3D-Visual Detection of Multiple Objects and Structural Features in Complex and Dynamic Indoor Environments

Marta Marrón, Miguel Ángel Sotelo, Juan Carlos García, David Fernández, Ignacio Parra
Electronics Department. University of Alcala
Escuela Politécnica
28871, Alcalá de Henares (Madrid)
SPAIN
[marta,sotelo,jcarlos,llorca,parra]@depeca.uah.es

***Abstract* – In this paper, it is presented an algorithm for processing visual data to obtain relevant information that will be afterwards used to track the different moving objects in complex indoor environments. In autonomous robots applications, visual detection of the obstacles in a dynamic environment from a mobile platform is a complicated task. The robustness of this process is fundamental in tracking and navigation reliability for autonomous robots. The solution exposed in the document is based on a stereo-vision system; so that 3D information related to each object position in the local environment of the robot is extracted directly form the cameras. In the proposed application, all objects, both dynamic and static, in the local environment of the robot but the structure of the environment itself are considered to be obstacles. With this specification a distinction between building elements (ceiling, walls, columns and so on) and the rest of items in the robot surroundings is needed. Therefore, a classification has to be developed altogether with the detection task. On the other hand, the obtained data can be used to implement a partial reconstruction of the environmental structure that surrounds the robot. All these algorithms explained in detail in the following paragraphs and visual results are also included at the end of the paper.**

I. INTRODUCTION

Visual tracking is one of the areas of greatest interest in robotics, as it is related with many topics such as visual surveillance or mobile robots navigation. Multiple approaches to this problem have been, therefore, developed by the researching community during the last decades. Among all these solutions an interesting classification can be done according to the method used to detect or extract information about the objects in the scene from the image:

1) If a static camera is used: In this situation, background subtraction is generally applied to extract the image information that corresponds to dynamic objects in the scene. This method is very spread ([3], [4], [1], [2]) among the works developed by the community research, mainly in surveillance applications.

2) If the specific model of the object to be tracked is known: This situation is very common in tracking applications, both using static cameras ([4], [1], [2]) or dynamic ones ([5], [6]). If looking for a concrete shape, colour or texture in the image, the detection process is computational more expensive, but the number of false alarms and the robustness of the detector are bigger than in the case of looking for any kind of objects in the scene.

All the works referred before solve the detection problem quite easily thanks to the application of the mentioned restrictions. In the work presented in this paper none of these specifications are completed.

The solution is then, more complicated:

1) Background subtraction cannot be used, as its visual appearance changes continuously.

2) Any element in the visual environment of the robot may be an obstacle, apart from the objects that belong to the building structure in which the robot is located.

In this situation, it seems to be necessary to develop a more generic classifier that organizes visual data included in the images in two clusters:

a) Measurements coming from obstacles.

b) Measurements coming from the environment.

Once the information is classified, data assigned to cluster a) can be used as an input in any of the tracking algorithms proposed by the scientific community, such as the one designed by the authors and presented in [8] and [7].

At the same time, data classified in the environment cluster can be used to do a reconstruction of the robot surrounding structure. This last process is especially interesting for the navigation system of the robot, as the partial reconstruction of the environment can be feedback to the navigation process in order to know the exact position of the mobile robot in a local moving task, or in order to build a local map of the path travelled by the robot.

Fig. 1 shows a functional description of the global tracking application described in previous paragraphs.

In the following points the ideas presented in this introduction will be detailed and some results of their final implementation will also be shown.

Some of the ideas presented in this paper have already been proven in a different application developed by the authors ([10]).

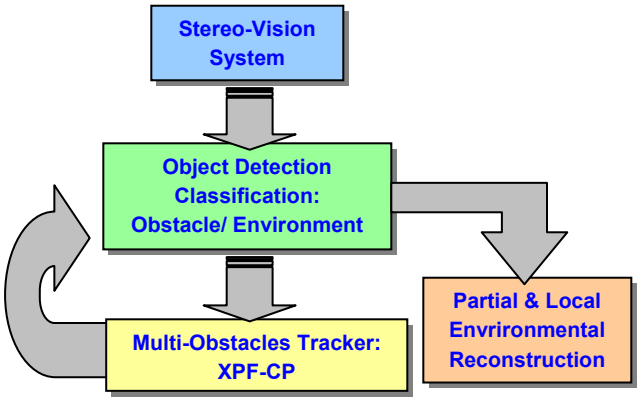

Fig. 1. General description of the global tracking application.

II. STEREO-VISION AND EPIPOLAR GEOMETRY

A stereo-vision process has been thought as the most adequate acquisition system for obtaining information about the dynamic environment at each sampling time. The main reasons of this decision are the following:

1) The amount of information that can be extracted from an image is much bigger than the one that can be obtained from any other kind of sensor, such as laser or ultrasound. This fact is especially interesting if probabilistic algorithms, like the one proposed by the authors of this paper in [8], are used for tracking objects with the obtained data. In fact, the first experiments with this type of trackers applied to the work described here were done by the authors in [9] with sonar data. As it is concluded in that document the algorithm lacks of robustness due to the poor amount of data extracted from the sonar system.

2) As the environmental configuration changes with time, the depth coordinate of the objects' position vector cannot be obtained with an only camera, and thus, a stereo-vision arrangement is needed. The acquisition system used is based on two synchronized statically arranged one next to the other. Analyzing left and right images captured by the sensors at the same time, a 3D position vector $\left[x_p, y_p, z_p\right]$ of every point in the scene (projected in $\left[u_l, v_l\right]$ and $\left[u_r, v_r\right]$ respectively in the left and in the right cameras) can be extracted through the epipolar geometry that relates the position of the two cameras.

Fig. 2 shows a functional description of the epipolar process, where the relevance of keeping constant and known extrinsic parameters ($\left[R_{lr}, T_{lr}\right]$) and intrinsic (mainly $\left[f_l, f_r\right]$) ones, through a calibration of the vision system, is denoted.

Both types of parameters are used to obtain the fundamental matrix (F_{lr}) that characterizes the stereo-vision arrangement in order to find matching points from the scene in the left and right frames.

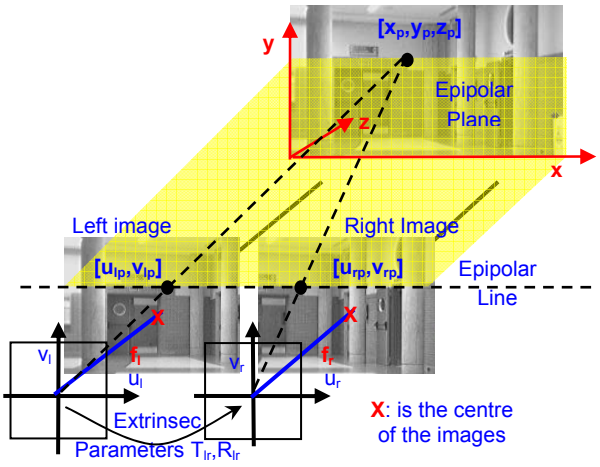

Fig. 2. Functional description of the stereo-vision data extraction process based on epipolar geometry.

The cameras platform has been calibrated and the intrinsic parameters for each camera, as well as the extrinsic parameters that relate them have been calculated in order to obtain the fundamental matrix (F_{lr}) that defines the system epipolar geometry. This way the perfect physically alignment between cameras is not necessary, because the calibration process defines mathematically the geometric relationships between the cameras [12].

The main problem of processing stereo-vision data, a part from the calibration one, is the high computational load of the epipolar matching for a dense number of points. Nevertheless, with recent hardware advances, real time stereo-vision processing has become possible in general purpose computers [11].

The stereo-vision system used in the experiments is formed by two black and white digital cameras synchronized with a Firewire connection and located in a static mounting arrangement, with a gap of 30cm between them, and at a height of around 1.5m from the floor.

III. THE DETECTION PROCESS

Different strategies have been tested in order to achieve the best results classifying the measurement set, obtained with the stereo-vision system presented before, both in execution time and in reliability.

The final solution chosen to solve the vision task previously exposed is described in this section. An alternative method is also presented after in this paper in order to make an efficiency comparison between them.

Fig. 3 shows the functional flowchart of the detection process finally designed by the authors with the specifications commented in the introduction:

A. The Canny Filter

The global classification process, shown in Fig. 3, is developed with each pair of frames (left and right) at each vision sampling time.

The left image is used to extract the pixels that may be interesting in the detection process, and therefore, those whose 3D position will be obtained through the right image and the epipolar geometry.

In order to reduce the execution time of the global detection process, these pixels are restricted to those extracted from the edges of the elements found in the left frame using a Canny filter.

The Canny edge detector [14] calculates the image gradient in order to highlight regions with high spatial derivatives. It is known as the optimal edge detector.

The Canny image provides a good representation of the discriminating features. Characteristics such as human heads, arms or legs, tables, doors, columns, etc. are visible and distinguishable also in quite crowded scenes and are easily extracted with the filter.

As it was exposed in the introduction of this paper, there is not a very concrete characteristic to look for in the images captured that specifically identifies an obstacle.

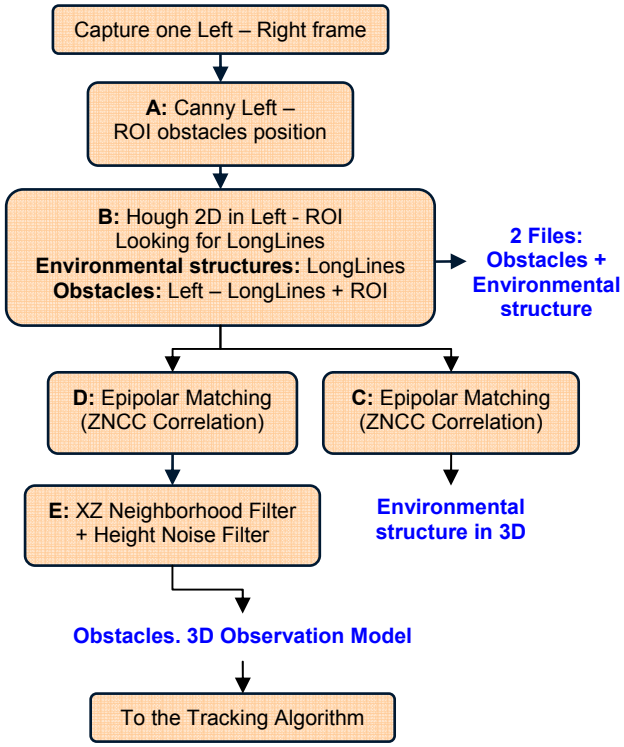

Fig. 3. Functional flowchart of the first detection and classification process designed.

On the other hand, edges corresponding with environmental structures have the common characteristic of forming long lines. Due to this reason, the detection process starts seeking structural shapes in the Canny image, as they can be found more easily in the Canny image.

As a preliminary process in order to robustly find structural features in the frame, the Canny left image is zeroed in the regions (ROIs) where an obstacle is expected to appear from the tracking algorithm results obtained in the previous time step (see Fig. 1).

B. Hough Transform

Hough transform is then used to search long lines or long line segments in the partial Canny image. These line segments will become in structural features if a matching point for each of its ends is found in the right frame with the epipolar geometry.

A probabilistic version of the Hough transform has been used in this application, to improve the process robustness.

C. Epipolar Correspondence of the Line Segments classified as Structural Features

To obtain its 3D coordinates, the position vector of the point in the left image is multiplied by the fundamental matrix (F_{lr}) of the stereo-vision arrangement to find the corresponding point in the right image along its epipolar line. In the searching process, the maximum distance between the corresponding points u coordinate ($|u_l - u_r|$) is limited in order to reduce the matching process computational cost.

The correspondence problem can be solved using a wide spectrum of matching techniques. Among them and recently the Zero Mean Normalized Cross Correlation (ZNCC) algorithm has performed most robustly [13] solving this task.

Fig. 4 shows the results obtained at the end of this step in an image extracted from a real experiment. In this image:

- Long line segments detected in the partial Canny image are shown in the upper image of the figure in different colours to distinguish them from the Canny pixels, drawn in write.
- Segments whose epipolar correspondence has been possible to find are drawn in the lower image. The segments are coloured in yellow and their ending points are coloured according to the depth of its position in the 3D space (red for the farther, green for the nearer). A green segment in the center-left part of the image is lost in the matching step, as it can be notice in the figure.

D. Epipolar Correspondence of the Points Classified as Obstacles

Long line segments are erased from the full Canny image firstly obtained from the left image, remaining in this partial edge left image the pixels of interest in the obstacles detection process.

At the same time, regions of interest (ROIs) from the initial Canny left image that were taken out to perform the structural features detection, are set back in the edge image, this time already free of objects classified as structural ones.

Epipolar matching is then developed with these pixels, in the way described in previous paragraphs, in order to obtain its 3D position in the scene.

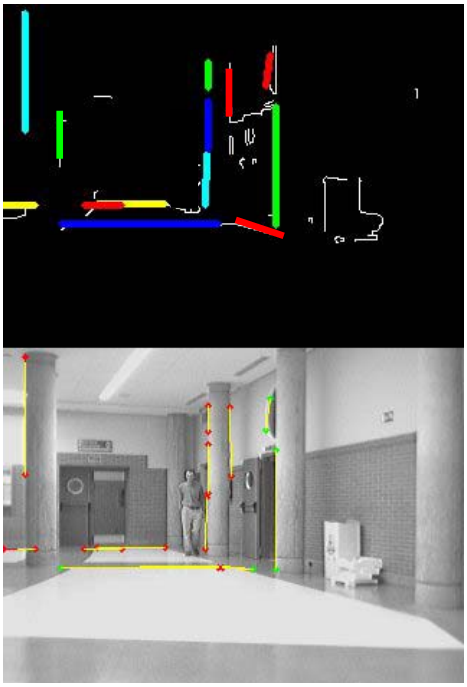

Fig. 4. Image obtained from a real experiment of the structural features detection process.

E. The Neighbourhood Filter

Despite the calibration of the stereo-vision arrangement, the correspondences between points in left and right images are often not correct due to occlusions and repetitive patterns and textures, generating outliers.

In order to reject these outliers, a neighbourhood filter is developed in the XZ plane to all points classified in the obstacles cluster 3D space. On the other hand, the height coordinate (Y) in each point 3D position vector is also used to filter the noise, increasing again the robustness of the detection process.

This way, a feasible set of points that characterizes the obstacles' position in the scene is obtained. These points will be used as the input observation model in the global tracking application (see Fig. 1 and Fig. 3).

IV. RESULTS

The global tracking algorithm based on the stereo-vision system described in this paper has been implemented on a mobile 4-wheeled platform. Different tests have been done in unstructured indoor environments, whose results are shown in this section.

The execution time of the developed classifier is around 60ms, which is an acceptable solution to implement a real time acquisition process of 15fps to 33fps with images of 320x240 pixels. This sampling time depends on the number of objects that exist in the scene and their position.

More specifically, the average rate of this period has a strong dependency on the number of points whose correspondence is searched, due to the ZNCC algorithm computational cost.

Some other algorithms have been tested in order to prove the efficiency of the detection and classification process presented in Fig. 3. The functional flowchart of one of them is shown in Fig. 5. The main difference between the functionality presented in that figure, and the one depicted in Fig. 3 is the order of the different steps in the process:

1) In the scheme shown in Fig. 5, the Canny image is not truncated to the frame areas where no predicted objects from the previous tracking loop are supposed.

2) Epipolar correlation is developed to all edge points in this global Canny image.

3) Hough transform is performed with the global set of edge points in the 3D space.

These 3 conditions do not change the final functionality of the detection and classification algorithm but degenerate its performance in different ways:

1) Not using the a-priori estimation of the objects' location in the scene makes more complicated to find structural elements of the environment in it, as edge points of both type of features are mixed in the initial Canny left image. The solution in Fig. 3 includes a preliminary classification step (A) that is not present in the one shown in Fig. 5. The first scheme is, therefore, more efficient than the second one.

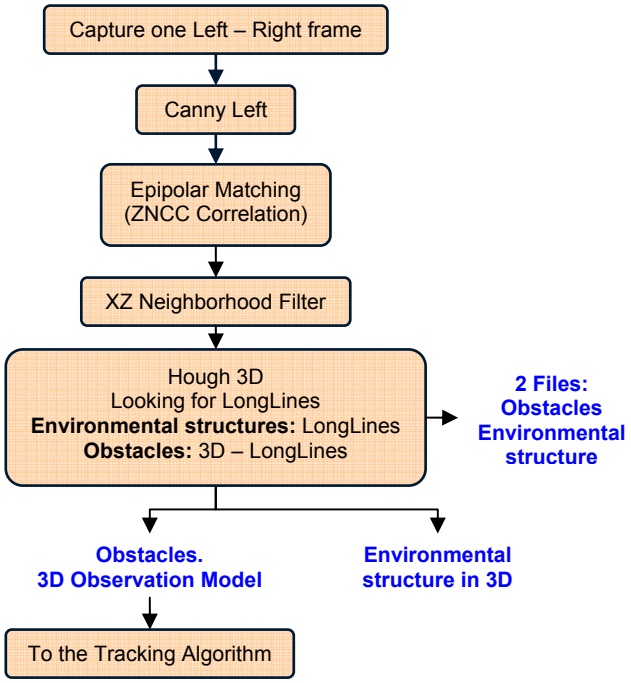

Fig. 5. Functional flowchart of a second detection and classification process tested.

2) Due to the difference expose in the previous paragraph, the correlation process has to be developed with a bigger number of points if using the algorithm in Fig. 5 (joining C and D steps) than if using the one in Fig. 3. The epipolar correspondence in the scheme shown in Fig. 3 is run not to every edge point in the image but only to the initial and final ones of every line segment proposed as a structural feature. Taking into account the load considerations related to the correlation process, this effect results in an increase of around 3 times in the execution time of the global detection and classification process.

3) As mentioned in previous paragraphs, the epipolar matching process usually generate false correspondences (also called outliers), resulting in erroneous 3D coordinate for the points in the image. Therefore, Hough transform gives worse results if it is run to the set of edge points in the 3D space, as it is in the algorithm in Fig. 5, than if it is executed in the 2D plane, as it is in the algorithm in Fig. 3. For this reason, the method described by Fig. 3 shows greater reliability in the classification process than the one in Fig. 5.

All vision processes have been developed using OpenCV libraries in independent sources, in order to achieve a modular software organization. Both algorithms have been developed in a PIV at 2.2GHz with 512Mbyes of RAM. In all cases the algorithm in Fig 3. gives better results than the one in Fig. 5. The results presented in the next figures are, therefore, obtained with the technique exposed in Fig. 5.

Furthermore, the global detection and classification algorithm have quite many parameters, that have been tuned empirically in case that they do not adjust automatically. A more complete description of the parameter tuning process is out of the aim of this paper.

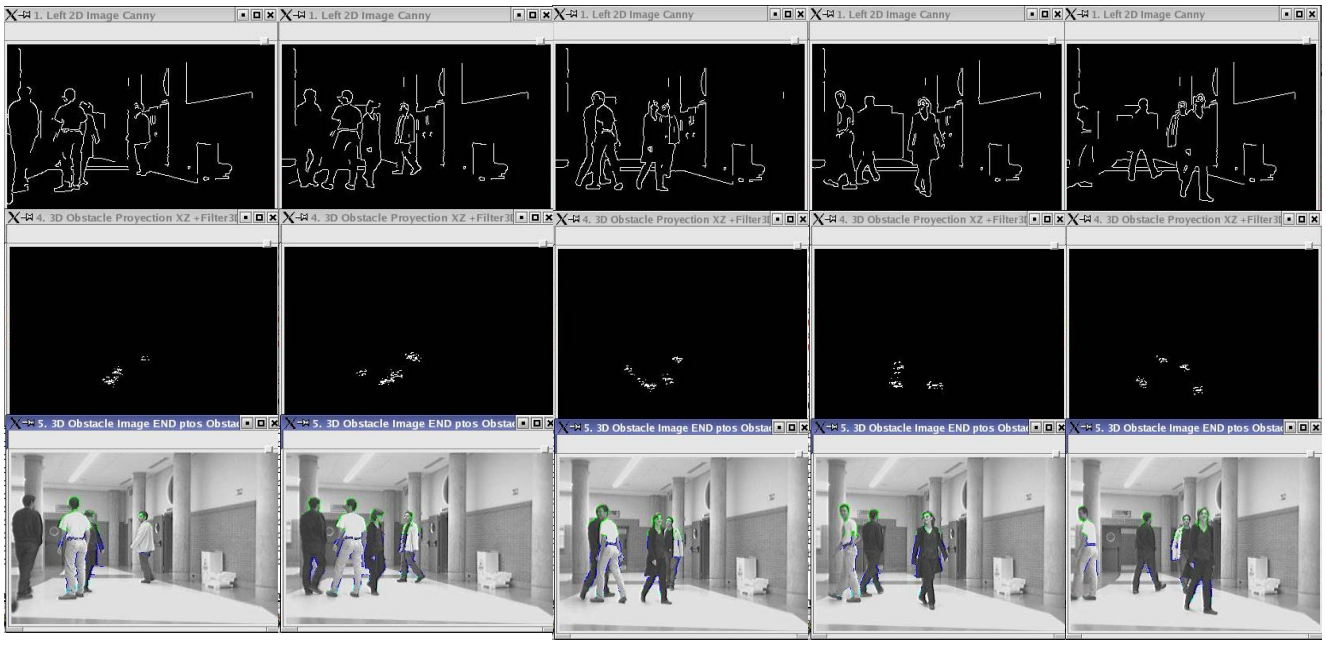

Fig. 6. Results of the classification algorithm in a real situation. Points classified in the obstacles cluster are shown in the images.

Fig. 6 shows the classifier output in one of the real time experiments, with 5 frames of a global sequence. Classification results for each one are displayed in 3 images organized vertically:

- The one on the top shows the edge image directly obtained applying the Canny filter to the left frame. Both obstacles and environmental structure borders are mixed in this image.
- The picture in the middle shows the result of the classification process in a XZ projection (x range is from -5m to 5m, and z range is from 0.2m to 20m) of the resulting 3D points. White dots represent the points classified in the obstacles cluster.
- The bottom one shows the final left image in which points classified in the obstacles cluster are highlighted with a colour according to the height where they are located in the 3D space.

As depicted in this figure, the obstacles in the resulting 3D space are represented by a uniformly distributed set of points. It can be concluded that the classification objective has been achieved, and that the resulting obstacle data set can be used in the tracker shown in Fig. 1.

Fig. 7 shows some other results of the detection and classification algorithm presented in this paper. In this figure, there are also 5 frames extracted from a real time experiment, in which the elements classified as structural features have been plotted:

- At the upper image, long line segments found by the detection process parallel to the robot walking plane are painted in green over the Canny left image.
- At the lower image, points included in the lines marked in the upper image, whose 3D position has been possible to obtain through the epipolar matching, are painted in green.

As it can be notice in both figures (6 and 7), the algorithm classifies in a very robust way features that are part of the building structure from those that are part of static or dynamic obstacles. The classification process is also robustly developed in crowded situations as the one shown in Fig. 6.

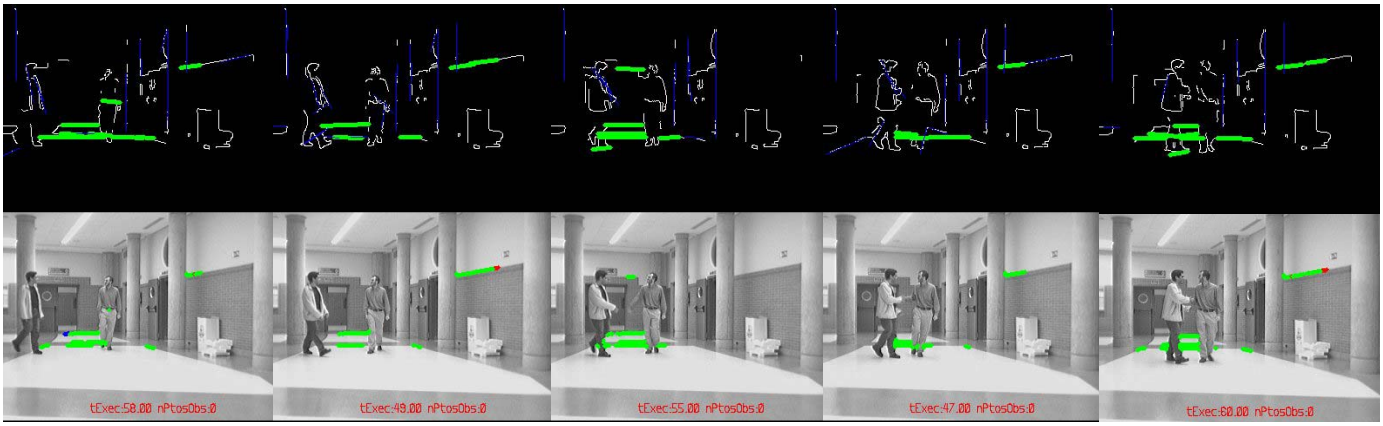

Fig. 7. Results of the classification algorithm in a real situation. Points classified in the structure cluster are shown in the images.

Neither the information related to the environmental structure nor the one related to the obstacles in it is processed by the application presented in this paper. A tracking algorithm as the one developed by the authors in [8], and a reconstruction algorithm still under development, are necessary to achieve the global objective presented in Fig. 1.

Though the frames presented in Fig. 6 and 7 have been acquired in a static position of the robot, the global tracking system is developed to be used in movement, as the feedback loop demonstrated in Fig. 1.

V. CONCLUSIONS

In this paper, an algorithm for detection and classification visual features in an indoor and complex environment is presented.

The detection and classification process is based on a synchronized stereo-vision system, and different vision techniques have been tested and validated in order to achieve the image process proposed.

The vision system has proven to be robust in different scenes and distances up to 20m.

As a future work, a pant-tilt unit can be added to the stereo-vision arrangement, in order to orientate the visual-field of the cameras to the most interesting area for the robot movement. There is an important problem related to this improvement: achieving an on-line calibration of a moving camera-arrangement is not an already solved question, and as it is mentioned in the paper, calibration is essential in order to obtain the 3D coordinate of any point in the scene through an epipolar matching process.

Results obtained with the proposed algorithm are shown in the figures included in the document, and prove that the objectives exposed have been achieved robustly and efficiently. The reliability of these results is especially important as they are thought to be used in tracking applications for robot autonomous navigation.

VI. ACKNOWLEDGMENTS

This work has been financed by the Spanish administration (CICYT: DPI2005-07980-C03-02).

VII. REFERENCES

[1] M. Isard, A. Blake. "Icondensation: Unifying low-level and high-level tracking in a stochastic framework", *Proceedings of the Fifth European Conference on Computer Vision (ECCV98),* Vol. 1, pp. 893-908, 1998.

[2] Y. Chen, T.S. Huang, Y. Rui. "Mode-based multi-hypothesis head tracking using parametric contours", *Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition (FGR02),* ISBN: 0-7695-1602-5, Washington, May 2002.

[3] Z. Khan, T. Balch, F. Dellaert. "A Rao-Blackwellized Particle Filter for Eigen Tracking", *Proceedings of the Third IEEE Conference on Computer Vision and Pattern Recognition (CVPR04),* pp. 980-986, Washington, June 2004.

[4] P. Torma, C. Szepesvári. "Sequential importance sampling for visual tracking reconsidered*", Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics,* ISBN: 0-9727358-0-1, Key West, January 2003.

[5] J.M. Odobez, D. Gatica-Perez. "Embedding motion model-based stochastic tracking", *Proceedings of the Seventeenth International Conference on Pattern Recognition (ICPR04),* Vol. 2, pp. 815-818, Cambridge, August 2004.

[6] K. Okuma, A. Taleghani, N. De Freitas, J.J. Little, D.G. Lowe. "A boosted particle filter: multi-target detection and tracking", *Proceedings of the Eighth European Conference on Computer Vision (ECCV04),* Lecture Notes in Computer Science, ISBN: 3-540-21984-6, Vol. 3021, Part I, pp. 28-39 Prague, May 2004.

[7] M. Marrón, J.C. García, M.A. Sotelo, E.J. Bueno. "Clustering methods for 3D vision data and its application in a probabilistic estimator for tracking multiple objects", *Proceedings of the Thirty-First Annual Conference of the IEEE Industrial Electronics Society (IECON05),* ISBN: 0-7803-9252-3, pp. 2017-2022, Raleigh, November 2005.

[8] M. Marrón, J.C. García, M.A. Sotelo, D. Fernandez, D. Pizarro. "XPFCP: An extended particle filter for tracking multiple and dynamic objects in complex environments", *Proceedings of the IEEE International Symposium on Industrial Electronics 2005 (ISIE05),* ISBN: 0-7803-8738-4. Vol. I-IV, pp. 1587-1593, Dubrovnik, June 2005.

[9] M. Marron, M.A. Sotelo, J.C. García. "Design and applications of an extended particle filter with a pre-clustering process, XPFCP", *Proceedings of the IEEE Conference on Mechatronics & Robotics 2004 (MECHROB04),* ISBN: 3-938153-50-X, Vol. 2/4, pp. 187-191, Aachen, September 2004.

[10] I. Parra, D. Fernández, M.A. Sotelo, P. Revenga, L.M. Bergasa, M. Ocaña, J. Nuevo, R. Flores. "Pedestrian recognition in road sequences", *Proceedings of the Fifth WSEAS International Conference on Signal Processing, Robotics and Automation (ISPRA06),* ISBN: 960-8457-41-6, pp: 273-278, Madrid, February 2006.

[11] D.M.Gavrila, V. Philomin. "Real-time object detection for smart vehicles", *Proceedings of the Seventh IEEE International Conference on Computer Vision (ICCV99)*, Vol. 1, pp. 87-93, Corfu, September 1999.

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

[13] B. Boufama. *Reconstruction tridimensionnelle en vision par ordinateur: Cas des cameras non etalonnees.* PhD Thesis, INP de Grenoble, 1994.

[14] F.J. Canny. "A computational approach to edge detection", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 8, n º 6, pp: 679–698, November, 1986.