done whenever the condition established as

$$pda\left(s_{i,k}^-, m_{j,k}\right) = e^{-\frac{\left(s_{i,k}^- m_{j,k}\right)^2}{2\sigma^2}} > 0.7 \tag{8}$$

is met, where $s_{i,k}^-$ represents the 3-D predicted position of the closest pedestrian i (the first three elements of vector $x_{i,k}^-$), $m_{j,k}$ is the 3-D position of the measured candidate j (the first three elements of vector $z_{j,k}$), and σ^2 is the covariance of the Gaussian distribution representing the predicted position of the target pedestrian (the following assumption has been made: $\sigma_x = \sigma_y = \sigma_z = \sigma$). Only candidates meeting (8) are validated by the system and enter the pretracking stage. Otherwise, the candidate is considered to be a new pedestrian appearing in the scene.

2) If the candidate is considered to be a new pedestrian, it is annotated in the tracked pedestrian list as a new element denoted by j, and its probability of being a pedestrian is initialized according to the classification value given by the SVM classifier at frame k ($S_{\text{distance-based},j,k}$), i.e.,

$$P(j_k) = \begin{cases} 1.0, & \text{if } 0.5 + D_{j,k} > 1.0 \\ 0.0, & \text{if } 0.5 + D_{j,k} < 0.0 \\ 0.5 + D_{j,k}, & \text{otherwise} \end{cases} \tag{9}$$

where $D_{j,k} = S_{\text{distance-based},j,k} - T$. The value of $P(j_k)$ is saturated to be limited between 0.0 and 1.0. After that, the position of the new pedestrian is initialized as $x_{j,k} = z_{j,k}$, and pretracking is activated.

A pedestrian entering the pretracking stage must be validated in several iterations before entering the tracking stage. The algorithm followed to implement pedestrian validation during pretracking is described in the following.

1) Let $s_{j,k}^-$ represent the predicted position of the prevalidated pedestrian j at frame k, and let $m_{j,k}$ represent the associated measure at frame k after performing Probabilistic Data Association. The probability of precandidate j to be considered as pedestrian at frame k, denoted by $P(j_k)$, is given by

$$\begin{aligned} P(j_k) &= P(j_k/j_{k-1})P(j_{k-1}) \\ &= C_n \cdot f(D_{j,k}) pda(s_{j,k}^-, m_{j,k}) P(j_{k-1}) \end{aligned} \tag{10}$$

where C_n is a normalizing factor.

2) The precandidate is validated as a pedestrian when its probability is above 0.8 during three consecutive iterations. Once a precandidate is validated, pretracking stops, and tracking starts.
3) Pretracking is stopped if the precandidate probability is below 0.5 during three consecutive iterations.

The same condition applies during tracking, i.e., tracking a pedestrian stops if its probability is below 0.5 during three consecutive iterations. The implementation of the multiframe validation and tracking algorithm described in this section permits the achievement of a compromise between robustness in new pedestrian detections and accuracy in pedestrian tracking.

VI. Experimental Results

The system was implemented on a Pentium IV PC at 2.4 GHz running the Knoppix GNU/Linux Operating System and Libsvm libraries [35]. Using 320 × 240 pixel images, the complete algorithm runs at an average rate of 20 frames/s, depending on the number of pedestrians being tracked and their position. The average rate has a strong dependency on the number of pixels being matched because of the correlation of computational cost, which consumes, on average, 80% of the whole processing time. The candidate selection system has proved to be robust in various illumination conditions, different scenes, and distances up to 25 m. The quality of the classification system is mainly measured by means of the DR/*false positive* ratio (DR/FPR). These two indicators are graphically bounded together in an *ROC*.

We created several training and test sets containing thousands of positive and negative samples (pedestrians and nonpedestrians, respectively) in different situations, as described in Section IV-A. In order to evaluate the influence of the positive/negative ratio in the training process, two different types of training sets were created. In the first type, the number of nonpedestrian samples was chosen to be twice the number of

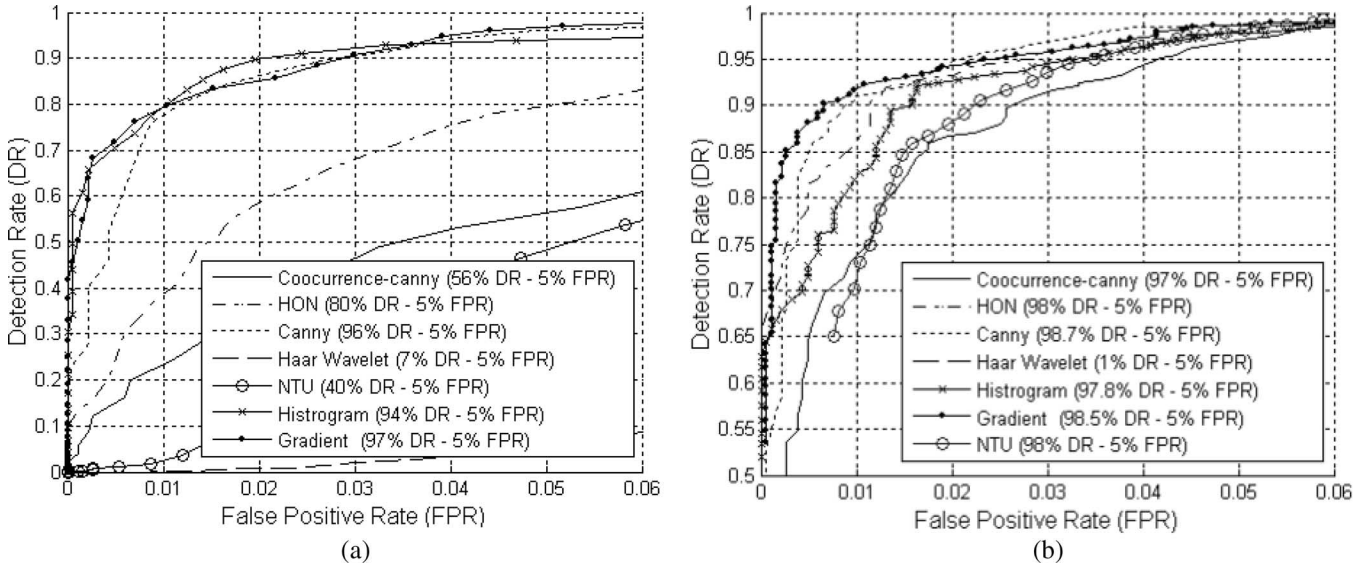


Fig. 5. *ROC* curves. (a) Holistic approach. (b) Components-based approach.

pedestrian samples. In the second type, the number of nonpedestrian and pedestrian samples was chosen to be the same. Positive samples (pedestrian samples) were extracted from recorded images acquired in real traffic conditions. Training sets were created both at daytime and nighttime using the TSetBuilder tool, which is specifically developed in this project for this purpose. By using the TSetBuilder tool, different candidate regions were manually selected from the image on a frame-by-frame basis. As previously mentioned, special attention was given to the selection of nonpedestrian samples. By selecting simple nonpedestrian samples (for instance, road regions), the system learns quickly but does not develop enough discriminating capability in practice as the attention mechanism may wrongly select a region of the image that might be very similar to a pedestrian. Accordingly, negative samples (nonpedestrian samples) in the training sets were neither randomly nor manually selected. The candidate selection mechanism described in Section II was used instead to automatically produce the negative training samples. The use of this mechanism endows the process with a strong realistic component. In the following sections, the results are compared and assessed using DR under certain FPRs.

The selection of the FPR value has been made to show performance in representative points where differences between curves can be optimally appreciated. FPR must be a value for which DR exhibits an acceptable value. This leads to selecting 5% in some cases or 10% in others. For cases in which 10% has been chosen, a value of 5% would not make sense since DR would be a really poor value in those conditions. In addition, FPR has been chosen as a value from which practically no cross points occur among the *ROC* curves of the different features. This means that a curve that is better than another at a given FPR_i remains better for almost all FPR values greater than the given FPR_i, as can be observed in the figures provided in this section. Accordingly, different FPR values have been selected for different types of tests in order to provide meaningful comparisons.

A. Holistic versus Component-Based

A first comparison is made in order to state the best performing approach among the holistic and component-based options. For this purpose, both the holistic and component-based classifiers were trained and tested using the same set. In particular, the training and test sets were designed to contain 10 000 and 3670 samples, respectively. These sets were created as subsets of DS and TDS. All samples were acquired in daytime conditions. As depicted in Fig. 5, the performance of the holistic approach for all feature extraction methods is largely improved in the component-based approach. In the component-based approach, the outputs of the six SVMs corresponding to the six candidate subregions are combined in a simple-distance classifier, as explained in Section IV. Almost every feature extraction method produces an acceptable result in the component-based approach, where the DR is between 97% and 98.7% at an FPR of 5% for all feature extraction methods, except for the Haar Wavelet. The DR ranges from 40% to 97% at an FPR of 5% in the holistic classifier. The Haar Wavelet is again below those figures. This shows that breaking the pedestrian into smaller pieces and specifically training the SVM for these pieces reduces the variability and lets the SVM generalize the models much better. It can then be stated, as previously agreed by other researchers, that the component-based approach clearly outperforms the global classifier.

B. Combination of Optimal Features

These results can further be improved by combining different feature extraction methods for different candidate subregions. The best performing features for each subregion are combined in a second classifier instead of applying the same feature extractor to all six subregions. In this paper, we used the same training and test sets as in Section VI-A. Fig. 6(a)–(f) shows the *ROC* curves for each separate subregion after computing the seven predefined features. As concluded in Section III-B,

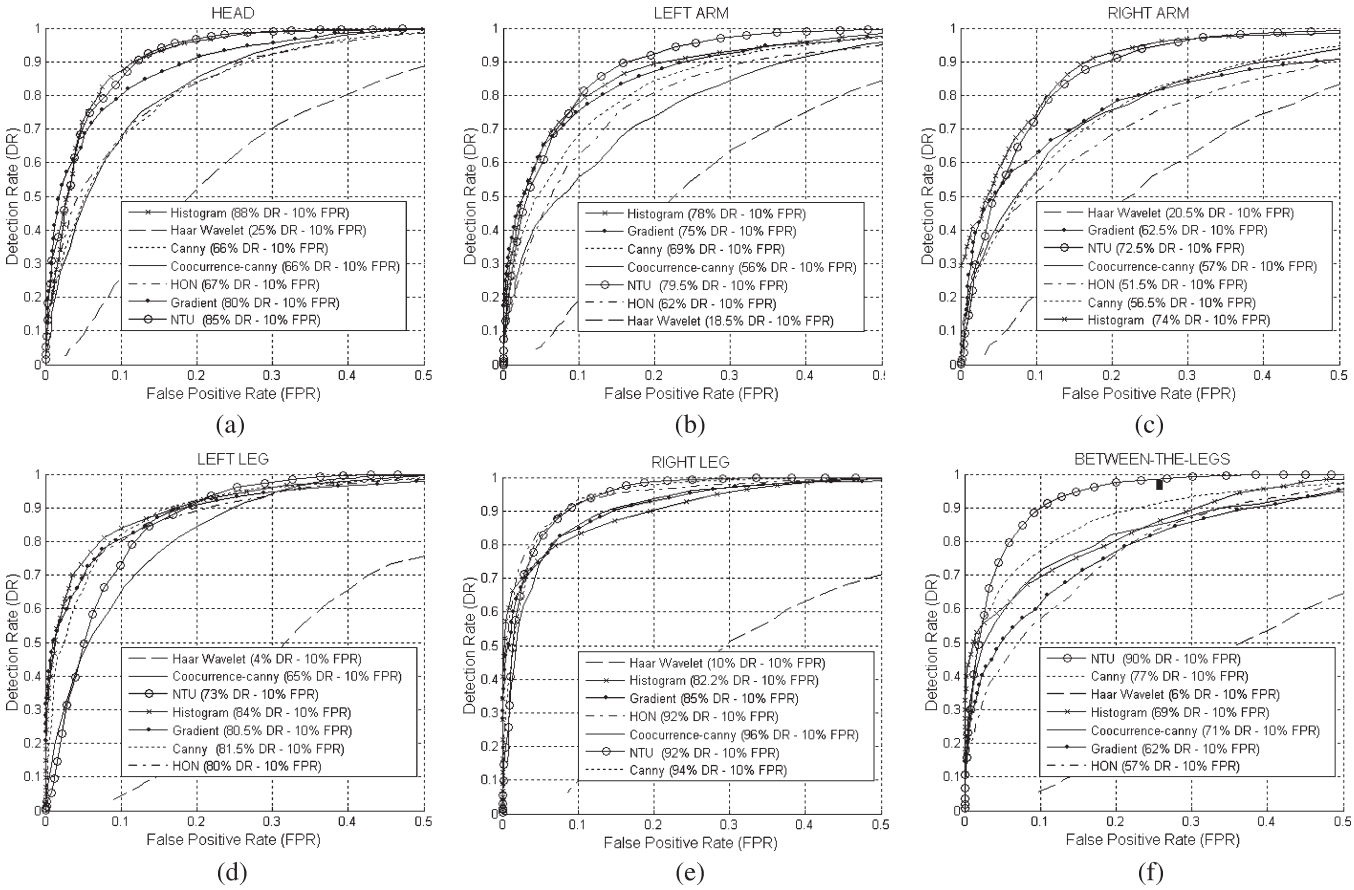


Fig. 6. *ROC* curves. (a) Head. (b) Left arm. (c) Right arm. (d) Left leg. (e) Right leg. (f) Between-the-legs.

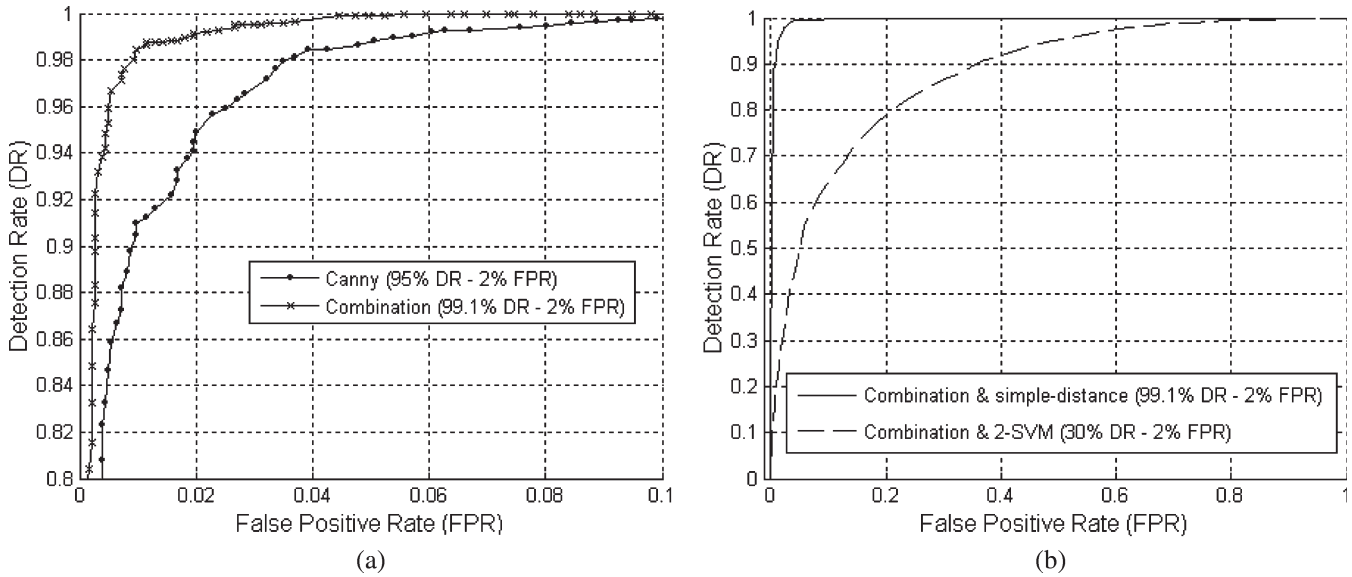


Fig. 7. *ROC* curves. (a) Comparison between features combination and Canny's extractor. (b) Comparison between simple-distance classifier and two-stage SVM.

the selection of optimal features for each subregion is carried out as follows: head—NTU, arms—Histogram, legs—HON, between-the-legs—NTU. The combined use of optimal features leads to a clear increase in the overall classifier performance with regard to individual feature extractors, as depicted in Fig. 7(a), where a DR of 99.1% is achieved for an FPR of 2%. These results improve the performance of Canny's detector, which is the best performing feature extractor (in the conditions of the experiment conducted and described in Section VI-A), which exhibits a DR of 95% at an FPR of 2%.

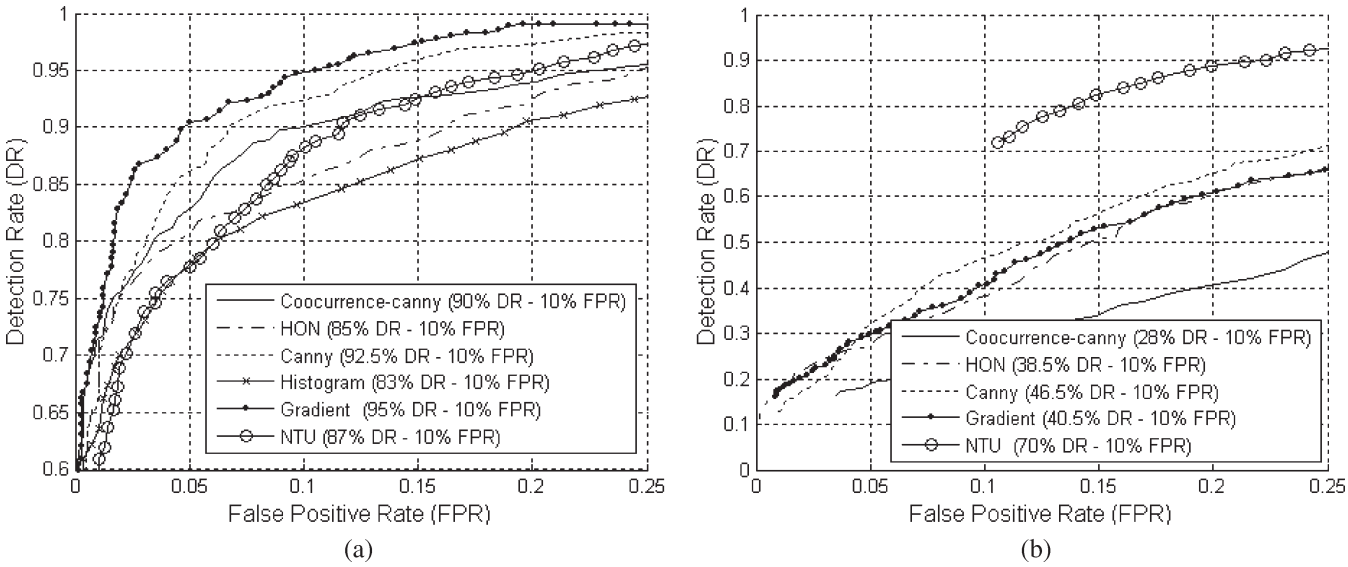


Fig. 8. *ROC* curves for nighttime pedestrian detection. (a) Classification of nighttime test samples using training set N (nighttime samples). (b) Classification of nighttime test samples using training set C (daytime samples).

C. Analysis of the Second-Stage Classifier

Another comparison has been studied in order to analyze the influence of the second-stage classifier that combines the information delivered by the six specifically trained SVM models. In the first approach, we have used a simple-distance criterion (i.e., distance to the hyperplane separating pedestrians from nonpedestrians) that computes the addition of the six first-stage SVM outputs and then decides the classification by setting a threshold. Another option has been tested by training a two-stage SVM (2-SVM). Once again, the same training and test sets as in Section VI-A were used in this experiment. The results achieved to date show that the simple-distance criterion clearly outperforms the 2-SVM classifier, as depicted in Fig. 7(b), where a comparison between both methods is shown when optimal feature extraction methods are applied. Thus, the simple-distance classifier exhibits a DR of 99.1% at FPR $= 2\%$, while the performance of the 2-SVM classifier is DR $= 30\%$ for the same FPR $= 2\%$. As a consequence of this, the combined use of component-based optimal feature extraction methods in a distance-based classifier is proposed as a reliable solution for pedestrian classification.

D. Effect of Illumination Conditions and Candidate Size

The need of separate training sets for day, night, and different candidate sizes is analyzed in this section. All training processes were carried out for both the 1/1 and the 1/2 positive/negative ratio in the training sets. Although the results attained after the experiments do not exhibit a dramatic difference in performance, a slightly superior behavior is obtained by using training sets following the 1/1 positive/negative ratio. Accordingly, the rest of the experiments shown here and in the next section were carried out using training sets with the same number of positive and negative samples. Nighttime samples were acquired only in illuminated urban and nonurban environments, where pedestrian detection remains feasible under the same conditions previously stated throughout this paper, i.e., below 25 m. Nonilluminated areas have not been considered in this analysis since pedestrian detection would not be possible beyond a few meters (6–8 m), and infrared cameras would be needed. As previously stated, the separation between short- and long-distance pedestrian detections has been empirically set to 12 m.

In the first experiment, an SVM classifier was trained using set G (containing all daytime samples) and tested using set TN. Next, a different SVM classifier was trained using set N (nighttime samples) and tested using the same set TN. The purpose of this experiment is to analyze the performance of nighttime classification using a global daytime classifier. The results of this experiment are depicted in Fig. 8. Observation of Fig. 8(a) reveals that nighttime pedestrian detection exhibits a high performance when training is carried out using a database containing nighttime samples. Thus, the DR is between 83% and 95% for all feature extraction methods at an FPR of 10%. Fig. 8(b) shows that nighttime pedestrian detection is not accurate when training is carried out using daytime samples (DR is between 23% and 70% at an FPR of 10%). In such a case, none of the proposed feature extraction methods exhibit acceptable operation as their performance is well below the N-based SVM classifier. In the first approach, our conclusion is that separate training sets for daytime and nighttime are definitely advisable for optimal classification. Illumination conditions are too different between day and night, which makes it difficult to maintain the same training set and the same classifier for all cases, since generalization becomes a really complex problem.

In the next experiment, three different SVM classifiers were trained using sets DS, DL, and G, respectively. The trained classifiers were tested against sets TDS and TDL. The purpose of this experiment is to test the necessity or convenience of training separate classifiers for short and long ranges at daytime. In the first step, all three classifiers were tested using TDS in order to demonstrate the influence of specialized classifiers

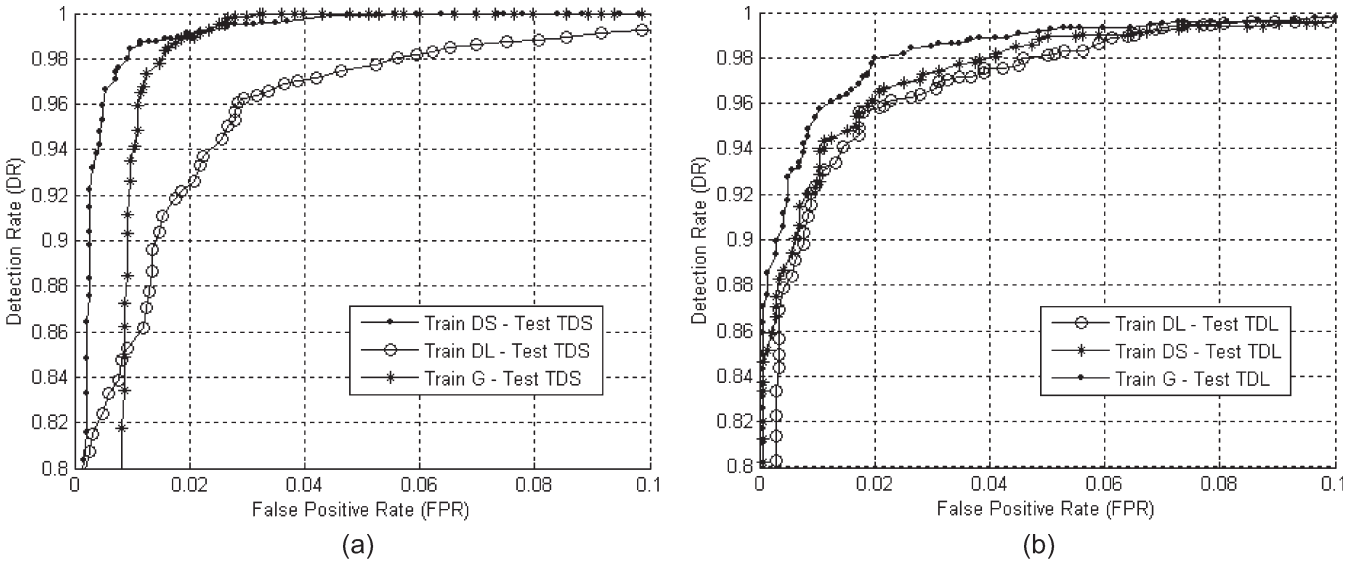


Fig. 9. *ROC* curves for daytime pedestrian detection. (a) Pedestrian detection at short distance. (b) Pedestrian detection at long distance.

in daytime short-range classification. The results are illustrated in Fig. 9(a). In the second step, the test is repeated using this time TDL to check how daytime long-range classification gets affected by learning specialization. Fig. 9(b) shows the results of this experiment. In both cases, the tests are executed using the optimal combination of features described in the previous section. As can be observed in Fig. 9(a), the classifier specialized in short-distance pedestrians exhibits only a bit better performance than the rest. Thus, the DR for a DS-based classifier (SVM classifier trained using set DS) is higher than the DR for a G-based classifier for an FPR below 2%, while the G-based classifier performs better for FPR greater than 2%. Similarly, the results depicted in Fig. 9(b) show that the G-based classifier clearly outperforms the rest of the classifiers for long-distance pedestrian detection. Despite the fact that short-distance pedestrian detection is slightly improved by using separate training sets, our conclusion, contrary to the initial intuition, is that a single SVM classifier trained with a single database containing all types of pedestrians at short and long distances proves to be more effective than separate classifiers for short and long distances, respectively. Let us state clearly that this statement remains applicable only for daytime pedestrian detection.

E. Effect of Bounding Box Accuracy

The accuracy exhibited in bounding candidates is limited, and in fact, a multiple-hypothesis generation for each detected candidate is encouraged to boost classifier performance, as described in Section II-C. Although this topic is usually not considered by most authors, in this section, we analyze the effect of badly bounded candidates in the performance of the recognition system. For this purpose, we devised an experiment in which an SVM classifier was trained using a training set of 3000 well-fitted (or tightly bounded) candidates (i.e., the bounding box of candidates fits the real position of the corresponding pedestrians in the image plane), while a different SVM classifier was trained using a training set containing 2000 badly bounded candidates. Next, the system is evaluated using a test set containing 1000 badly bounded candidates, which is the most usual situation in online real operation. The results of this experiment are illustrated in Fig. 10. Fig. 10(a) depicts the performance obtained after testing a set of badly bounded samples using a classifier trained on badly bounded samples. Practical results show that the performance remains nearly unaffected for HON, Canny, and Cooccurrence–Canny extractors, where a DR of 92%, 81%, and 80%, respectively, is obtained at an FPR of 5%. Quite the opposite, other methods exhibit a clear decrease in performance. Fig. 10(b) shows the performance obtained after testing a set of badly bounded samples using a classifier trained on well-fitted (or tightly bounded) samples. In this case, all methods exhibit much worse figures since none of the proposed extractors succeed in providing a DR above 83% (for the case of HON, which is the best performing one) at an FPR of 5%. The analysis of these results suggests that choosing the optimal feature extraction methods just in terms of DR and FPR can lead, in practice, to a decrease in recognition performance. It seems advisable to carry out a strategy in which badly bounded candidates will be deliberately introduced in the training set. Additionally, an MC generation stage has been developed in order to generate several candidates for each originally selected hypothesis to at least assure some well-fitted candidates that match the samples used for training.

F. Global Performance

The performance of the global system is evaluated in a set of sequences recorded in real traffic conditions. Some of the sequences were acquired in urban environments and others in nonurban areas. The purpose of this evaluation is to assess the combined operation of the attention mechanism and the SVM-based classifier, including the MC generation strategy, and a multiframe validation stage using Kalman filtering. The results obtained in the experiments are listed in Table I. For each row in the table, the following information is provided: type of environment (urban or nonurban; the nonurban sequences

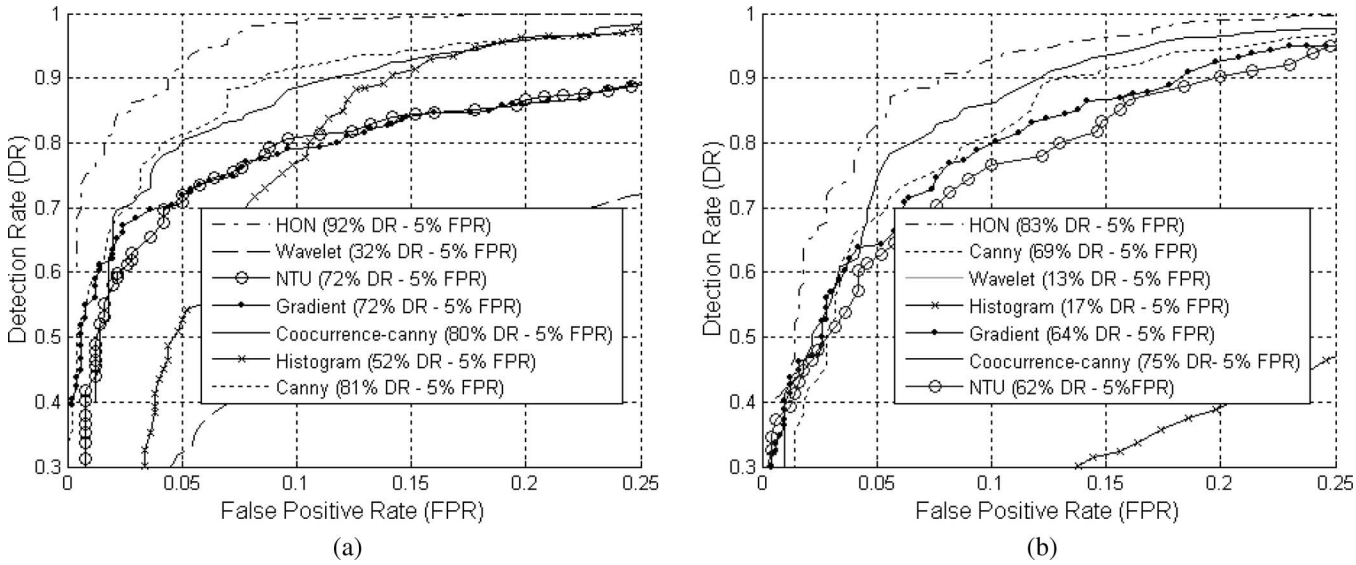


Fig. 10. *ROC* curves for bounding box accuracy. (a) Classification of badly bounded samples using training set containing badly bounded samples. (b) Classification of badly bounded samples using training set containing only tightly bounded samples.

TABLE I
GLOBAL PERFORMANCE EVALUATED IN A SET OF SEQUENCES

Environment	Duration	Detected	Missed	False alarms
Urban	20 min	138	10	11
Non urban	72 min	163	3	5

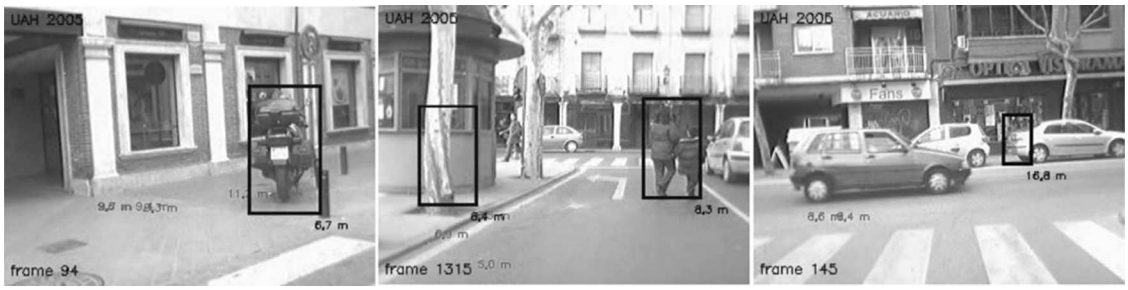


Fig. 11. Examples of false-positive detections.

were recorded in open roads as well as in the campus of the University of Alcalá), duration of the sequence, number of detected pedestrians (only pedestrians below 25 m are considered), number of missed pedestrians, and number of false alarms (F/A) issued by the system. Let us remark that the generation of false alarms is also subject to multiframe validation in order to avoid glitches. Accordingly, a false alarm takes place only when a false positive persistently occurs in time. The global system was implemented according to the following features: subtractive clustering candidate selection, component-based SVM using the six subregions described in Section IV, combination of features according to the description provided in Section III-B, multiple SVM for daytime and nighttime classification, MC generation to compensate for the bounding box accuracy effect, and multiframe validation using Kalman filtering.

The analysis of results reveals that performance is quite different in urban and nonurban environments. Thus, the pedestrian detection system exhibits a ratio of 11 false alarms in 20 min of operation in urban scenarios. This yields a ratio of 33 false alarms per hour. Similarly, the DR is 93.24% in urban environments, where ten pedestrians were missed by the system. Let us clarify the fact that all missed pedestrians were partially occluded or completely out of the vehicle path. Concerning the 11 false alarms produced by the system, they were caused by three motorbikes, two trees, four lampposts and other urban furniture, one wastebasket, and one fence. Fig. 11 depicts three examples of false detections. In all false alarm cases, there was a misclassified real object causing the false alarm. Concerning nonurban environments, three pedestrians were missed by the system in 72 min of operation. In all cases, pedestrians were far from the car (20 m or beyond) and wore clothes that produced almost no contrast with the background. This yields a DR of 98.19% in nonurban scenarios, where images are not so heavily corrupted with clutter. Similarly, five false alarms occurred in the sequences, which are mainly due to

lampposts and trees located by the edge of the road, yielding an average ratio of four false alarms per hour. As happens in urban environments, false alarms are caused by real objects. Although these figures are still unacceptable for the deployment of a real pedestrian detection system, the results described in this paper point to the possible application of a robust pedestrian protection system in roads and other open environments. In any case, the results can be largely improved by incorporating motion- and position-dependent features.

VII. Conclusion and Future Work

In this paper, we have carried out a comparative study of feature extraction methods for vision-based pedestrian detection. Candidates are first selected by an attention mechanism based on subtractive clustering and stereo vision. This helps reduce the number of false candidates and enhance the performance of the recognition stage. In order to reduce the variability of pedestrians, the learning process has been simplified by decomposing selected candidates into six local subregions that are easily learned by individual SVM classifiers. The component-based approach has been demonstrated to outperform the global classifier in practice. In addition, the combination of different feature extraction methods for different subregions leads to an increase in classifier performance. Accordingly, the so-called optimal features have been identified for each subregion and combined in a more discriminant components-based classifier. Likewise, the effects of illumination conditions and candidate size have been studied. Several training and test sets have been created for empirically demonstrating the suitability of multiple classifiers for daytime and nighttime at short and long ranges, respectively. At nighttime, the use of the pedestrian-detection system is limited to well-illuminated areas. Another important factor, usually disregarded by most authors, is the effect of the candidate bounding box accuracy. Experimental results support the use of features based on contrast or edges, such as HON or cooccurrence over Canny images, as well as the development of an MC generation strategy, in order to assure that the issuance of some well-fitted candidates matches the samples used for training. Finally, we have presented the global performance of the system described in this paper, including candidate selection, MC generation, candidate detection using SVM, and pedestrian multiframe validation and tracking using Kalman filtering.

Although experimental results show that progress is being made in the right direction, further improvement needs to be made before deploying a really robust vision-based pedestrian detection system for assisted driving in real traffic conditions. For this purpose, motion-based and position-dependent features will be incorporated, which aim at enhancing the shape-based pedestrian detection algorithm developed in this paper. Two further actions are being carried out at the moment in order to improve the presented system. Additional classifiers are being introduced to detect motorbikes, urban furniture, and so on. This measure aims at decreasing the false alarm rate. The system is also being ported to an Apple MiniMac computer where optimization using ALTIVEC is being carried out in order to reduce the correlation computational cost.

References

[1] A. Shashua, Y. Gdalyahu, and G. Hayun, "Pedestrian detection for driving assistance systems: Single-frame classification and system level performance," in *Proc. IEEE Intell. Veh. Symp.*, Parma, Italy, Jun. 2004, pp. 1–6.

[2] A. Broggi, M. Bertozzi, A. Fascioli, and M. Sechi, "Shape-based pedestrian detection," in *Proc. IEEE Intell. Veh. Symp.*, Dearborn, MI, Oct. 2000, pp. 215–220.

[3] D. M. Gavrila and V. Philomin, "Real-time object detection for smart vehicles," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, 1999, pp. 87–93.

[4] F. Xu, X. Liu, and K. Fujimura, "Pedestrian detection and tracking with night vision," *IEEE Trans. Intell. Transp. Syst.*, vol. 6, no. 1, pp. 63–71, Mar. 2005.

[5] F. Xu and K. Fujimura, "Pedestrian detection and tracking with night vision," in *Proc. IEEE Intell. Veh. Symp.*, Versailles, France, Jun. 2002, pp. 21–30.

[6] B. Fardi, U. Schuenert, and G. Wanielik, "Shape and motion-based pedestrian detection in infrared images," in *Proc. IEEE Intell. Veh. Symp.*, Las Vegas, NV, Jun. 2005, pp. 18–23.

[7] M. Bertozzi, A. Broggi, A. Lasagni, and M. D. Rose, "Infrared stereo vision-based pedestrian detection," in *Proc. IEEE Intell. Veh. Symp.*, Las Vegas, NV, Jun. 2005, pp. 24–29.

[8] M. Bertozzi, A. Broggi, R. Chapuis, F. Chausse, A. Fascioli, and A. Tibaldi, "Shape-based pedestrian detection and localization," in *Proc. IEEE ITS Conf.*, Shanghai, China, Oct. 2003, pp. 328–333.

[9] C. Papageorgiou and T. Poggio, "A trainable system for object detection," *Int. J. Comput. Vis.*, vol. 38, no. 1, pp. 15–33, 2000.

[10] 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.

[11] D. M. Gavrila, J. Giebel, and S. Munder, "Vision-based pedestrian detection: The protector system," in *Proc. IEEE Intell. Veh. Symp.*, Parma, Italy, Jun. 2004, pp. 13–18.

[12] H. Nanda and L. Davis, "Probabilistic template based pedestrian detection in infrared videos," in *Proc. IEEE Intell. Veh. Symp.*, Versailles, France, Jun. 2002, pp. 15–20.

[13] H. Cheng, N. Zheng, and J. Qin, "Pedestrian detection using sparse Gabor filter and support vector machine," in *Proc. IEEE Intell. Veh. Symp.*, Las Vegas, NV, Jun. 2005, pp. 583–587.

[14] F. Suard, A. Rakotomamonjy, A. Bensrhair, and V. Guigue, "Pedestrian detection using stereo-vision and graph kernels," in *Proc. IEEE Intell. Veh. Symp.*, Las Vegas, NV, Jun. 2005, pp. 267–272.

[15] U. Franke and S. Heinrich, "Fast obstacle detection for urban traffic situations," *IEEE Trans. Intell. Transp. Syst.*, vol. 3, no. 3, pp. 173–181, Sep. 2002.

[16] C. Curio, J. Edelbrunner, T. Kalinke, C. Tzomakas, and W. V. Seelen, "Walking pedestrian recognition," *IEEE Trans. Intell. Transp. Syst.*, vol. 1, no. 3, pp. 155–163, Sep. 2000.

[17] U. Franke, D. Gavrila, S. Gorzig, F. Lindner, F. Puetzold, and C. Wohler, "Autonomous driving goes downtown," *IEEE Intell. Syst. Their Appl.*, vol. 13, no. 6, pp. 40–48, Nov./Dec. 1998.

[18] L. Zhao and C. E. Thorpe, "Stereo and neural network-based pedestrian detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 1, no. 3, pp. 148–154, Sep. 2000.

[19] M. Szarvas, A. Yoshizawa, M. Yamamoto, and J. Ogata, "Pedestrian detection with convolutional neural networks," in *Proc. IEEE Intell. Veh. Symp.*, Las Vegas, NV, Jun. 2005, pp. 224–229.

[20] G. Grubb, A. Zelinsky, L. Nilsson, and M. Rilbe, "3d vision sensing for improved pedestrian safety," in *Proc. IEEE Intell. Veh. Symp.*, Parma, Italy, Jun. 2004, pp. 19–24.

[21] D. Fernández, I. Parra, M. A. Sotelo, L. M. Bergasa, P. Revenga, J. Nuevo, and M. Ocaña, "Pedestrian recognition for intelligent transportation systems," in *Proc. ICINCO*, Barcelona, Spain, Sep. 2005, pp. 292–297.

[22] P. Viola, M. J. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," in *Proc. ICCV*, 2003, pp. 734–741.

[23] C. Hernández, "Sistema de asistencia a la conducción de vehículos de carretera mediante la detección y aviso de la salida del carril," M.S. thesis, Univ. Alcalá, Madrid, Spain, 2005.

[24] M. A. Sotelo, F. J. Rodríguez, and L. Magdalena, "Virtuous: Vision-based road transportation for unmanned operation on urban-like scenarios," *IEEE Trans. Intell. Transp. Syst.*, vol. 5, no. 2, pp. 69–83, Jun. 2004.

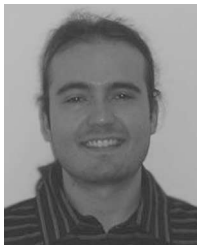
[25] M. A. Sotelo, F. J. Rodríguez, L. Magdalena, L. M. Bergasa, and L. Boquete, "A color vision-based lane tracking system for autonomous driving on unmarked roads," *Auton. Robots*, vol. 16, no. 1, pp. 95–116, Jan. 2004.

[26] R. Labayrade, C. Royere, D. Gruyer, and D. Aubert, "Cooperative fusion for multi-obstacles detection with use of stereovision and laser scanner," in *Proc. Int. Conf. Adv. Robot.*, 2003, pp. 1538–1543.
[27] B. Boufama, "Reconstruction tridimensionnelle en vision par ordinateur: Cas des cameras non etalonnees," Ph.D. dissertation, INP de Grenoble, Grenoble, France, 1994.
[28] S. Chiu, "Fuzzy model identification based on cluster estimation," *J. Intell. Fuzzy Syst.*, vol. 2, no. 3, pp. 267–278, 1994.
[29] F. J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-8, no. 6, pp. 679–698, Nov. 1986.
[30] R. M. Haralick, "Statistical and structural approaches to texture," *Proc. IEEE*, vol. 67, no. 5, pp. 786–804, May 1979.
[31] N. Dalai and B. Triggs, "Histogram of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. and Pattern Recog.*, 2005, pp. 886–893.
[32] L. Wang, "Texture unit, texture spectrum and texture analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 28, no. 4, pp. 509–512, Jul. 1990.
[33] J. Nuevo. (2005). *Tsetbuilder Tutorial, Technical Report*. [Online]. Available: ftp://www.depeca.uah.es/pub/vision/SVM/manual.pdf
[34] J. C. Christopher, "A tutorial on support vector machines for pattern recognition," *Data Mining Knowl. Discov.*, vol. 2, no. 2, pp. 121–167, Jun. 1998.
[35] C. C. Chang and C.-J. Lin. (2001). *LIBSVM: A Library for Support Vector Machines*. [Online]. Available: http://www.csie.nut.edu.tw/ cjlin/libsvm

Ignacio Parra Alonso was born in Madrid, Spain, in 1979. He received the M.S. degree in telecommunications engineering from the University of Alcalá (UAH), Madrid, in 2005. He is currently working toward the Ph.D. degree in egomotion computing onboard a road vehicle at UAH.

He is currently a member of the research staff at UAH. His areas of interest include intelligent transportation systems, intelligent vehicles, artificial vision, and operating systems. He has run several courses on GNU/Linux.

Mr. Parra Alonso was the recipient of the Master Thesis Award in eSafety from ADA Lectureship at the Technical University of Madrid in 2006.



David Fernández Llorca was born in Madrid, Spain, in November 1980. He received the M.S. degree in telecommunications engineering from the University of Alcalá (UAH), Madrid, in 2003. He is currently working toward the Ph.D. degree at UAH.

He is currently a Teaching Assistant at UAH. His research interests are mainly focused on the application of image processing to eSafety and ITS applications.

Mr. Llorca was the recipient of the Master Thesis Award in eSafety from ADA Lectureship at the Technical University of Madrid in 2004 and the Best Telecommunication Engineering Student award, also in 2004.

Miguel Ángel Sotelo (M'02) received the Dr. Ing. degree in electrical engineering from the Technical University of Madrid, Madrid, Spain, in 1996 and the Ph.D. degree in electrical engineering from the University of Alcalá (UAH), Madrid, in 2001.

From 1993 to 1994, he was a Researcher at the Department of Electronics, UAH. He is currently an Associate Professor at the UAH. He is the author of more than 100 refereed publications in international journals, book chapters, and conference proceedings. His research interests include real-time computer vision and control systems for autonomous and assisted intelligent road vehicles.

Dr. Sotelo is a member of the IEEE ITS Society and the ITS-Spain Committee. He has been serving as Auditor and Expert at the FITSA Foundation for R+D Projects in the domain of automotive applications since September 2004. He was the recipient of the Best Research Award in the domain of Automotive and Vehicle Applications in Spain in 2002, the 3M Foundation Awards in the category of eSafety in 2003 and 2004, and the Best Young Researcher Award from the UAH in 2004.

Luis M. Bergasa (M'04–A'05) received the M.S. degree in electrical engineering from the Technical University of Madrid, Madrid, Spain, in 1995 and the Ph.D. degree in electrical engineering from the University of Alcalá (UAH), Madrid, in 1999.

He is currently an Associate Professor at the Department of Electronics, UAH. He is the author of more than 80 publications in international journals, book chapters, and conference proceedings. His research interests include real-time computer vision and its applications, particularly in the field of the robotics, assistance systems for elderly people, and intelligent transportation systems.

Dr. Bergasa is a member of the Computer Science Society. He was the recipient of the Best Research Award of the 3M Foundation Awards in the Industrial category in 2004 and of the Best Spanish Ph.D. Thesis Award in Robotics from the Automatic Spanish Committee in 2005 as Director of the work.

Pedro Revenga de Toro received the Technical degree in telecommunications engineering from the University of Alcalá (UAH), Madrid, Spain, in 1989 and the Dr. Ing. degree in electronics engineering from the Technical University of Valencia, Valencia, Spain, in 2000.

Since 1990, he has been a Lecturer in the Department of Electronics, UAH. He is the author of more than 50 refereed publications in international journals, book chapters, and conference proceedings. His research interests include robotics, multisensorial integration, control electronics, parallel processing systems, and mobility assistance systems.

Jesús Nuevo received the M.S. degree in telecommunications engineering from the University of Alcalá (UAH), Madrid, Spain, in 2004, where he is currently working toward the Ph.D. degree.

His research interests include computer vision, autonomous vehicles, pattern recognition, and machine learning.

Manuel Ocaña received the Ing. degree in 2002 and the Ph.D. degree from the University of Alcalá (UAH), Madrid, Spain, both in electrical engineering.

From 2002 to 2005, he was a Researcher in the Department of Electronics, UAH, where he is currently an Associate Professor. He is the author of more than 20 refereed publications in international journals, book chapters, and conference proceeding. His research interests include robotics localization and navigation, assistant robotics and computer vision, and control systems for autonomous and assisted intelligent vehicles.

Dr. Ocaña was the recipient of the Best Research Award for the 3M Foundation Awards in the category of eSafety in 2003 and 2004.

Miguel Ángel García Garrido received the Industrial Engineering middle degree and the Electronic Engineering degree from the University of Alcalá (UAH), Madrid, Spain, in 1998 and 2001, respectively. He is currently working toward the Ph.D. degree in artificial vision at the UAH.

Since 2003, he has been a Lecturer in the Electronics Department, UAH. His research interests are in the area of intelligent transportation systems, including, driver assistance systems.