Cooperation of Passive Vision Systems in Detection and Tracking of Pedestrians

David Lefée, Stéphane Mousset, Abdelaziz Bensrhair and Massimo Bertozzi

Abstract— **This work presents a cooperative approach for detecting and tracking pedestrians in an urban environment. Its originality lies in the cooperation of two vision systems. A monocular vision system retrieves feature elements and these elements are visualized. However, false detection can occur due to objects whose outline is similar to that of a pedestrian. This problem is solved by the introduction of an auto-adaptive stereovision algorithm that recovers all the vertical 3D segments of the scene. This cooperation supplies a fast and robust method for detecting pedestrian presence. Then, it allows pedestrian tracking through multiple images.**

I. INTRODUCTION

Several vision-based approaches are used for the detection of obstacles in urban environments. When only pedestrians are to be detected, a widely used approach is to look for specific patterns, such as movement features [1], shapes [2], or colours [3]. In these cases, the processing can be entirely based on monocular vision. Although the pedestrians can be detected in this configuration, the distance at which they are detected cannot be accurately computed without the help of other sensors, or without knowledge of the camera calibration or other outside information. Information about the depth is essential, both to eliminate the background scene, and to warn the driver about an imminent object located on the road. A system such as [4] shows that stereoscopic vision is reliable and useful for extracting and interpreting feature elements in general situations. Feature element extraction has been applied with success in intelligent real-time vehicles [5] [6]. The system presented in this work integrates the research work developed by the Dipartimento di Ingegneria dell'Informazione, Universitá di Parma (Italy) and the PSI laboratory, INSA of Rouen (France). The former uses a specific model for pedestrians. The results of the computation are fed to the second system which does not use a specific model. The two systems have been integrated into the GOLD (Generic Obstacle and Lane Detection) system.

This paper is organized as follows: section 2 introduces the monocular vision system, and section 3 describes the steps involved in building the 3D curves. Section 4 shows how the two systems cooperate. Some results are given in section 5. Section 6 ends the paper with some final remarks.

David Lefée, Stéphane Mousset, Abdelaziz Bensrhair are with the PSI laboratory, INSA of Rouen, Mont Saint Aignan 76131 France david.lefee@insa-rouen.fr

Massimo Bertozzi is with the Dipartimento di ingeneria dell'Informazione, Universit di Parma, I-43100 Italy bertozzi@ce.unipr.it

II. DETECTION BASED ON MONOCULAR VISION

The University of Parma has developed a monocular vision system for obstacle detection. Monocular images of the scene are acquired, then analysed by the GOLD system which is implemented on ARGO. ARGO is an experimental autonomous vehicle with automated driving capacity

A. Pedestrian detection

The goal of pedestrian detection is to locate those objects with components similar to a human shape. Due to camera movement, and changes in lighting, pedestrian detection is a non-trivial task. The detection algorithm is based on the following considerations [7]: localisation in a particular region of the scene, vertical edges, vertical symmetry axis, size and aspect specific to pedestrians. Given these assumptions, the localisation of pedestrians proceeds as follows: first an area of interest is identified on the basis of practical considerations. Inside this area, a Sobel mask is used to extract the outlines of objects. From the phases of the outlines, two graphs are obtained: the first contains vertical edges while the second has horizontal edges. Then, the background is eliminated. Since a second view of the scene is supplied by the stereoscopic vision system (section 3), it is used to compare the background of the two images. Then, one symmetry map is built from the grey level image, one from the vertical edges and another from the horizontal edge images [7]. A pedestrian has only vertical edges, so the horizontal edges are negligible. In order to reflect this, the horizontal edge map is combined with the others using a negative coefficient, to form a single map.

B. Construction of the bounding box

The monocular vision system is based on the following configuration:

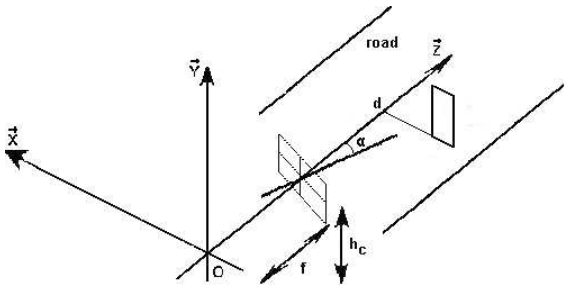

Fig. 1. *Configuration of the monocular vision system*

Where α is the angle camera of tilt in relation to the road, and h_c is the height of the camera. Due to the configuration adopted, the distance d is determined by:

$$d = \frac{h_c.(1 + tan(\alpha).\frac{p_y}{f}.(\frac{N-1}{2} - n))}{tan(\alpha) - \frac{P_y}{f}.(\frac{N-1}{2}) - n}$$

with f is the focal length of the lens, and p_y is the distance between two consecutive pixels in a column. N is the number of lines in an image, and n is the line image.

The objects with symmetrical and structural constraints are then located on the image; the size needs to be computed to detemine whether the object could represent a pedestrian. A bounding box is determined by finding the boundary of the object's side and base.

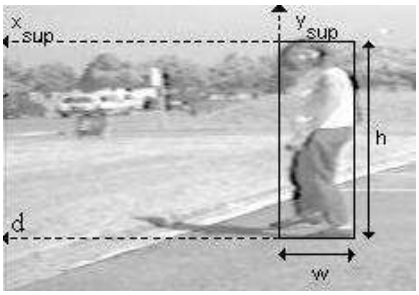

Fig. 2. *Parameters of the bounding box*

The left and right side of the bounding box are sought separately from the vertical edge map. In order to limit the size of the search, a proportional inverse weighting is given to each column, according to the position of the symmetry axis. Following the defined model of a pedestrian, the sides are represented by the arms. The system is calibrated, so the base is also sought from the vertical edge card [7]. The position of the bounding box boundary corresponds to the distance of the objects, and is given by *the equation (1)* . The summit of the bounding box is determined by the general shape of the bounding box.

Finally, amongst candidate pedestrians, only those with a shape corresponding to a human shape are kept.

III. The Stereoscopic Vision System

The PSI laboratory, INSA of Rouen, has designed a passive stereovision sensor in order to build the 3D curves of the environment, and compute motion according to the optical axes of the cameras.

A. Configuration

The sensor is made up of a rigid body, two identical lenses and two Philip *VMC* 3405 camera modules whose optical centres are 127 mm apart. A *Pict-Port Stereo H4S* Letron Vision frame grabber controls these two cameras, and acquires both images (728 x 568) simultaneously. The clock on the frame grabber AD-converter is the pixel clock of one of the two cameras. It is a timing signal used to divide the incoming lines of the video signals into pixels. With such a clock, maximum resolution can be reached and alias effects are avoided. Futhermore, the two camera lens units are set up so that their optical axes are parallel and, in order to respect an epipolar constraint, the straight line joining the two optical centres is parallel to each horizontal line in the images.

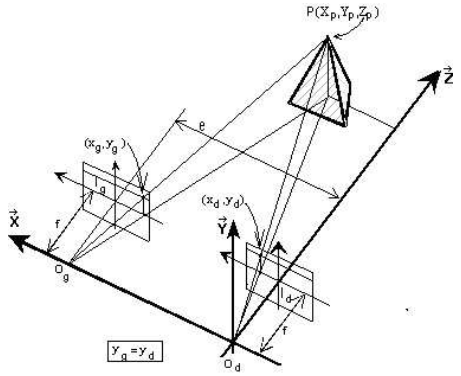

Fig. 3. *Stereoscopic vision system modeling*

Let P_L and P_R be two stereo-corresponding points of a $3D$ point P of an object (*Fig.3*). Let $(X_L Y_L)$, $(X_R Y_R)$ and (XYZ) be their coordinates. $(X_L Y_L)$ and $(X_R Y_R)$ are given in pixels, (XYZ) is given in meters. Then, due to the epipolar configuration,$Y_L = Y_R$ and $disp = (X_R - X_L)$. Based on this configuration, depth information is given in meters by:

$$Z = \frac{fe}{p_x disp}$$

where e is the distance between the two optical centers, p_x is the distance between two consecutive pixels in a line, and its value is 0.065 mm. The focal length of the two lenses, f, equals 16 mm. The horizontal disparity of two stereo-corresponding points is given by $disp$ in pixels.

B. Processing of 3D curves

Stereovision needs three-step processing: right and left image feature extraction, a feature matching step, then reconstruction of the depth. From the depth map processing, a fast and robust algorithm is introduced to compute the vertical 3D curves of a real scene [8]. 3D curves are defined as a sequence of connected 3D points. Like the 3D points, a 3D curve is defined by its two stereo curves, and constructed from coordinates X_{Ri} of declivities in the right image. The uncertainties about the building allow us to define a 3D curve as an element of a volume with the following dimensions: 3 pixels in width and 2 pixels in depth. Then, from the *a priori* knowledge of the environment, the small curves are eliminated and the others are treated as a 3D segment.

Here on *Fig. 4* is, an evaluation of 3D curve processing.

Pair of images	525
Mean number of 3D curves	45,6
Mean processing time	600 ms
Maximal depth for extracting a pedestrian curve	40 m

Fig. 4. *Evaluation of 3D curve construction*

On it, there is the number of stereoscopic image pairs which have been analysed, the number of relevant 3D curves for each pair of images, the mean processing time for each pair of images and the maximal depth of extraction of 3D pedestrian curves.

This algorithm has been designed to detect all the obstacles of the scene [9], and has been succesfully tested for pedestrian detection.

IV. Cooperation between the vision systems

In this section, an extension of the pedestrian detection module, using the two previous vision systems, is presented. On the one hand, the cooperation between these processes eliminates a lot of false detection by adding a new primitive: vertical 3D curves [8]. On the other hand, it allows the tracking of pedestrians through sequences of images. We use the GOLD system pedestrian data, as well as the 3D data representing obstacles and supplied by the stereoscopic vision system.

A. Pedestrian detection

A priori knowledge found by the GOLD system such as the size, the vertical edges and the vertical symmetry axis, are computed from the right image of the stereoscopic vision system. As a result, a bounding box which characterizes a pedestrian is constructed. Its parameters are the height h, the width w, the upper left coordinates x_{sup} and y_{sup} and the depth d (*Fig.2*). However, these primitives are not exclusively pedestrian features. So, the algorithms supply a lot of false detections. But, as these primitives are necessary but not sufficient, it is indispensable to add another primitive to limit the detections: the 3D curves. The 3D curves have been designed from a structured environment, but they are also used for pedestrian detection. On the whole, a pedestrian has only vertical edges; the horizontal edges are negligible. And, due to the uncertainties of the depth measurements, the generally vertical 3D curves of the pedestrian can be modelled solely by vertical curves. Without the uncertainties, this would not be possible.

The bounding box positions are compared with the 3D curve positions on the reference image. The process consists in the validation of the bounding box, by checking the 3D curves inside it [10]. If one of these criteria is not met, the bounding box is eliminated. So, optimal detection is achieved both by the pedestrian detection and by the presence of 3D curves.

A lot of false detection is thus eliminated; on a test sequence of 80 images, the detection for object existing in the scene filmed was 77%. 8% of these detected objects were non-pedestrians.

B. Estimation of the pedestrian position

In almost all cases, the bottom of the bounding box is not correctly positioned on the human shape: there is false computation of the depth of the object by the GOLD system. So, the stereoscopic vision system is used to retrieve the depth value [11]. Since the layout shape is Gaussian, the extreme values are rejected if they are not in a virtual 3D bounding box, defined by:

$$P_n = p/p \in R_n; \overline{disp} - \alpha.\sigma \leq disp(p) \leq \overline{disp} + \alpha.\sigma$$

where R_n is the set of 3D outline points, defined from the projection of the bounding box number n. $\overline{disp}$ is the mean of the points included in R_n, and σ is their standard deviation. α is a discrimination parameter.

The 3D points that are not eliminated are elements of the 3D bounding box. Then, the minimum depth is obtained by:

$$disp_{ref} = min(disp(p), p \in p_n)$$

The minimum depth of 3D curves is the most accurate value and it represents the reference distance.

Then, *equation 1* and *equation 2* supply the line position n as a function of the *refence distance*, and give :

$$n = \frac{h_c + h_c.tan(\alpha).\frac{p_y}{f}.\frac{(N-1)}{2} - \frac{f.e}{p.disp_{ref}}.(tan(\alpha) - \frac{p_y}{f}.\frac{(N-1)}{2})}{\frac{p_y}{f}.\frac{f.e}{p_x.disp_{ref}} + \frac{p_y}{f}.h_c.tan(\alpha)}$$

with $disp_{ref}$ the reference distance. This relation retrieves the pedestrian's position in the scene.

The depth information allows pedestrian position to be located in a two-dimensional graph (see *Fig.5*)

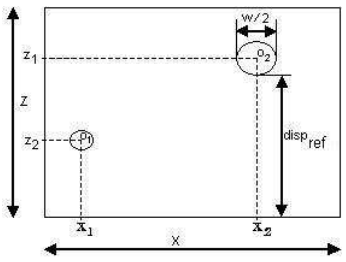

Fig. 5. *Bird's eye view*

in which the vertical axis represents the depth, and the horizontal axis represents the transversal position of pedestrians, in relation to the two cameras. Then, a pedestrian is characterized by a circle, with the coordinates:

$$O_c = (x, disp_{ref} + \frac{w}{2})$$

with x, the medium position and w the size of the bounding box; they are supplied by the monocular vision system. $disp_{ref}$ is the disparity reference and given by the stereoscopic vision system.

This two-dimensional graph follows the pedestrians as they move.

C. Temporal matching of circles

One image supplies a lot of information about the presence of pedestrians. However, some problems remain:

- multi-answers for a single pedestrian
- false detection
- non-detection

The robust solution proposed, links two temporally shifted images of the scene. The circles which have been located in the two images are matched *via* some circle attributes.

1) Selection of circle attributes: With the attributes a distance between the circles, which is not the single physical distance between them, can be computed.

In order to match the circles temporally, some *a priori* attributes have been defined. The most constant attributes of the acquired image at two different moments are selected. The attributes kept are:

- the position x and the size of the bounding box. They are supplied from the monocular vision system
- the mean of the disparities and the mean of the grey levels. They are supplied by the stereoscopic vision system.

On a test sequence of 50 images, the shape of the attributes is represented on *Fig*-6.

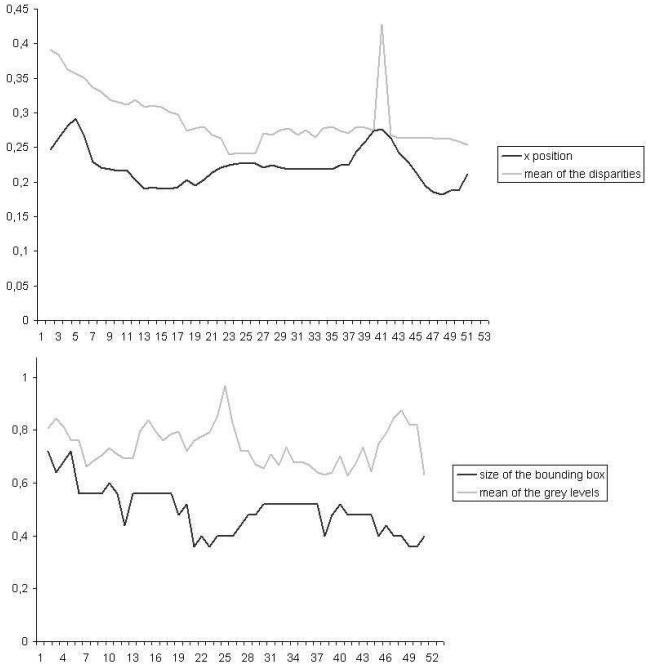

Fig. 6. *Attributes shape*

The first graph follows the pedestrians as they move (x and z). The second graph represents geometric features.

The next step concerns the computation of the distance. The local information is defined:

$$d_a = \sqrt[2]{\sum_{i=0}^{n_a-1} \Delta_i^2}$$

with Δ_i, the normalized attributes and defined by $\Delta_i = \frac{\Delta_i}{max\Delta_i}$, in which $\max(\Delta_i)$ is the maximum value of Δ_i. n_a is the whole of attributes.

The distances are compared for each pair of circles. Thus, they represent the similarity levels of the attributes.

2) Nearest neighbour: To improve the general matching, a global information criterion is applied [12]; a quality value is introduced to measure the similarity between the attributes of the t -1 and the t images. The quality value is computed to minimize the global distance of the circle attributes: the more the global distance tends to zero, the better the matching.

t+1 / t	N1	N2	N3	...	N
N1	d11	d12	d13	...	d1n
N2	d21	d22	d23	...	d2n
N3	d31	d32	d33	...	d3n
...	...	...	...	...	...
N	dn1	dn2	dn3	...	dnn

Fig. 7. *Global matching*

The number of circles at the time t is represented on the horizontal axis, and the number of circles at the time t -1 is represented on the vertical axis. d_{ij} is the distance between each pair of circles. The matching proceeds as follows:

- the minimum distance is sought for each column
- it is compared to a threshold:
- below the threshold, there is matching between the i line circle and the j column circle
- and above this threshold, there is no matching and the circle label is tagged; a penalty is introduced for the non-matching circles
- the similarity relations between the two circle images are examined. Thus, due to the unicity constraint we have imposed, the corresponding column and line are deleted
- the matching is carried out for all the lines of the chart
- the non matching circles are also tagged

Finally, a chart with only the minimum distance is obtained. This algorithm retrieves the best matching of the circles, and the non-matching circles are kept.

V. Experimental results

We have tested our algorithms on a large set of real images. Here on *Fig.8* are some results concerning the localisation and the tracking of pedestrians.

The images represent the right images of a complex outside scene. A lot of elements, such as cars, pedestrians, road signs and road markings, appeared both in the background, and in the foreground of the images. Moreover, these elements are located in shadowy areas. The images have been acquired at 256 x 288 x 8 bits.

The ($a1$) image has been acquired at time t -2, and represents the reference image, ($a2$) at time t -1, and ($a3$) at time t. The results of the pedestrian localisation are shown on ($b1$). The results of the pedestrian tracking are shown on ($b2$) at time t -1 and, and ($b3$) at time t.

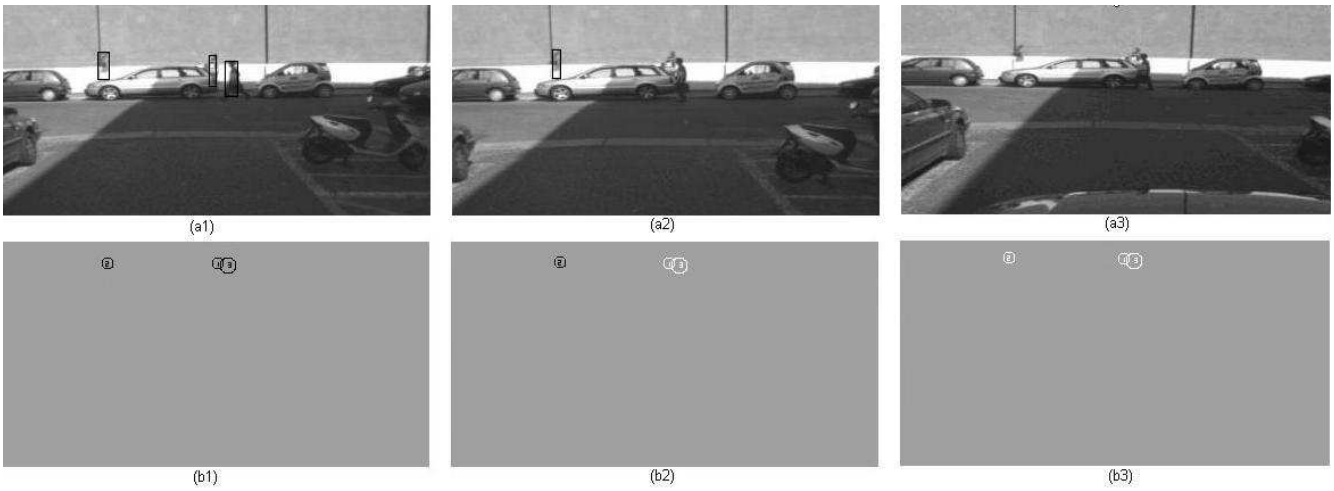

Fig. 8. *Global matching*

The pedestrians detected by the monocular vision system are represented by a black bounding box on ($a1$) and ($a2$). On ($a2$), a non-pedestrian is detected and ($a3$), no objects are detected.

The final result of the cooperation locates pedestrians and represents them by circles. A black circle corresponds to a correct matching between the ($b2$),($b3$) images and the reference image. In this case, there is a strong confirmation of the pedestrian presence. When no match is found, the circle is tagged. The pedestrian is located with a white circle. Then, a label has been applied on each circle.

In this way, the pedestrians that have been partially detected in the image sequences are vizualised thanks to a "memory effect" between the images.

VI. CONCLUSIONS AND FUTURE WORK

A. Conclusions

This work presents a vision-based system for detecting and tracking pedestrians in urban environments. The procedure is carried out through the cooperation between two systems using the GOLD data and the sterescopic system. This cooperation eliminates false detection by adding 3D information about pedestrians.

The pedestrian depth is computed in order to locate the pedestrian's position accurately, and is used to follow the pedestrian as he moves. The efficiency of this cooperation is shown in the results we have obtained on real scenes; the tests showed the system to be reliable and robust with respect to noise caused by shadows, or varying lighting conditions.

However, some problems remain. The final selection of pedestrians is too strict, so a few pedestrians are eliminated. Due to stochastic processes such as noise, or objects with vertical 3D curves, some false detections cannot be eliminated.

B. Future Work

To complete our methods, we want to improve the algorithms.

We have now choosen to use the Kalman Filter to estimate position. In fact, this filter can model moving objects. It could also be interesting to give us more accurate measurements, and from one estimated position at time t -1, to estimate the new position.

VII. ACKNOWLEDGMENTS

This work is the result of collaboration between the *dipartimento di ingeneria dell'Informazione of the University of Parma*, and the PSI laboratory, INSA of Rouen.

REFERENCES

[1] R. Cutler and L. Davis, "Robust real-time periodic motion detection, analysis, and applications," *IEEE Trans. On Pattern Analysis and Machine applications*, Vol 22, No 8, 2000, pp. 781-796.

[2] S.M. SMITH, "ASSET-2: Real-time motion segmentation and object tracking," *Real Time Imaging Journal*, Vol 40, 1998, pp.21-40.

[3] D. Harville, T. Darrelle, G. Gordon and J. Woodfill, "Integrated person tracking using stereo, color and pattern detection," *IEEE Conf. On Computer Vision and pattern Recognition*, 1998, pp. 601-608.

[4] D. Harwood, I. Haritauglu and L. Davis, "w^4s: A real-time system for detecting and tracking people in 2,5D," *European Conf. On Computer Vision*, 1998, pp. 877-892.

[5] G. Conte, M. Bertozzi, A. Broggi and A. Fascioli, "Vision-based automated vehicle guidance: the experience of the Argo vehicle," *Tecniche di Intelligenza Artificiale e Pattern Recognition per la Visione Artificiale*, 1998, pp. 35-40.

[6] D.M. Gavrilla, “Multi feature hierarchical template matching using distance transforms,” *IEEE Conf. On Pattern Recognition*, 1998.
[7] M. Bertozzi, A. Broggi, M. Sechi and A. Fascioli, “Shape-based pedestrian detection,” *IEEE Conf. On Intelligent Vehicles*, 2002.
[8] A. Broggi, A. Fascioli, S. Mousset, A. Bensrhair, M. Bertozzi and G. Toulminet, “Stereo vision-based feature extraction for vehicle detection,” *IEEE Conf. On Intelligent Vehicles*, 2002.
[9] G. Toulminet, S. Mousset and A. Bensrhair, “Fast and Accurate Stereo Vision-Based Estimation of 3D position and Axial Motion of Road Obstacles,” *IJIG*, Vol 4, 2004, pp.1-27.
[10] D. Lefée, S. Mousset, A. Bensrhair, “Approche coopérative entre systèmes monoculaire et stéréoscopique pour la détection de piétons en temps réel,” *IX Journes francophones des jeunes chercheurs en vision par ordinateur*, 2003, pp 251-260.
[11] D. Lefée, S. Mousset, A. Bensrhair and A. Fascioli, “Détection et suivi de piéton par systémes de vision passifs dans un environnement urbain,” *Transportation Innovation and Lane Detection* , Vol 2, 2003, pp.469-477.
[12] Z. Zhang, “Le problème de la mise en correspondance: L état de l art,” *INRIA*,No 2143,1993.