Pedestrian Recognition in Road Sequences

I. PARRA, D. FERNÁNDEZ, M. A. SOTELO, P. REVENGA, L. M. BERGASA, M. OCAÑA, J. NUEVO, R. FLORES
Department of Electronics
University of Alcalá
Escuela Politécnica Superior, Campus Universitario s/n, Alcalá de Henares, 28871
SPAIN
parra,llorca,sotelo,revenga,bergasa,mocana,jnuevo,flores@depeca.uah.es http://www.depeca.uah.es

Abstract: - This paper presents a vision-based pedestrian recognition method in the framework of Intelligent Transportation Systems. The basic components of pedestrians are first located in the image and then combined with a SVM-based classifier. This poses the problem of pedestrian detection and recognition in real, cluttered road images. Candidate pedestrians are located using a subtractive clustering attention mechanism. A distributed learning approach is proposed in order to better deal with pedestrians variability, illumination conditions, partial occlusions and rotations. An extensive comparison has been carried out using different feature extraction methods, as a key to image understanding in real traffic conditions. A database containing thousands of pedestrian examples extracted from real traffic images has been created for learning purposes. The results achieved up to date show interesting conclusions that suggest a combination of methods as an essential clue for optimal recognition performance

Key-Words: - Pedestrian recognition, ITS, vision, SVM, feature extraction, road vehicles

1 Introduction

For most authors it is clear that vision provides the main clues for pedestrian recognition although other sensors, such as laserscanners, have also been tested [1]. To ease the pedestrian recognition task a candidate selection mechanism is normally applied. The selection can be implemented by performing an object segmentation either in the 3D scene or in the 2D image plane. The first solution requires the use of stereo vision [2] [3], while the second one tackles the problem of candidate selection using a single camera (monocular vision). Only a few authors succeed in dealing with the problem of monocular pedestrian recognition to some extent [4]. The main problem with candidate selection mechanisms in monocular systems is that they are bound to yield a large amount of candidates per frame, in average, in order to ensure a low false negative ratio (the number of pedestrians that are not selected by the attention mechanism). Another problem in monocular systems is the fact that depth clues are lost unless some constraints are applied, such as the flat terrain assumption, which is not always applicable. This problem can be easily solved by using stereo vision systems, although other problems arise such as the need of maintaining callibration and the high computational cost required to implement dense algorithms. Nonetheless, with recent hardware advances, real-time dense stereo vision becomes increasingly feasible for general-purpose processors [5].

In this work, we present a solution for pedestrian recognition at daytime. Other systems already exist for pedestrian detection using night vision [6]. Nighttime detection is carried out using infrared cameras as long as they provide better visibility at night and under adverse weather conditions. However, infrared cameras do not provide the perfect solution. Thus, the use of infrared cameras is quite an expensive option that can make mass production an untractable problem, especially for the case of stereo vision systems where two cameras are required. They provide images that strongly depend on both weather conditions and the season of the year. For instance, a cyclist could not be perceived in the image provided by an infrared camera in a rainy day if the cyclist is wet. Additionally, infrared cameras need recalibration every year. In principle, the algorithm described in this paper is applied to visible cameras. Nonetheless, as soon as the technology for night vision cameras production becomes cheaper and more mature the results could be easily extended to a stereo night vision system.

Various approaches have been proposed in the literature based on shape analysis. Some authors use feature-based techniques, such as recognition by vertical linear features, symmetry, and human templates [7] [8], Haar wavelet representation [9] [10], hierarchical shape templates on Chamfer distance [11], correlation with probabilistic human templates [12], and principal component analysis [13]. Others use neural network-based classifiers [14]. In the last years, Support Vector Machines

(SVM) have been widely used by many researchers [9] [10] [4] [3] [15] as they provide a supervised learning approach for object recognition as well as a separation between two classes of objects. This is especially useful for the case of pedestrian recognition. Combinations of shape and motion are used as an alternative to improve the classifier robustness [16] [4]. Some authors have demonstrated that the recognition of pedestrians by components yields superior performance than the recognition of the entire body [10] [15]. Additionally, in [6] an interesting discussion is presented about 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.

SVM-based classifiers fall into the category of object detection techniques characterised by example-based learning algorithms. This type of technique can provide a solution to the pedestrian recognition problem as long as a sufficiently large number of pedestrians examples are contained in the database and the examples are representative of the pedestrian class in terms of variability, illumination conditions, position and size in the image. Example-based techniques are easy to use with objects composed of distinct identifiable parts arranged in a well-defined configuration. A learning approach based on components [10]is more efficient for object recognition in real cluttered environments than holistic approaches [9] as long as the first can deal with partial occlusions and is less sensitive to object rotations. However, in spite of their ability to detect objects in real images, we propose to reduce the pedestrians searching space in an intelligent manner based on the road image, so as to increase the performance of the detection module. Accordingly, road lane markings are detected and used as the guidelines that drive the pedestrian searching process. The area contained by the limits of the lanes determines the zone of the real 3D scene where pedestrians are searched for. The objects found in the searching area are passed on to the pedestrian recognition module. This helps reduce the rate of false positive detections. In case that no lane markings are detected, a basic area of interest is used instead covering the front part ahead of the ego-vehicle. The description of the lane marking detection system is provided in [17]. The rest of the paper is organised as follows: section II describes the candidate selection mechanism. Section III provides a description of several feature extraction methods. The comparative results achieved up to date are presented in section IV. Finally, section V summarizes the conclusions and future works.

2 Candidate Selection

We have developed a calibrated stereo vision platform and calculated the intrinsic parameters for each camera, as well as the extrinsic parameters between them, in order to obtain the fundamental matrix that defines the system epipolar geometry. This way the perfect physically alignment between cameras that implies the assumption of parallel epipolar lines, is not necessary, because the stereo calibration process defines mathematically the geometric relationships for the cameras [18]. The first task is image preprocessing which has two steps: normalize intensity values, to correct for differences between the two images, and eliminate radial and tangential distortion. Once here, we apply a Canny algorithm for feature extraction on the left image. The Canny image provides a good representation of the discriminating features of pedestrians. Features such as heads, arms and legs are visible and distinguishable and are not affected by colours or intensity. It provides clear indications about discriminating zones for the pedestrian recognition system.

Our approach creates a 3D map whose origin is placed at the left camera. Using the fundamental matrix for each Canny's detected point we search the corresponding point in the other image along its epipolar line (fixing the maximum distance between corresponding points in order to reduce the matching computational cost). The correspondence problem can be solved by using a wide spectrum of matching techniques. Most recent successes have been in area-based algorithms, specifically the *Zero Mean Normalized Cross Correlation* has performed most robustly [19].

Despite this the correspondences are often not correct due to occlusions and repetitive patterns and textures. According to the previous statements we need a filtering criteria in order to reject outliers. We create a bird's eye map and first, we extract 3D points within the pedestrian searching area (after the road lanes marking detecting system[17]). Secondly, road surface points (road markings) and high points, above 2m, are removed. Finally we filter the XZ map according to a neighbourhood criterion. As depicted in Figure 1, the appearance of pedestrians in 3D space is represented by an uniformly distributed set of points. Data clustering techniques are related to the partitioning of a data set into several groups. The common approach of clustering techniques is to find clusters centers that will represent each cluster.

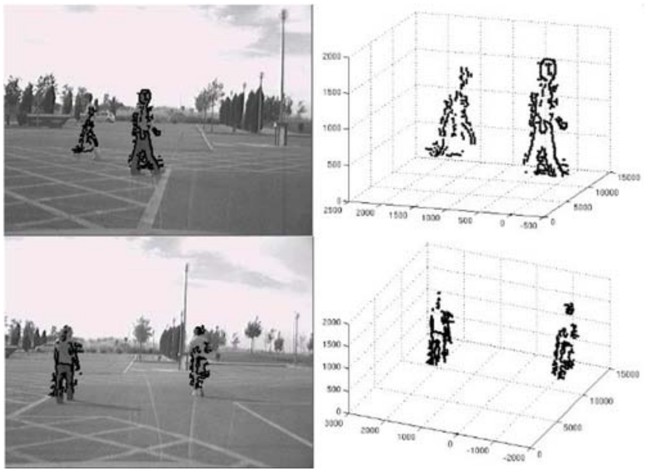

Figure 1. Left: 2D points overlay on left image. Right: 3D points location.

Normally, the number of clusters is known beforehand. This is the case of *K-means* based algorithms. In our case the number of clusters is unknown, outlier effects have to be reduced or completely eliminated and it is neccesary to define specific space characteristics in order to group different pedestrians in the scene. For these reasons we use the *Subtractive Clustering* [20] that is applied in *Fuzzy Model Identification Systems*. The idea is to find regions in the feature space with high densities of data points. The point with the highest number of neighbours is selected as a cluster centre. The data points within a prespecified neighborhood radius are then removed (subtracted), and the algorithm looks for a new point with the highest number of neighbours. We carry out this algorithm using a 3-dimensional neighbourhood radius [15]. Pedestrian classification will be done in 2D in the ROI defined by the image projection of the 3D candidate regions. Figure 2 depicts the multicandidate regions of interest generated by the clustering mechanism in a sequence of images

3 Feature Extraction

The appearance of pedestrians in the scene presents a high intraclass variability (moving longitudinally, moving laterally, stationary, different shapes, clothes, etc.). In consequence, it makes sense to use a by-components learning approach in which each pedestrian body part is independently learnt by a specialized classifier in a first learning stage. The body local parts are then integrated by another classifier in a second learning stage. The use of independent classifiers in a distributed manner simplifies the learning process.

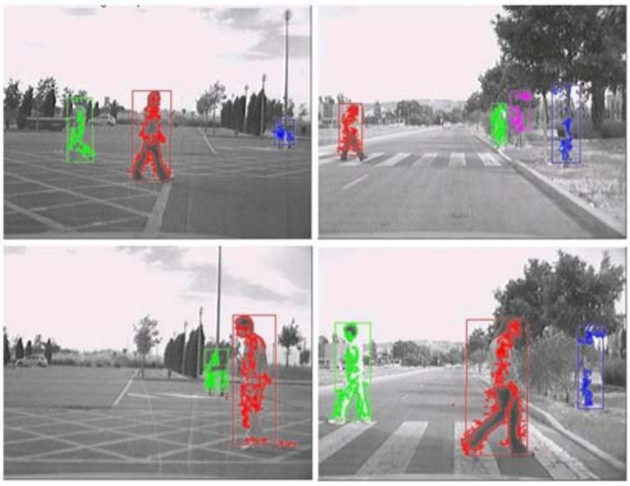

Figure 2. Generation of candidate regions of interest in a sequence of images.

Otherwise, it would be difficult to attain an acceptable result using a holistic approach. We have considered a total of 6 different sub-regions for each candidate region of interest which has been fit to a size of 24×72 pixels. The locations of the six-regions have been chosen in an attempt to detect coherent pedestrian features as depicted in Figure 3. A set of features must be extracted from each sub-region and fed to the classifier. The choice of the most appropriate features for pedestrian characterization remains a challenging problem nowadays. We show that performance depends crucially on the features that are used to represent the pedestrians. Several feature extraction methods have been tested: canny edge detector, cooccurrence matrix over canny edge image and over normalized gray image, gray intensity differences histogram over the normalized image, image gradient magnitude and orientation and finally texture unit number.The canny edge detector [21] finds the image gradient to highlight regions with high spacial derivatives. It is known to many as the optimal edge detector. Edge detecting an image significantly reduces the amount of data and filters out useless information, while preserving the important structural properties in an image. The feature vectors created have a variable dimension depending on the zone size in pixels.

The coocurrence [22] can be specified in a matrix of relative frequencies $P_{i,j}$ with which two neighbouring pixels separated by distance d at orientation θ occur in the image, one with gray or binary tone i and the other with gray or binary tone j, depending on whether we compute coocurrence over gray level image or canny image. Resulting matrices are symmetric and can be normalized by dividing each entry in a matrix by the number of neighboring pixels used in computing that matrix.

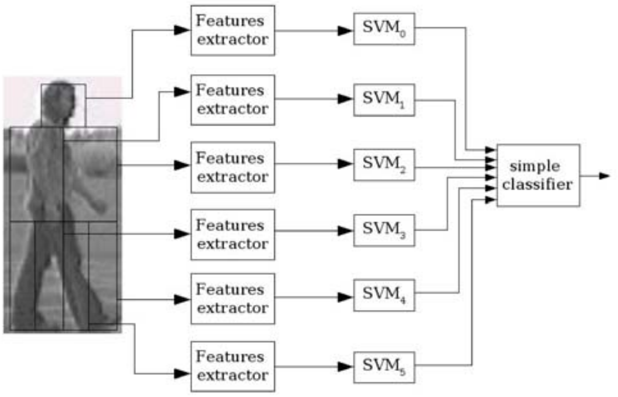

Figure 3. Global structure of the two-stage SVM classifier.

In our approach we use one pixel distance and four orientations (0°,45°,90°,135°). Its features vectors have four 2×2 coocurrence matrices when computing over canny image, and 4 32×32 coocurrence matrices in case we compute over 32 gray levels normalized image. he gray histogram computes the relatives frequencies of the intensity differences between neighbouring pixels along four orientations over 128 gray levels normalized image. This generates four 128-lengthed vectors per subregion. For the gradient magnitude and orientation we calculate the spatial derivatives of the image, g_x and g_y and compute its magnitude. Then we calculate the orientation from $\theta = atan(g_x, g_y)$. The resulting vector has twice the window size in pixels. The texture unit number was proposed in [23]. The local texture information for a pixel can be extracted from a neighbourhood of 3×3 pixels, which represents the smallest complete unit in the sense of having eight directions surrounding pixels. The corresponding texture unit is defined by a set containing eight elements. The NTU process generates a vector with the same size as the zone in pixels.

Support Vector Machines (SVM) classifiers, proposed by [24] have yielded excellent results in various data classification tasks, including people detection [9]. The SVM algorithm uses structural risk minimization to find the hyperplane that optimally separates two classes of objects. We use it in order to classify each candidate as either pedestrian or non-pedestrian. The global training strategy is carried out in two stages, as depicted in Figure 3. In a first stage, separate SVM-based classifiers are trained using individual training sets that represent a subset of a sub-region. Each SVM classifier produces a theorical output between -1 (non-pedestrian) and +1 (pedestrian). Accordingly, it can be stated that this stage provides classification of individual parts of the candidate sub-regions. In a second step, the outputs of all classifiers are merged in a simple classifier which makes a decision based on a distance criterion in order to provide the final classification result. Once here, each candidate classified as pedestrian is dynamically tracked by a Kalman filter [25] which decreases the false negative rate.

4 Experimental Results

The system was implemented on a Pentium IV at 2.4 Ghz running the Knoppix GNU/Linux Operating System. With 320×240 pixel images resolution, the complete algorithm runs at an average rate of 20 frames/s depending on the number of pedestrians being tracked and their position. Specifically the average rate has a strong dependency on the number of correlated points because of the correlation computacional cost, which consumes 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 25m, developing a practical false-negative rate of 0%, after the kalman filtering. Once the selection of pedestrians as candidates is granted the false-positive rate is expected to be corrected by the SVM classifier. We created several databases containing thousans of samples of pedestrians and non-pedestrian in different situations. The number of pedestrians samples in the training sets were usually chosen to be similar to the number of non-pedestrian samples although some training sets were created with a different possitive/negative ratio to evaluate its influence in the performance of the classifier. These candidates were extracted from recorded images acquired in real experiments onboard a road vehicle under real traffic conditions. All training sets were created at day time conditions using the TSetBuilder tool, specifically developed in this project for this purpose. By using the TSetBuilder tool different candidate regions are manually selected in the image on a frame-by-frame basis. Special attention was given to the selection of non-pedestrian samples. If we select simple non-pedestrian examples (for instance, road regions) the system learns very quickly but it does not develop enough discriminating capability in practice, as the attention mechanism can select a region of the image that might be very similar to a pedestrian but it is not a pedestrian in reality. The training of all SVM classifiers was performed using the free-licence LibTorch libraries for Linux. Different SVM classifiers were trained for each one of the feature extraction methods.

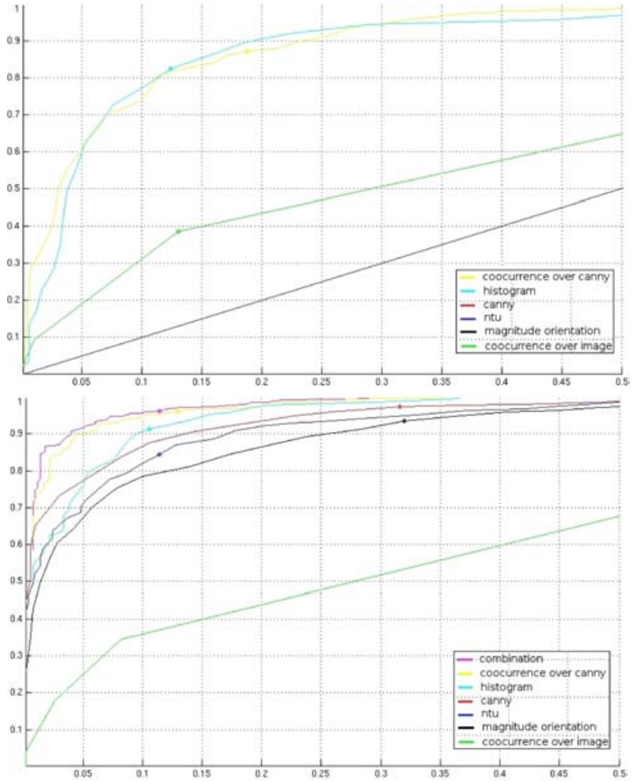

Figure 4. ROC performance. Top: holistic approach, Bottom: by components approach.

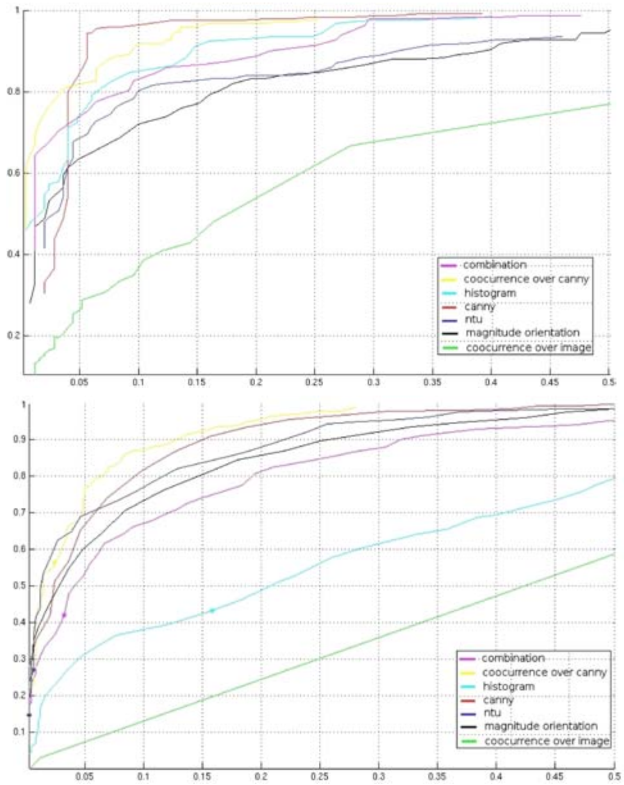

Figure 5. ROC performance. Top: 2nd classifier with SVM, Bottom: bounding box accuracy

In the holistic approach most of the feature extraction methods fed too much information to the SVM which was not able to generalize a model, leading to useless classifiers. The performance of the few ones which produced trained models were largely improved by the by components approach, as we can see in the Receiver Operation Curves (ROC) in Figure 4, while the by components approach managed to train almost every feature extraction method. This shows that breaking the pedestrian into smaller pieces and training the SVM specifically for these pieces reduces the variability and lets the SVM generalize the models much better. The coocurrence matrices over canny edge extractor image developed the best single frame performance with a 90% of detection rate and a 5% of false positive rate. But we could even improve these results by combining different feature extraction methods, the best for each one of the six splitted zones, in a second classifier getting a 90% of detection rate with a 3% of false positive rate (Figure 4). Now we are developing an adaptative normalization value computed by summarizing the different values obtained on running time. We have also studied the importance of the second classifier which combines the information delivered by the six specifically trained SVM models.

The results achieved up to date show that the simple classifier performs best in most cases, but also certain feature extraction methods have been highly improved as depicted in Figure 5. In most vision based detection system the candidates are first estimated in one image and then delivered to be validated in a second stage. The accuracy they can show bounding their candidates is limited, and in fact most of them perform a multiple candidate generation for each one of the candidates extracted from the first stage to ensure the best performance in the classifier. In order to reduce the effect of these badly bounded candidates and allow our system generate very few candidates we have studied the effects of the bounding box accuracy for the different feature extraction methods. We have trained models with well fitted candidates and tested these models over testing sets with badly bounded candidates. The results show that some feature extraction methods as coocurrence over canny image are nearly unaffected while most of them decrease their performance (Figure 5). This suggests that choosing your feature extraction method just in terms of detection rate and false positive rate can lead, in real working conditions, to decreasing its performance or to be forced to generate a huge number of candidates in order to get at least one well fitted candidate. The performance of the single-frame recognition process is largely increased by using multiframe validation. The probability of a candidate region being classified as pedestrian is modelled as a Bayesian random variable.

5 Conclusions

We have performed a comparative study of feature extraction methods for vision-based pedestrian recognition. The learning process has been simplified by decomposing the candidate regions into 6 local sub-regions that are easily learned by individual SVM classifiers. The distributed approach has yielded, superior performance compared to the holistic version. It has been shown that combining different feature extraction methods improves the system performance, which opens a wide spectrum of combinations in order to enhance the classification. It has also been proved the importance of carrying out a comparative study of feature extraction methods to evaluate its robustness against other variables. In this way we have proved the importance of the bounding box accuracy in the candidate selection proccess. Currently we are studying the efect of the distance to the candidate in the SVM learning capability. The content of the training sets will be largely increased by including new and more complex samples. In addition, a gait recognition process will be introduced to enhance the shape-based pedestrian detection algorithm. Finally a Single Instruction Multiple Data (SIMD) optimization will be developed in order to reduce the correlation computacional cost.

References:

[1] Fuerstenberg, K.C., Dietmayer, K.J., Willhoeft, V.: Pedestrian recognition in urban traffic using a vehicle based multilayer laserscanner. *In: In Proc. IEEE Intelligent Vehicles Symposium*, Versailles, France, June 2002 (2002)

[2] Gavrila, D.M., Giebel, J., Munder, S.: Vision-based pedestrian detection: The protector system. *In Proc. IEEE Intelligent Vehicles Symposium*, pp. 13-18, Parma, Italy, June 14-17 (2004)

[3] Grubb, G., Zelinsky, A., Nilsson, L., Rilbe, M.: 3d vision sensing for improved pedestrian safety. *In Proc. IEEE IV Symposium*, pp. 19-24, Italy (2004)

[4] Shashua, A., Gdalyahu, Y., Hayun, G.: Pedestrian detection for driving assistance systems: single-frame classification and system level performance. *In Proc. IEEE Intelligent Vehicles Symposium*, pp. 1-6, Parma, Italy (2004)

[5] Sunyoto, H., Mark, W., Gavrila, D.M.: A comparative study of fast dense stereo vision algorithms. *In Proc. IEEE Intelligent Vehicles Symposium*. (2004)

[6] Xu, F., Liu, X., Fujimura, K.: Pedestrian detection and tracking with night vision. *In: IEEE Transactions on ITS*, vol 6 No. 1, March 2005 (2005)

[7] Broggi, A., Bertozzi, M., Fascioli, A., Sechi, M.: Shape-based pedestrian detection. *In Proc. IEEE Intelligent Vehicles Symposium*. (2000)

[8] Bertozzi, M., Broggi, A., Chapuis, R., Chausse, F., Fascioli, A., Tibaldi, A.: Shape-based pedestrian detection and localization. *In Proc. IEEE ITSC*. (2003)

[9] Papageorgiou, C., Poggio, T.: A trainable system for object detection. *In: Intl J. Computer Vision*, Vol. 38, No. 1, pp. 15-33 (2000)

[10] Mohan, A., Papageorgiou, C., Poggio, T.: Example-based object detection in images by components. *In: IEEE Transactions on Pattern Analisis and Machine Intelligence*, Vol. 23 No. 4 (2001)

[11] Gavrila, D.M., Philomin, V.: Real-time object detection for smart vehicles. *In: Computer Vision*, 1999. The Proceedings of the Seventh IEEE International Conference on, ISBN: 0-7695-0164-8 (1999)

[12] Nanda, H., Davis, L.: Probabilistic template based pedestrian detection in infrared videos. *In Proc. IEEE Intelligent Vehicles Symposium*. (2002)

[13] Franke, U., Gavrila, D., Gorzig, S., Lindner, F., Puetzold, F., Wohler, C.: Autonomous driving goes downtown. *In: Intelligent Systems and Their Applications, IEEE*, vol 13, pp 40-48 (1998)

[14] Zhao, L., Thorpe, C.E.: Stereo and neural network-based pedestrian detection. *In Proc. IEEE Transactions on ITS*, Vol.1, No.3 September (2000)

[15] Fernández, D., Parra, I., Sotelo, M.A., Bergasa, L.M., Revenga, P., Nuevo, J., Flores, R.: Pedestrian recognition for intelligent transportation systems. *In Proc. ICINCO*, Barcelona, Spain, September 2005 (2005)

[16] Viola, P., Jones, M.J., Snow, D.: Detecting pedestrians using patterns of motion and appearance. *In ICCV 2003*, pp 734-741 (2003)

[17] Sotelo, M.A., Nuevo, J., Bergasa, L.M., Ocana, M.: Road vehicle recognition in monocular images. *In Proc. ISIE 2005*, Duvrobnik, Croatia June 2005 (2005)

[18] Xu, G., Zhang, Z.: Epipolar Geometry in Stereo, Motion and Object Recognition: A Unified Approach. *1st edn. Kluwer Academic Publishers*, London (1996)

[19] Boufama, B.: Reconstruction tridimensionnelle en vision par ordinateur: Cas des cameras non etalonnees. *In: PhD thesis*, INP de Grenoble, France (1994)

[20] Chiu, S.: Fuzzy model identification based on cluster estimation. *In: J. of Intelligent and Fuzzy Systems*, vol. 2, no. 3, pp. 267-278, 1994 (1994)

[21] Canny, F.J.: A computational approach to edge detection. *In: IEEE Trans PAMI*. (1986)

[22] Haralick, R.M.: Statistical and structural approaches to texture. *In: Procdedings of the IEEE*, vol 67, No 5, May 1979 (1979)

[23] Wang, L.: Texture unit, texture spectrum and texture analysis. *In: IEEE Transactions on Geosciences and Remote Sensing*, Vol.28, No4, pp. 509-512 (1990)

[24] Vapnik, V.: The nature of statistical learning theory, *Springer Verlag* (1999)

[25] Kalman, R.: A new approach to linear filtering and prediction problems. *In: Trans. ASME Journal of Basic Engineering*, vol. 82, no. 1,pp. 35 45 (1960)