PEDESTRIAN RECOGNITION FOR INTELLIGENT TRANSPORTATION SYSTEMS

D. Fernández, I. Parra, M. A. Sotelo, L. M. Bergasa, P. Revenga, J. Nuevo, M. Ocaña
Department of Electronics. University of Alcalá
Alcalá de Henares, Madrid, Spain
Email: llorca,parra,michael,bergasa,revenga,jnuevo,mocana@depeca.uah.es

Keywords: Pedestrian Recognition, Support Vector Machines, Stereovision, Intelligent Transportation Systems.

Abstract: This paper describes a binocular vision-based pedestrian recognition System. 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. The performance of the pedestrian recognition system is enhanced by a multiframe validation process. By doing so, the detection rate is largely increased. A database containing hundreds of pedestrian examples extracted from real traffic images has been created for learning purposes. We present and discuss the results achieved up to date.

1 INTRODUCTION

This paper describes a binocular vision-based pedestrian recognition system in the framework of Intelligent Transportation Systems (ITS) technologies. In our approach, the basic components of pedestrians are first located in the image and then combined with a SVM-based classifier. The challenge is to use a couple of FireWire digital cameras as input, in order to achieve a low cost final solution that meets the requirements needed to undertake serial production. The digital cameras provide range measurements using the laws of stereo vision. Some previous works use available sensing methods such as laserscanner (Fuerstenberg et al., 2002), stereovision (Gavrila et al., 2004) (Grubb et al., 2004), or a combination of both (Labayrade et al., 2003). Only a few works deal with the problem of monocular pedestrian recognition using pattern recognition techniques (Shashua et al., 2004). Pedestrian recognition is a challenging problem in real traffic, cluttered environments. This is a complex problem as long as it requires that the object class exhibits high interclass and low intraclass variability. In addition, pedestrian recognition should perform robustly under variable illumination conditions, variable rotated positions, and even if some of the pedestrian parts or limbs are partially occluded.

Object recognition techniques can be classified into three major categories, as described in (Mohan et al., 2001). The first category is represented by model-based systems in which a model is defined for the object of interest and the system attempts to match the model to different parts of the image in order to find a fit. Unfortunately, pedestrians can be regarded as quite a variable class that makes it impossible to define a model that represents the class in an accurate, general way. In consequence, model-based systems are of little use for pedestrians recognition purposes. The second category are image invariance methods which perform a matching based on a set of image pattern features that, supposedly, uniquely determine the object being searched for. Pedestrians do not exhibit any deterministic image pattern relationships because of its large variability (size, pose and so forth). For this reason, image invariance methods are not a viable option in order to solve the pedestrian recognition problem. The third category of object detection techniques is characterised by example-based learning algorithms. The salient features of a class are learnt by the system based on a set of examples. This type of technique can provide a solution to the pedestrian recognition problem as long as the following conditions are met.

- A sufficiently large number of pedestrians examples are contained in the database.

- The examples are representative of the pedestrian class in terms of variability, illumination conditions, position and size in the image.

Example-based techniques have been previously used in natural, cluttered environments for pedestrian detection (Shashua et al., 2004) (Gavrila et al., 2004). In general, these techniques are easy to use with objects composed of distinct identifiable parts arranged in a well-defined configuration. A distributed learning approach based on components (Mohan et al., 2001) is more efficient for object recognition in real cluttered environments than holistic approaches (Papageorgiou and Poggio, 2000). Distributed learning techniques can deal with partial occlusions and are 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 (Sotelo et al., 2005). The rest of the paper is organised as follows: section II provides a description of the candidate selection mechanism. Section III describes the pedestrian recognition system. The results achieved up to date are presented in section IV. Finally, section V summarizes the conclusions and future work.

2 CANDIDATE SELECTION

We have developed a calibrated stereo platform and calculated the intrinsic parameters for each camera, and the extrinsic parameters between them, in order to obtain the fundamental matrix that defines the system epipolar geometry. This way the perfect physically aligning 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 (Xu and Zhang, 1996).

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, as depicts Figure 1. Features such as heads, arms and legs are visible and distinguishable and are not affected by colours or intensity. It gives us some indications about discriminating zones for the pedestrian recognition system.

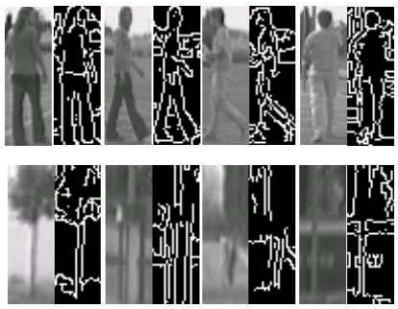

Figure 1: Some Canny images examples. Upper row: pedestrians examples. Bottom row: non pedestrians examples

In order to extract 3D scene information some authors use disparity map techniques combined with the v-disparity segmentation (Grubb et al., 2004) (Labayrade et al., 2003). This option was discarded because of the disadvantages associated with disparity computation algorithms: prior to disparity map generation the image pair has to be rectified to ensure good correspondence matching. In addition the information for performing generic obstacles detection is defined with a vertical line into the v-disparity image. This implies managing very little information to detect obstacles, which works for big object detection as vehicles, but could not be enough for smaller object detection such as pedestrians. After solving the correspondence problem, our approach creates a 3D points map which 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 cost of matching).

The correspondence problem can be solved using a wide spectrum of matching techniques. But most recent successes have been in area-based algorithms. Specifically the *Zero Mean Normalized Cross Correlation* has performed most robustly (Boufama, 1994). This algorithm seeks -for a point given on the left image- the larger correlation response for a point of the right image, taking into account the relevance of the window size. As the window size decreases, the discriminatory power of the area-based criterion is decreased and some local maximum in ZMNCC could have been found in the search regions. Moreover, continually increasing the window size causes the perfor-

mance to degrade because of occlusion regions and smoothing of disparity values across depth boundaries. In consequence the correspondences are often not correct.

According to the previous statements we need a filtering criteria in order to reject outliers. We create a XZ map (bird's eye map) and first, we extract 3D points within the pedestrian searching area (after the road lanes marking detecting system). Second, road surface points (road drawings) and high points, above 2m, are removed. And finally we filter the XZ map according to a neighbourhood criterion. Figure 2 depicts the filtering criteria.

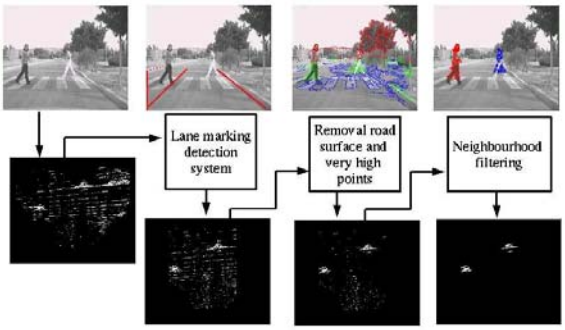

Figure 2: Filtering criteria and XZ maps.

As we can see in Figure 3, the appearance of pedestrians in 3D space is represented by an uniformly distributed set of points. Data clustering techniques is concerned with the partitioning of a data set into several groups such that the similarity within a group is larger than that among groups. The common approach of clustering techniques is to find clusters centers that will represent each cluster and 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* (Chiu, 1994) that is applied in *Fuzzy Model Identification Systems* and is based on a measure of the density of data points. 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 centre for a cluster. 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 $r_a = (r_{ax}, r_{ay}, r_{az})$. Since each data point is a candidate for a cluster centre, a density measure at data point $\mu_k = (x_k, y_k, z_k)$ is de-

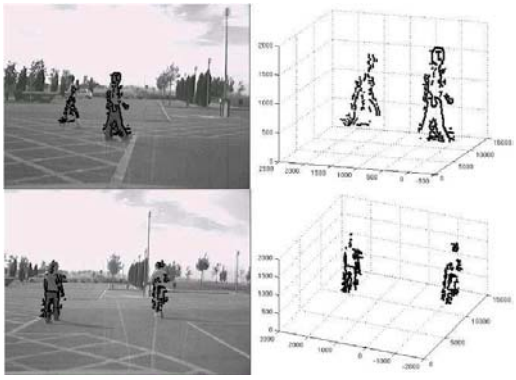

Figure 3: Left: 2D points into left image. Right: 3D points location.

fined as

$$D_k = \sum_{j=1}^{K} exp\left(-\frac{||\mu_k - \mu_j||}{(r_a/2)^2}\right) \quad (1)$$

Let μ_C be the point with highest density and D_C its density measure. Next, the density measure for each point μ_k is revised by the formula

$$D'_k = D_k - D_C exp\left(-\frac{||\mu_k - \mu_C||}{(r_b/2)^2}\right) \quad (2)$$

where r_b defines a neighbourhood to be reduced in density measure and it is normally larger than r_a to prevent closely spaced cluster centres, typically $r_b = 1.5r_a = (1.5r_{ax}, 1.5r_{ay}, 1.5r_{az})$.After the density measure for each point is revised, the next cluster centre is selected and all the density measures are revised again. The process is repeated until a sufficient number of cluster centres are generated. After applying subtractive clustering to a set of input data, each cluster represents a candidate. Pedestrian classification will be done in 2D in the ROI defined by the image projection of the 3D candidate regions. Figure 4 depicts the multicandidate regions of interest generated by the clustering mechanism in a sequence of images. Nonetheless, this figure is bound to change depending on traffic conditions.

3 PEDESTRIAN RECOGNITION

The appearance of pedestrians in the scene presents a wide variability (moving longitudinally, moving laterally, stationary, etc.). In consequence, it makes sense to use a distributed 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

Figure 4: Generation of candidate regions of interest in a sequence of images.

second learning stage. The proposed approach can be regarded as a hierarchical one. By using independent classifiers in a distributed manner the learning process is simplified, as long as a single classifier has to learn individual features of local regions in certain conditions. 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 first sub-region is located in the head. The arms and legs are covered between the second and fifth regions. In addition we define a region located between the legs, covering an area with relevant information depending on the pedestrian pose. The locations of the six-regions have been chosen in an attempt to detect coherent pedestrian features as depicted in figure 5.

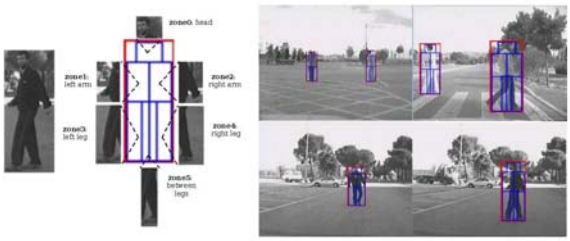

Figure 5: Left: composition of a candidate region of interest into 6 sub-regions. Right: examples in a sequence of images.

A set of features must be extracted from each sub-region and fed to the classifier. These are expected to be invariant to local shifts of candidate region of interest caused by change of pose and articulation of the pedestrian's arms and limbs. Several features extractors have been proved: co-occurrence matrix over canny edge extraction and over 32 levels normalized image, normalized orientation histogram over the 128 levels normalized image, image gradient magnitudes and orientation and finally texture unit number (NTU). The co-occurrences matrices over canny edge extraction are computed by the accumulated addition of the 4 possible bits combinations on the image $(00, 01, 10, 11)$ and yield one 2×2 matrix for each direction. As we have choosed 4 orientations $(90°, 45°, 0°, 315°)$ we get 4 matrices per sub-region and therefore a 16-element vector. When we compute the co-occurrence over the normalized image instead dealing with the cany edge extraction, the image has been normalized to 32 levels so that the co-occurrence matrices are not too large. By doing this we get 4 32×32 matrices per sub-region which seems a much more reasonable size. The normalized orientation histogram adds the difference magnitude between pixels in the 4 orientations delivering, this way, 4 128-lengthed vectors per sub region. Image gradient magnitudes and orientation have been directly fed to the classifier and their size depends on the sub-region's one. Finally, NTU extracts the local texture information of a neighbourhood of pixels (Wang, 1990) and the vectors size also depends on the sub-region size.

The Support Vector Machines (SVM) classifier, proposed by (Vapnik, 1999) have yielded excellent results in various data classification tasks, including people detection (Papageorgiou and Poggio, 2000). 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. 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 an 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 majority criterion in order to provide the final classification result. Once here, each candidate classified as pedestrian is dynamically tracked by a Kalman filter which decreases the false negative rate.

4 RESULTS

The system was implemented on a Pentium IV at 2.4 Ghz running the Knoppix 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 pedestrian being tracked

Table 1: SVM classification results.

	Distributed SVM Classifier			Holistic SVM Classifier		
	detection rate	false positive rate	false negative rate	detection rate	false positive rate	false negative rate
Coocurrence over normalized image	0.7437	0	0.2563	0.7789	0	0.2211
Coocurrence over canny image	0.8643	0	0.1357	0.8593	0.0653	0.0754
Canny image	0.7940	0	0.2060	0.7236	0	0.2764
Magnitude orientation	0.7236	0	0.2764	0.7136	0	0.2864
Normalizad Orientation Histogram	0.9246	0	0.0754	0.8894	0.0402	0.0704
Texture Unit Number	0.8593	0	0.1407	0.7136	0	0.2864

and their position. Specifically the average rate have 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 a database containing 1000 samples of pedestrians and non-pedestrian in different situations. The number of pedestrians samples in the training sets was chosen to be similar to the number of non-pedestrian samples. These ones 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 (Nuevo, 2005), 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. We obtained different detection rates depending on the feature extractor as depicted in Table 1 in a test set containing 500 images. 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. Accordingly, its value is recomputed at each frame as a function of the outputs provided by the single-frame classifier and by a Kalman filter used for pedestrian tracking. Figure 6 shows an example of pedestrian detection and tracking.

5 CONCLUSION

We have developed a binocular multi-frame pedestrian classification system based on Support Vector Machines (SVM). 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 complete classifier can be regarded as a hierarchical one. The distributed approach has yielded, superior performance, over the same data set,compared to the holistic classifier version. The results achieved up to date with a set of 1000 samples are encouraging. Nevertheless they still need to be improved before being safely used as an assistance driving system onboard road vehicles in real traffic conditions. For this purpose, the content of the training sets will be largely increased by including new and more complex samples that will boost the classifier performance, in particular when dealing with difficult cases. We aim at enhancing the classifier ability to discriminate those cases by incorporating thousands of them in the database. In addition, the attention mechanism will be refined in order to provide more candidates around the original candidate region. This will reduce the number of candidate regions that only contain a part of a pedestrian, i.e., those cases in which the entire pedestrian is not completely visible in the candidate region due to a misdetection of the attention mechanism. Finally, a gait recognition process will be introduced in order to enhance the shape-based pedestrian detection algorithm.

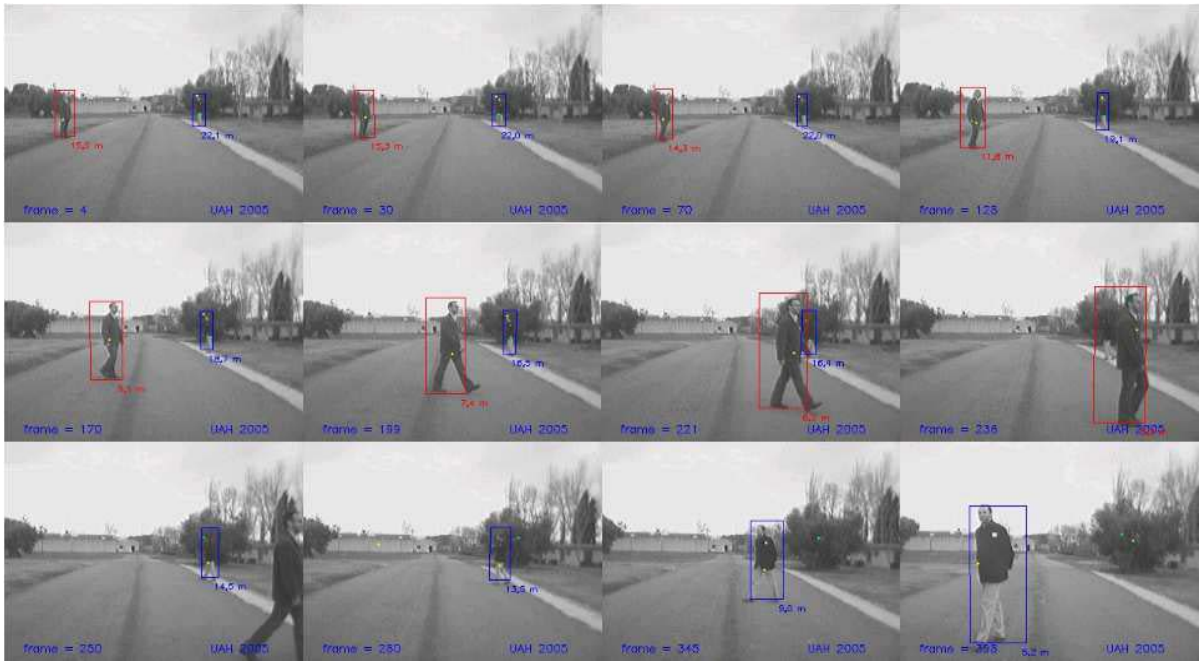

Figure 6: Pedestrian detection and tracking in a sequence of images.

ACKNOWLEDGMENT

This work has been funded by Research Projects CICYT DPI2002-04064-05-04 and FOM2002-002 (Ministerio de Fomento, Spain).

REFERENCES

Boufama, B. (1994). Reconstruction tridimensionnelle en vision par ordinateur: Cas des cameras non etalonnees. In *PhD thesis*. Institut National Polytechnique de Grenoble, France.

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

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

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

Grubb, G., Zelinsky, A., Nilsson, L., and Rilbe, M. (2004). 3d vision sensing for improved pedestrian safety. In *In Proc. IEEE Intelligent Vehicles Symposium*. pp. 19-24, Parma, Italy.

Labayrade, R., Royere, C., Gruyer, D., and Aubert (2003). Cooperative fusion for multi-obstacles detection with use of stereovision and laser scanner. In *International Conference on Advanced Robotics*. pp. 1538-1543.

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

Nuevo, J. (2005). Testbuilder tutorial. technical report 2005. ftp://www.depeca.uah.es/pub/vision/SVM/manual.pdf.

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

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

Sotelo, M. A., Nuevo, J., Bergasa, L. M., and Ocana, M. (2005). Road vehicle recognition in monocular images. In *submitted to ISIE 2005*. Duvrobnik, Croatia June 2005.

Vapnik, V. (1999). The nature of statistical learning theory. Springer Verlag.

Wang, L. (1990). Texture unit, texture spectrum and texture analysis. In *IEEE Transactions on Geosciences and Remote Sensing*. Vol. 28, No 4, pp. 509-512 (90-19).

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