People detection and stereoscopic analysis using MAS

S. Rodríguez*, O. Gil*, F. de la Prieta*, C. Zato*, J.M. Corchado*, P. Vega*, M. Francisco*
* University of Salamanca, Plaza de la Merced s/n, 37008 Salamanca, Spain
{srg, oscar.gil, fer, carol_zato, corchado, pvega, mfs}@usal.es

***Abstract*—This paper presents a multiagent system that can process stereoscopic images and detect people with a stereo camera. In the first of two phases, the system creates a model of the environment using a disparity map. It can be constructed in real time, even if there are moving objects present in the area (such as people passing by). In the second phase, the system is able to detect people by combining a series of novel techniques. A multi-agent system (MAS) is used to deal with the problem. The system is based on cooperative and distributed mechanisms and was tested under different conditions and environments.**

I. INTRODUCTION

For many years, stereoscopic vision has received a considerable amount of attention from a psychophysic perspective. More recently the scientific community has demonstrated an increasing interest in the study of artificial vision. Image processing applications are varied and include such aspects as remote control, the analysis of biomedical images, character recognition, virtual reality applications, and enhanced reality in collaborative systems, among others. Although image analysis and people detection is a well explored topic, the use of multiagent technology in this area has become the focal point of important interest [2][15]. The availability of commercial hardware to solve the lowlevel problems of stereo processing has turned them into an attractive sensor to develop intelligent systems. Stereo vision provides a type of information that offers several advantages in the development of human–machine applications. On one hand, the disparities information is less susceptible to illumination changes than the information provided by a single camera. Furthermore, the possibility of knowing the distance from the camera to the person is highly useful for locating and detecting of people and objects in a determined area.

This paper presents a system that is capable of processing stereoscopic images and detecting people with a stereo camera that is placed in an under-head position. In the first phase, the system creates a model of the environment using a disparity map. The model can be constructed in real time, even when there are moving objects present in the area (such as people passing by). For this reason, it is an appropriate tool for using on mobile devices (such as mobile robots). Our system analyzes and detects people by combining a series of novel techniques and raw images detectors such as Sum of Absolute Differences (SAD) or Gradient Orientation Histograms (HOG).

The remainder of this paper is structured as follows. Section 2 explains the basis of the background modelling and foreground techniques. Section 3 shows how the multiagent systems performs stereo processing and people detection. Section 4 presents the experiment carried out, while Section 5 presents the conclusions and suggests possible future work.

II. BACKGROUND

The primary concepts used in the development of this system include stereoscopic data handling, as well as all related analysis and detection processes, and multi-agent technology.

A. Stereoscopic analysis

Traditionally, the use of stereoscopy as a technique for reconstructing images has dealt with two problems. Using a two-dimensional pair of images with spatial coordinates (u,v), the left image (L) and right image (R), the correspondence problem attempts to find which two pixels mL(uL,vL) from the left image and mR(uR,vR) from the right image correspond to the same pixel M in three-dimensional space (X,Y,Z). Once these pixels have been found, the reconstruction problem attempts to find the coordinates for pixel M[9]. Addressing the correspondence problem is undoubtedly the most difficult task. Since there are generally several possibilities for choosing the element in image R that corresponds to the element in image L, the stereo correspondence problem is said to be ambiguous. It is necessary to determine which characteristics can be applied in order to reduce the ambiguity as much as possible.

The ultimate goal of reconstruction is to find the coordinates for pixel M (x,y,z) based on the coordinates from the projections for the same point over the images (uL,vL) and (uR,vR) [9].

Regarding the problem for obtaining correspondence, there are several strategies that can be classified in different ways [6]. The disparity calculation allows us to obtain the depth for each of the pixels on the image, obtaining one single image as the disparity map. Given that there is a direct correlation between the depth of the objects in an image and the disparity with a stereo pair, we can use the information from the disparity map as relative values for the depth of the objects. The techniques based on features obtain high quality primitives (edge, segments, curves, regions, etc.) that store a set of properties that remain unchanged with the projection [11]. Area based techniques consider the captured images to be a transferred two-dimensional signal. For each one of the pixels in the image, they try to make a transfer, minimizing certain criteria (correlation). One of the most simple techniques is the Sum of Absolute Differences

978-1-4244-7652-7/10/$26.00 ©2010 IEEE

Authorized licensed use limited to: UNIVERSIDAD DE SALAMANCA. Downloaded on July 08,2010 at 16:24:31 UTC from IEEE Xplore. Restrictions apply.

(SAD), since it operates exclusively with whole numbers. The bookstore that was used in the project (Triclops SDK [14]) establishes a correspondence between the images using this technique.

B. People detection

An object detector can be considered a combination of a set of image characteristics and a detection algorithm. Regarding the use of stereo vision to detect people, there are two approaches that are commonly referred to in written work: the first is to learn to recognize the types of image clusters similar to those that are common in a given object. This approach can be referred to, in general terms, as one based on parts. The other approach is the detection of the full body using descriptors based on Vector Support Machines (VSM) as a classifier [13]. The human form as been shown to be a difficult "object" to detect because of the significant variability in its appearance, clothes, and lighting conditions. The first thing that needs to be done is to detect a set of general characteristics inherent to the human form that can be readily identifiable, even under difficult circumstances with difficult lighting conditions [13] [10]. One of the first algorithms that was used for the detecting characteristics of a image and was helpful in recognizing object, was SIFT (Scale Invariant Feature Transform) [10]. The characteristics of SIFT are, based on the appearance of an object in specific points of interest, and are unaltered by rotation or a change of scale. However, they present certain difficulty in dealing with a large database, and have not shown the best results in recognizing people [10] [12]. The present study will use the Histogram of Oriented Gradients (HOG) [4]. The fundamental idea is that the appearance of objects and the shape of an image can be described by the distribution of gradient intensity or direction. The application of these descriptors can be achieved by dividing the image into small connected regions, called cells. A histogram of oriented gradients is compiled for every cell and the pixels contained within each one. The application of these histograms represents the descriptor [5] [4]. The HOG descriptor has several advantages over other descriptor methods. As Dalal and Triggs [5] observed, this descriptor maintain an almost vertical position. The HOG descriptor is then especially suited for detecting humans in images [5] [4].

C. Multi-agent systems

The use of deliberative BDI (Belief, Desire, Intention) agents [3][15] is essential in the development of the platform we are proposing. Apparently, the human visual system deals with a high level of specialization when it comes to classifying and processing the visual information that it receives, such as reconstructing an image by texture, shadow, depth, etc. Computationally, it is difficult to compete with such specialization and separate from an image only the relevant information for any particular purpose. In response to this problem, we propose implementing an algorithm over a distributed agent-based architecture that will allow visual information contained in an image to be processed in real time. Because the system is capable of generating knowledge and experience, the effort involved in programming multiple tasks will also be reduced since it would only be necessary to specify overall objectives, allowing the agents to cooperate and achieve the stated objectives.

D. Our approach

This paper presents a system able to analyze sequences of stereoscopic images and detect people. Our approach is built starting from several groups of agents whose properties and missions must be able to:

- Use a stereoscopic camera for inputting data in the system.
- Obtain images from the stereoscopic camera as well as from a physical storage space.
- Analyze the images obtained in order to calculate the distance between the camera and the elements found within the area represented by the image.
- Display the intermediate steps of the analysis. To do so, the images that correspond to these steps will be presented.
- Make a graphical representation of the depth and location of the elements found within the area.
- Detect human figures in images taken by a stereoscopic camera, using characteristic extraction algorithms.
- Relate each of these partial results in order to obtain a global response from the multi-agent system.

The present study is focused on these points. This study proposes a distributed collaborative approach that can deal with problems by using a multi-agents system (MAS) based on cooperative self-organization methods.

III. MAS BUILDING

The different process are implemented over a distributed agent-based architecture, which allows it to run tasks in parallel using each service as an independent processing unit. The architecture would allow a stereoscopic image processing system to carry out its own phases, which could be distributed among the agents. This way each of the tasks, including data gathering, proprocessing, filtering and reconstruction, as well as human form detection, could be carried out. In addition to the specialized agents, there is an agent platform for monitoring and supervising the correct functioning of the system. A description and initial proposal for this architecture can be found in [15]. The system is comprised of a set of agents with defined roles that share information and services. The analysis of images supposes a complex process where each agent executes its task with the information available at each moment.

Figure 1 shows the structural composition of the system in which the platform agents interact with each other and with the device sensors in order to carry out the objectives listed in the previous section. The data obtained from the stereoscopic camera are entered into the system and shared between the agents that will use specific services to process the data (filtering, preprocessing, disparity analysis, etc.) [15]. As shown in the figure, the system exits can be located in either the high-density disparity map that is obtained from the distance between the camera and the objects, the numerical representation of these distances, their three-dimensional representation in real time, and/or the detection of human forms in the specified area.

Authorized licensed use limited to: UNIVERSIDAD DE SALAMANCA. Downloaded on July 08,2010 at 16:24:31 UTC from IEEE Xplore. Restrictions apply.

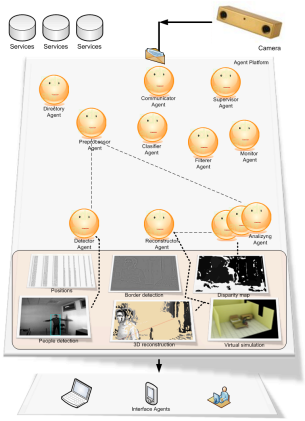

Figure 1. Structural composition of the system

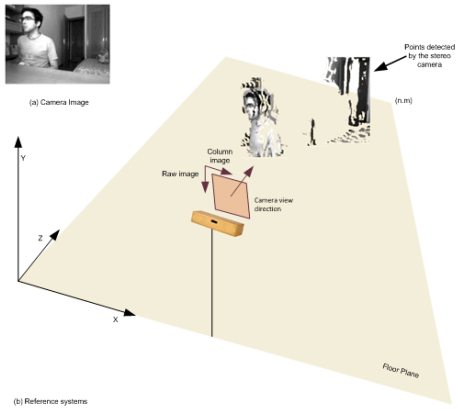

Figure 2. (a) Image of the right camera captured by stereo system. (b) Three-dimensional reconstruction of the scene showing the reference systems employed.

This section presents the core of stereo processing and detection processing. It is divided into two parts. The first part explains the basis of stereo calculation and how the 3D points captured by the stereo camera are translated to another reference system more appropriate for our purposes. The second part explains how the system is used to extract information about the location of the individuals.

A commercial stereo camera [15] was employed in this work because it can capture two images from slightly different positions (stereo pair) that are transferred to the computer to calculate a disparity image containing the points matched in both images. Knowing the extrinsic and intrinsic parameters of the stereo camera it is possible to reconstruct the three-dimensional position. The stereo calculation is made with the Triclops library[14], which defaults to a pairing algorithm based on the Sum of Absolute Differences (SAD). As it is the method used in the libraries provided by Point Grey [14], we chose to implement it, along with a proposal for optimizing the algorithm via the parallelization of the tasks by the algorithm [15].

SAD operates exclusively with whole numbers. Given a pixel with coordinates (x, y) in the left image, a correlation index C(x, y, s) is calculated for each displacement s for the correlation window in the right image. To calculate the correlation index, $C(x,y,s) = \sum_{u=-w,v=-w}^{u=w,v=w} | I_l(x+u,y+v) - I_r(x+u+s,y+v) |$ where 2w + 1 is the size of the window centered on the pixel located at position (x, y) and I_l, I_r are the gray values for the pixels in the left and right images respectively. The disparity $d_l(x, y)$ between the left and right image pixels is defined as displacement s which minimizes the correlation index: $d_l(x,y) = arg\ min_s\ C(x,y,s)$.

Fig. 2(a) shows an example of a scene captured with our stereo camera (the image corresponds to the right camera). Fig. 2(b) shows the three-dimensional reconstruction of the scene captured using the points detected by the stereo camera. The ‘‘world’’ and camera reference systems have been superimposed in Fig. 2(b), which shows that the number of points acquired by a stereo camera can be very high (they are usually referred to as point cloud). For that reason, many authors perform a reduction of the amount of information by orthogonally projecting them onto a 2D plan-view map [7].

The image analysis provides a point cloud in which each point represents a píxel in the image that indicates the position of the coordinates XYZ. The starting point of the coordinates used to represent the image is taken from the right-side reference point of the camera. The x-axis is horizontal, i.e., the axis that joins the camera’s two reference points. The y-axis is the vertical axis that follows the camera’s orientation. The z-axis measures the distance to the camera and is the axis that is perpendicular to the reference point.

People detection and stereo processing are treated as separate processes in this study. Every time a new image is captured, the system must first apply stereo processing to obtain the distances of the objects in the image. After that, the system can decide to apply the people detection to the same image. To achieve this goal, the HOG [4][5] algorithm, along with the classifiers used for training and validating the dataset, was applied, in addition to the test case that included the set of images taken with the stereoscopic camera.

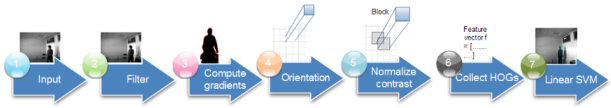

Figure 3. Detection phases

As explained in section 2, there are two fundamental issues regarding the detection of objects: the extraction of the most relevant characteristics, and the learning obtained from the classification [13]. Instead of using the raw image directly, it is common to use characteristics that are based on points, stains or gauss differences, intensities, gradients, color, textura, or a combination of these methods [8][1][16]. The developed system provides a set of services used specifically for extracting relevant characteristics [8][1][16], that comprise the base of the HOG descriptor and are used as the initial step of detection[15]. The software specifically uses the R-HOG block descriptor, which superimposes a square or rectangular block on the network cells. Each block is normalized independently. The primary phases carried out by the system during its detection process can be seen in Figure 3.

Authorized licensed use limited to: UNIVERSIDAD DE SALAMANCA. Downloaded on July 08,2010 at 16:24:31 UTC from IEEE Xplore. Restrictions apply.

• During the first and second phase, the data are input from the camera, and the filtering process required for reducing the lighting and shading effects are applied. The third stage computes first order image gradients. These capture contour, silhouette and some texture information, while providing further resistance to illumination variations. The system provides the most appropriate form detector, according to Sobel, Canny, etc. [16][1][15].

• The fourth stage aims to produce an encoding that is sensitive to local image content while remaining resistant to small changes in pose or appearance. The adopted method combines radient orientation information locally in the same way as the SIFT[10] feature. The image window is divided into small spatial regions, called "cells". For each cell we accumulate a local 1-D histogram of gradient or edge orientations over all the pixels in the cell. This combined cell-level 1-D histogram forms the basic "orientation histogram" representation. Each orientation histogram divides the gradient angle range into a fixed number of predetermined bins. The gradient magnitudes of the pixels in the cell are used to vote into the orientation histogram.

• The fifth stage computes normalization, which takes local groups of cells and contrast normalizes their overall responses before passing to next stage. Normalization introduces better invariance to illumination, shadowing, and edge contrast. It is performed by accumulating a measure of local histogram "energy" over local groups of cells that we call "blocks". The result is used to normalize each cell in the block. Typically each individual cell is shared between several blocks, but its normalizations are block dependent and thus different. As a result, the cell appears several times in the final output vector with different normalizations. There are two main blocks: rectangular R-HOG blocks and circular C-blocks. The R-HOG blocks are generally square and represented by three parameters: the number of cells per each block, the number of pixels per each cell, and the number of channels per cell histogram. The their experiment on human detection, Dalal and Triggs [4], determined the optimal parameters to be: 3×3 blocks of 6×6 pixel cells, with 10.4% miss rate. Four different block normalization schemes were evaluated for each of the above HOG [5]. If v ix the non-normalized vector that contains all of the histograms for one block, $\|v\|_K$ the k-norm for k=1, 2, and ε a small normalization constant to avoid division by zero. The four schemes are: (i) L2-norm, $v \leftarrow v/\sqrt{\|v\|_2^2+\varepsilon^2}$;(ii) L2-Hys, L2-norm followed by clipping (limiting the maximum values of **v** to 0.2) and renormalizing; (iii) L1-norm, $v \leftarrow v/(\|v\|_1+\varepsilon)$;(iv) L1-sqrt, $v \leftarrow v/\sqrt{v/(\|v\|_1+\varepsilon)}$. L1-norm followed by square root essentially treats the descriptor vectors as probability distributions, using the Bhattacharya distance between them.

• The next step collects the HOG descriptors from all blocks of a dense overlapping grid of blocks covering the detection window into a combined feature vector for use in the window classifier.

• The final step in recognizing forms using HOG is to feed the descriptors with some recognition system based on supervised learning. The SVM classifier is a binary classifier that looks for the optimalhyperplane as a decision making function. Once it has been trained with the images that are contained in a particular object, the SVM classifier can make decisions with regards to the presence of an object, such as a human being, in the testing images (which are different from the training images). To this end, a modified SVMLight packet from Dalal and Triggs [5] was used, obtaining the results shown in the following figures, in which the human form can be detected in different positions.

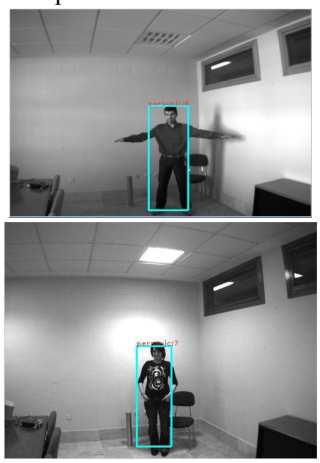

Figure 4. Results in the process of detection human forms

IV. EXPERIMENT AND RESULTS

In explaining the model we have provided examples of its performance. A broader experimentation was done to test processing and detection of different people under different lighting conditions and different distances from the stereo vision system. We employed 640x480 sized images and sub-pixel interpolation to enhance the precision in the stereo calculation. The operation frequency of our system is near 10 Hz on a 3.2 Ghz Pentium IV computer running with Windows XP. The camera has the following characteristics [14]: 640x480 pixel sensors, monochrome, 3.8mm focal distance, capable of capturing 48 photograms per second, 120mm line base, 6 pine IEEE-1394 (FireWire) interface connection. The images were taken from a height of 1.6m with a 6fps velocity, obtaining approximately 400M coded data in AVI and PGM format (16 bits per image). More than the fifty percent of the computing time is dedicated to image capturing and stereo computation (about 50 ms) and the rest to detection (about 40 ms). This indicates that the proposed system is fast enough to be used in real time applications.

In order to evaluate the system's capacity, a variety of tests were performed. The system improved the processing capability compared to other centralized systems, given that the distributed agents approach makes it possible to carry out processing tasks individually and with different techniques (selection of edge detectors, filters, final objectives: obtaining distance, detecting forms). Because the system is perfectly modulized, the tasks can be carried out simultaneously and or in a distributed manner.

For computing disparity, we used a stereo algorithm that allowed for real-time computation of dense disparity maps [15]. Standard stereo calibration and external calibration were applied the first time the system was installed in the environment. Then the system can work

Authorized licensed use limited to: UNIVERSIDAD DE SALAMANCA. Downloaded on July 08,2010 at 16:24:31 UTC from IEEE Xplore. Restrictions apply.

TABLE I.
ACCURACY RESULTS

Position	Distance	Angle	Avg. Err		Std. Dev.	
			Natural	Fluorescent	Natural	Fluorescent
P1	3.20 m	00º	62 mm	62 mm	74 mm	74 mm
P2	3.00 m	00º	60 mm	60 mm	58 mm	58 mm
P3	2.50 m	00º	54 mm	54 mm	28 mm	28 mm
P4	3.20 m	15 º	85 mm	85 mm	96 mm	96 mm
P5	3.00 m	22 º	81 mm	83 mm	82 mm	84 mm
P6	2.50 m	32 º	77 mm	83 mm	60 mm	64 mm
P7	3.20 m	-15 º	81 mm	81 mm	99 mm	99 mm
P8	3.00 m	-22 º	81 mm	83 mm	90 mm	92 mm
P9	2.50 m	-32 º	77 mm	83 mm	62 mm	66 mm

without any manual configuration as long as the camera settings (e.g., camera position in the environment, lenses, focal lengths) are not changed. The system has been tested in many situations and different conditions. Here we report a summary of different experimental results performed during the course of our research.

To measure the precision of the system we marked 9 positions in the environment at different distances, illumination and angles from the camera, and measured the distance returned by the system of a person standing at these positions. Although this error analysis is affected by imprecise positioning of the person on the markers, the results of our experiments, in Table 1 averaging 40 measurements for each position, show a precision in localization (i.e., average error) of less than 10 cm, but with a high standard deviation, which denotes the difficulty of obtaining a precise measurement. However, this precision is usually sufficient for many applications, like the one considered in domestic scenarios[15].

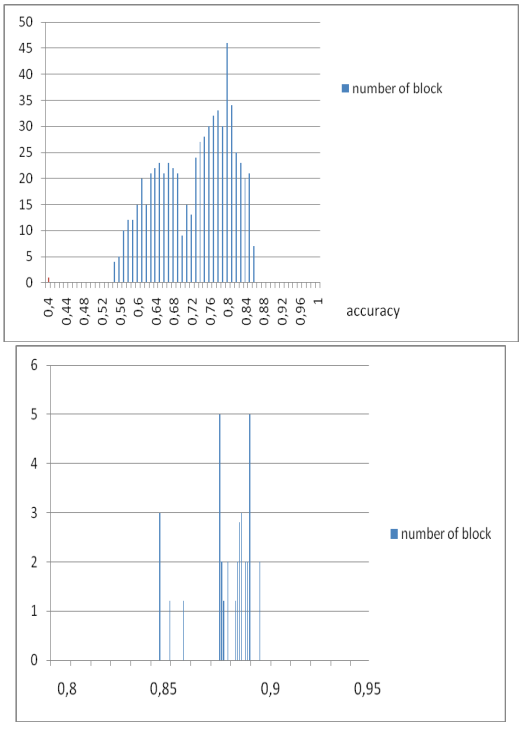

Figure 5. Distribution of accuracy of 16*16 block and Distribution of accuracy of 40*80 blocks

Finally, as our main contribution is the integration of the HOG feature into the multiagent platform, we evaluated the importance of using different sized blocks. Figure 5 shows two histograms. Each histogram shows the distribution of the classification accuracy (measured by the same error rate) of blocks using the linear SVM. The first histogram is of all blocks of size 16 × 16 pixels and the second histogram is of blocks of size 40 × 80 pixels. The figure clearly shows that the 16 × 16 blocks are the least informative and that increasing the block population does contribute significantly to the performance of our system.

V. CONCLUSIONS AND FUTURE WORK

This paper has presented a stereo processing system that integrates several capabilities into an effective and efficient multiagent platform: stereo image processing, distance calculation, real time graphical representation of depth, the identification of elements found within the area and human detection. Experimental results show good performance and high robustness to many problems: shadows, global illumination changes, background changes, people not moving, etc. Two major lines of investigation should be followed in the future in order to extend the applicability of the proposed system: outdoor environments and larger, more crowded areas. In the first case, it is necessary to consider more sophisticated techniques for background modelling and updates, such as multi-modal representation for the background and dynamic thresholding for background subtraction. In following the second line of investigation, we plan to include the system in a bigger architecture that could deal with large and crowded environments, which is probably beyond the scope of this method. This bigger architecture would control and autonomous mobile robot and be tested for its suitability for human–machine applications that require the ability to operate in real time and in changing conditions.

ACKNOWLEDGMENT

This research has been partially supported by the project PET2008_0036 and FEDER funds.

REFERENCES

[1] 1. Canny, J.:. A computational approach to edge detection. IEEE Trans Pattern Analysis and Machine Intelligence, 8 (6), pp. 679-698. (1986)

Authorized licensed use limited to: UNIVERSIDAD DE SALAMANCA. Downloaded on July 08,2010 at 16:24:31 UTC from IEEE Xplore. Restrictions apply.

[2] 2. Castanedo F., García J. , Patricio M.A. and Molina J.M., Designing a Visual Sensor Network Using a Multi-agent Architecture, ASC 978-3-642-00486-5, 430-439 (2009)

[3] 3. Corchado, J.M., Glez-Bedia, M., de Paz, Y., Bajo, J. and de Paz, J.F.: Replanning mechanism for deliberative agents in dynamic changing environments. Computational Intelligence 24 (2), pp. 77-107. (2008)

[4] 4. Dalal N and Triggs B. Histograms of Oriented Gradients for Human Detection, IEEE Conference Computer Vision and Pattern Recognition , USA, pg. 886 - 893, (2005)

[5] 5. Dalal N., Triggs B. and Schmid C.: Human detection using oriented histograms of flow and appearance.Book: European Conference on Computer Vision (ECCV), vol. II, pages 428-441 (2006)

[6] 6. Dhond U.R, Aggarwal J.K.: Structure From Stereo - A Review, IEEE Trans. on Systems, Man. and Cybernetics, Vol. 19, No.6, Nov/Dec (1989)

[7] 7. Harville M., Stereo person tracking with adaptive plan-view templates of height and occupancy statistics, Image and Vision Computing 2 127–142 (2004).

[8] 8. Lindeberg T.: Feature detection with automatic scale selection. International Journal of Computer Vision 30. Páginas 77-116 (1998)

[9] 9. López-Valles et al. :Revista Iberoamericana de Inteligencia Artificial. Vol. 9. No.27 , pp. 35-62. ISSN: 1137-3601. (2005)

[10] 10. Lowe, David G.:Object recognition from local scale-invariant features. Int. Conf. on Computer Vision. 2. pp. 1150–1157.doi: 10.1109/ICCCV.1999.790410 (1999)

[11] 11. Marr D. y Poggio T.: A computational theory of human stereo vision. Proc R Soc Lond, pp. 301-328 (1979)

[12] 12. Nazlı y Pınar Histogram of oriented rectangles: A new pose descriptor for human action recognition. Image and Vision Computing 27. 2009. Pág. 1515–1526. (2009)

[13] 13. Pedersoli M. , Gonzàlez J., Chakraborty B. and Villanueva J.J., Enhancing Real-Time Human Detection Based on Histograms of Oriented Gradients, Computer Recognition Systems 2, ASC, Springer Berlin, ISBN 978-3-540-75174-8, (2007)

[14] 14. Point Grey Research Inc. http://www.ptgrey.com/ (2009)

[15] 15. Rodríguez S., De Paz J.F., Bajo J., Tapia D.I., and Pérez B. Stereo-MAS: Multi-Agent System for Image Stereo Processing.IWANN'09. Ed. J. Cabestany et al.,LNAI, Springer Verlag. ISBN: 978-3-540-87655-7. Vol 5517. Pág. 1256–1263 (2009)

[16] 16. Sobel, I. & Feldman,G., A 3x3 Isotropic Gradient Operator for Image Processing, presentado en la conferencia Stanford Artificial Project (1968).

Authorized licensed use limited to: UNIVERSIDAD DE SALAMANCA. Downloaded on July 08,2010 at 16:24:31 UTC from IEEE Xplore. Restrictions apply.