Proceedings of the IEEE ICRA 2009
Workshop on People Detection and Tracking
Kobe, Japan, May 2009

A dioptric stereo system for robust real-time people tracking

Ester Martínez and Angel P. del Pobil
Robotic Intelligence Lab
Engineering and Computer Science Department
Jaume-I University
Castellón
Spain
Email: { emartine, pobil } @icc.uji.es

***Abstract*—We address and solve a number of problems in the context of a robot surveillance system based on a pair of dioptric (fisheye) cameras. These cameras provide a hemispherical field of view that covers the whole robot workspace, with some advantages over catadioptric systems, but there is little previous work about them. Then, we had to devise and implement a number of novel techniques to achieve robust tracking of moving objects in dynamic, unknown environments from color image sequences in real time. In particular, we present a new two-phase adaptive background model that exhibits a robust performance when there are unexpected changes in the scene such as sudden illumination changes, blinking of computer screens, shadows or changes induced by camera motion or sensor noise. The system is also capable of tracking the detected objects when they are not in movement. We also deal with fisheye camera calibration to estimate both intrinsic and extrinsic parameters, as well as the estimation of the distance between the system and the detected objects with our dioptric stereo system. Experimental results are reported.**

Robotics research, from its begining, has been always focused on building robots which help human beings in their daily tasks while both of them coexist in the same environment. That means new robot generations have to deal with dynamic, unknown environments, unlike industrial robots which act in a restricted, controlled, well-known environment. For that reason, one of the key issues in this context is to be aware of what is happening around.

In fact, robot performance in any real environment requires to detect people and/or other objects, particularly if they are moving, in the robot's workspace. On the one hand, interaction tasks require detection and identification of the objects with which to interact. On the other hand, the safety of all elements present in the robot workspace should be guaranteed at any time, specially when they are human beings. Thus, it is important that the robot quickly detects the presence of any moving element to be able to properly react to the element movements.

So, among the available robot sensors, cameras might be suitable for this goal, since they are an important source of information. Nevertheless, it is not straightforward to successfully deal with a non-constrained environment by using traditional cameras due to its limited field of view. That constrain could not be removed by combining several images captured by rotating a camera or strategically positioning a set of them, because it is necessary to establish any feature correspondence between many images at any time. This processing entails a high computational cost which makes them fail for real-time tasks.

An effective way is to combine mirrors with conventional imaging systems [1] [2] [3]. The obtained devices are called catadioptric systems. Moreover, if there is a single viewpoint, they are referred as central catadioptric systems [4]. This is a desired feature in such imaging systems since it describes world-image mapping. In fact, a single viewpoint implies that all rays go through a 3D point and its projection on the image plane goes through a single point in the 3D space. Conventional perspective cameras are devices of a single viewpoint, for example. Although the central catadioptric imaging can be highly advantageous, they unfortunately exhibit a dead area in the centre of the image what can be an important drawback in some applications.

With the aim of overcoming all the above drawbacks, a dioptric system was used. Dioptric systems, also called fisheye cameras, are systems which combine a fisheye lens with a conventional camera [4] [5]. Thus, a conventional lens is changed by one of these lenses which has a short focal length what allows cameras to see objects in an hemisphere. Although fisheye devices present several advantages in front of catadioptric sensors such as no presence of dead areas in the captured images, a unique model for this kind of cameras does not exist unlike central catadioptric ones [6].

In this work, we have focused on dioptric systems to implement a robot surveillance application for fast and robust tracking of moving objects in dynamic, unknown environments. Although our final goal is to design an autonomous, mobile manipulation robot system, here we present the first stage: novel techniques for robust tracking of moving objects in dynamic, unknown environments from color image sequences such that manipulation tasks could be safely performed in real time when the robot system is not moving. For that, three different related problems have been tackled:

- moving object detection
- object tracking
- distance estimation from the system to the detected objects

First of all, a new robust adaptive background model has been designed. It allows the system to adapt to different unexpected changes in the scene such as sudden illumination changes, blinking of computer screens, shadows or changes induced by camera motion or sensor noise. Then, tracking process from two omnidirectional images takes place. Finally, the estimation of the distance between the system and the detected objects must be done by using an additional method. In this case, information about the 3D localization of the detected objects with respect to the system was obtained from a dioptric stereo system.

Thus, the structure of this paper is as follows: the new robust adaptive background model is described in Section 2, while in Section 3 the tracking process is introduced. An epipolar geometry study of a dioptric stereo system is presented in Section 4. Some experimental results are presented in Section 5, and discussed in Section 6.

I. Moving Object Detection

Research in human and object detection has taken a number of forms. Well-known segmentation techniques from a taken image are thresholding or frame subtraction. However, on the one hand, it is difficult to deal with threshold selection when it is working with an unknown, dynamic environment and targets to track can have different features. Actually, the uncertainty provided by those specified work conditions also makes that automatic threshold search methods, mainly based on histogram properties, fail [7]. In that way, other experiments to obtain a robotic assistant in which a person is detected and then followed by a mobile robot [8] [9] [10] [11] have been carried out. Nevertheless, in spite of the fact that existing algorithms are very fast and easy to use, image processing for object identification is very poor since it is color- and/or face-based. This restricts their utility because it is not viable to track objects of a particular color, which has also to be significantly different from the background, or it constrains people to always face the vision system.

On the other hand, although the image difference method provides a good detection of changing regions in an image, it is important to pay attention to several uncontrolled changes in the system environment which can produce multiple false negatives and make the system fail. These dynamic, uncontrolled changes can be divided into:

- minor dynamic factors, such as, for example, blinking of computer screens, shadows, mirror images on the glass windows, curtain movement or waving trees, as well as changes induced by camera motion, sensor noise, non-uniform attenuation or atmospheric absorption, among other factors
- sudden changes in illumination such as switching on/off a light or opening/closing a window

Different research has been developed to adapt to this changes. One of the most common is the background subtraction approach, which has been proposed by several researchers [12] [13] [14] [15] [16]. Basically, a background model, which is built after observing the scene several seconds, is used to identify moving objects by thresholding the new frame with respect to the built background model. However, this approach presents two important drawbacks:

- everything observed when the background model is being built is considered background
- no sudden change in illumination occurs during the whole experiment

It is important to take into account that, unlike most of them which each background pixel is represented by a Gaussian distribution, Stauffer and Grimson [17] presented adaptive background mixture models. However, as it was pointed out in [18], some issues have to be solved.

Therefore, a novel algorithm is proposed here. It is divided into two different phases, as can be seen in Figs. 1 - 2:

1) In the first phase, an initial background model is obtained by observing the scene during several seconds. However, unlike most background estimation algorithms, another technique for controlling the activity within the robot workspace is performed. With the aim of reducing the computational and time cost, this control is performed by means of a simple difference technique. In that way, there is no danger to damage people who approach the robot while this initial model is being built. Thus, basically, in this phase, a simple frame-difference approach is performed in order to detect moving objects within the robot workspace. Then, two consecutive morphological operations are applied to erase isolated points or lines caused by the dynamic factors mentioned above. In this point, two different tasks are carried out:
    - On the one hand, adaptive background model is updated with the values of the pixels classified as background in order to adapt it to some small changes which do not represent targets
    - On the other hand, a tracking process, which is explained in the next sections, is performed
2) In the second phase, detection and identification moving object process starts. When a human or another moving object enters in a room where the robot is, it is detected by means of a two-level processing:
    - pixel level, in which the adaptative background model, initially built in the previous phase, is used to classify pixels as foreground or background. It is possible because each pixel belonging to the moving object has an intensity value which does not fit into the background model. That is, the used background model associates a statistical distribution (defined by its mean color value and its variance) to each pixel of the image. Then, when an interest object enters and/or moves around the robot workspace, there will be a difference between the background model values and object's pixel values. Actually, a criterion based on stored statistical information is defined to deal with this classification and it can be expressed as follows:

$$b\left(r,c\right)=\begin{cases}1 & \text{if } \left|i\left(r,c\right)-\mu_{r,c}\right|>k\times\sigma_{r,c}\\ 0 & \text{otherwise}\end{cases} \quad (1)$$

where $b\left(r,c\right)$ is the binary value of the pixel $\left(r,c\right)$ to be calculated, $i\left(r,c\right)$ represents pixel brightness in the current frame, $\mu_{r,c}$ and $\sigma_{r,c}$ are the mean and standard deviation values calculated by the background model respectively and k is a constant value which depends on the point distribution

- frame level, whereby the raw classification based on the background model is improved as well as the model is adapted when a global change in illumination occurs. A proper combination of substraction techniques has been implemented. In that way, a different segmentation process is applied at frame level and it is used to improve the segmentation carried out at pixel level. Furthermore, this processing allows the system to identify global illumination changes. That is, it is assumed that an significant illumination change has taken place when there is a change in more pixels than a half of the image size. When an event of this type occurs, a new adaptative background model is built because if it was not done, the application would detect background pixels as moving objects, since the model is based on intensity values and a change in illumination produces a variation of them.

As in the previous phase, after properly segmenting an image, two consecutive morphological operations are applied to erase isolated points or lines caused by small dynamic factors. Later, pixels classified as background are incorporated to the adaptive background model, while foreground pixels are processed by applying a tracking method.

II. Tracking Moving Targets

Once targets to be tracked have been identified, the next step is to track them. For that, first of all, a connected-component labeling algorithm is performed. However, due to segmentation errors, it might be obtained more than one labeled component for the same target. Thus, a merge algorithm, based on neighbourhood and feature similarity, is applied. Then, a minimum rounded rectangles are generated. After that, with the aim of performing the corresponding tracking, a pattern is built from each of them. In this case, a pattern is intended as the data structure such that allows the system to track the moving objects by means of matching an object in two consecutive frames even when it suffers a partial or whole occlusion.

Thus, in our case, a pattern is composed of two different things:

- a representative image of the target, that is, it is not possible to directly compare two images of the same object when they are provided by an omnidirectional image. It is due to the fact that every object has different

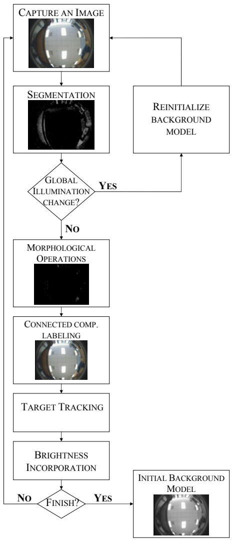


Fig. 1. Phase-1 flowchart of implemented two-phase algorithm for moving object detection in unknown, dynamic environments

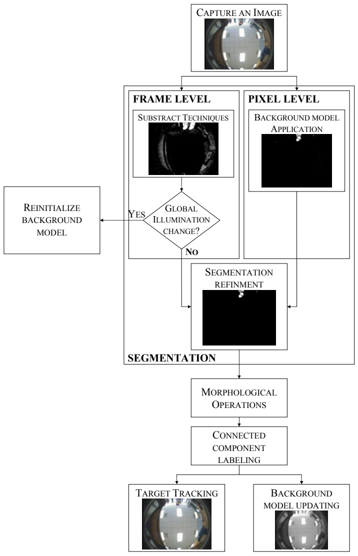
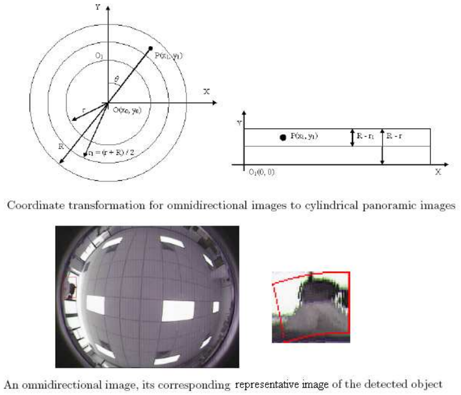

Capture an Image
FRAME LEVEL
PIXEL LEVEL
Substract Techniques
Background model Application
Reinitialize background model
Yes
Global Illumination change?
No
Segmentation refinment
SEGMENTATION
Morphological Operations
Connected component labeling
Target Tracking
Background model updating

Fig. 2. Phase-2 flowchart of implemented two-phase algorithm for moving object detection in unknown, dynamic environments

orientation depending on its position inside the scene. So, several rotations would be necessary in order to correctly match the images of the same object in two different frames. Thus, it is necessary to apply a transformation from the circular omnidireccional image to a perspective one (see Fig. 3. This is done only for each region detected as object of interest since transformation of the whole image could become very high time-consuming.

- a feature array whose elements contain information about brightness and blob width and height, among other things, used to properly match to images of the same object in two consecutive frames as well as two stereo images



Coordinate transformation for omnidirectional images to cylindrical panoramic images

An omnidirectional image, its corresponding representative image of the detected object

Fig. 3. Representative perspective image from the labeled omnidirectional image

Therefore, on the one hand, representative images are compared with the extracted from the previous frame or the another stereo image. In this way, a pixel-similarity likehood between representative images is obtained. On the other hand, a feature-similarity likehood is generated from feature array comparison. Both likehoods are properly combined to match two images from the same object in consecutive frames or frames taken from a dioptric stereo system.

III. Stereo System

For approximately determining the distance from the system to an objective we need to estimate the correspondence between omnidirectional images, that is, the epipolar equation.

From the point of view of stereo vision, an epipole is defined as the projection of the camera center on the image plane of another camera. Unlike traditional cameras, two epipoles are visible, since any camera is within the field of view of each other. For that reason, in the case of omnidirectional cameras, it is not necessary to use a third external object for stereo calibration. This is the idea in which Zhu et al [15] [16] based to implement a virtual stereovision system with a flexible baseline in order to detect, track and localize moving human subjects in unknown indoor environments. In the literature, other approaches developed for catadiotric systems can be found [16] [19] [20] [21]. However, even though there is almost no work with dioptric stereo systems, we have implemented a process to estimate the distance from the dioptric stereo system to the detected objects.

The guidelines of the distance estimation method are as follows:

- a matching process between images taken by different cameras is done. First, the adaptive background model at two levels is independently performed for the images captured by each camera. Then, each detected blob is described by means of a feature array whose elements contain information about brightness and blob width and height, among other things. Next, each feature array is compared with all the detected blobs in the frame taken by the other camera in the system, while the matching process included in the implemented moving object detection method is simultaneously performed. Thus, a similarity likelihood is calculated and a matching decision is made based on it.
- detection of the other camera in each frame, as it is depicted in Fig. 4. In fact, this step is necessary to be performed only once because the baseline of the stereo system is fixed.
- estimation of the distance with respect to each camera from triangulation geometry, as it is shown in Fig. 4. The triangle to solve is determined by the projection ray of a 3D point of the real world on each plane image as well as the projection ray of the center of the other camera of the system. It is possible thanks to the calibration step since the projection rays can be estimated, as mentioned above.

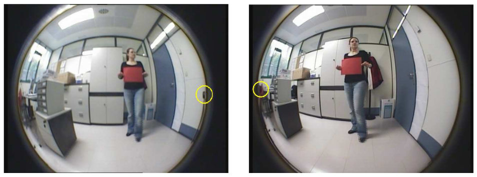


Fig. 4. Camera detection in images captured by the two cameras

IV. Experimental Results

Two different experiments have been carried out to check the designed application performance. First, the performance of the moving object detection was evaluated by using only one fisheye camera. After that, the obtained estimation of the detected object distance through our dioptric stereo system was analysed. In this section, some of these results are provided.

A. Experimental set up

For both kinds of experiments carried out, a mobile manipulator which incorporates a visual system composed of 2 fisheye cameras mounted on the robot base, pointing upwards to the ceiling, to guarantee the safety in its whole workspace. Figs. 5 depicts our experimental setup, which consists of a mobile Nomadic XR4000 base, a Mitsubishi PA10 arm, and two fisheye cameras (*SSC-DC330P third-inch color cameras* [22]

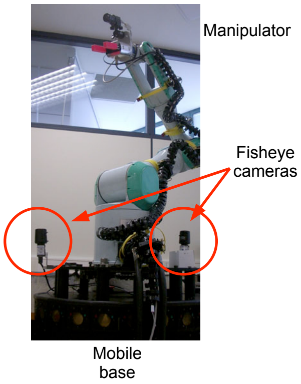


Fig. 5. Experimental setup: external view of the arm and cameras.

with *fish-eye vari-focal lenses YV2.2x1.4A-2*, which provide 185 degree field of view).

Thus, on the one hand, a single fisheye camera was used to evaluate the moving object detection performance. The robot system was located in the center of our laboratory where almost all the space was covered and where most of the uncontrolled, dynamic factors named above were present (e.g. blinking of computer screens, shadows, mirror images on the glass windows or variations in illumination due to the different time of the day or the switch on/off a light). On the other hand, the dioptric system was used. In both cases, the images to process were acquired in 24-bit RGB color space with a $640x480$ resolution.

B. Moving Object Detection Evaluation

As it was pointed out, the first series of experiments were to evaluate the performance of the novel adaptative, robust background model. For that, illumination conditions and object positions were changed. Two sequences of the images as a result of applying the novel updated background model under the same illumination conditions is depicted in Fig. 6. As it can be seen, the method is able to visually track moving objects without constraints such as clothes color or illumination.

In a similar way, illumination conditions were changed and, as it is shown in Fig. 7, the obtained results were also successful.

V. Conclusions

In this paper, a robust visual application to detect and track moving objects within a robot workspace has been presented based on a pair of fisheye cameras. These cameras have the clear advantage of covering the whole workspace without resulting in a time consuming application, but there is little previous work about this kind of devices. Consequently, we had to implement novel techniques to achieve our goal.

Thus, the first subgoal was to design a process to detect moving objects within the observed scene. After studying several factors which can affect the detection process, a novel adaptive background model has been implemented where contraints such as waiting a period of time to build the initial background or illumination conditions do not exist. In a similar

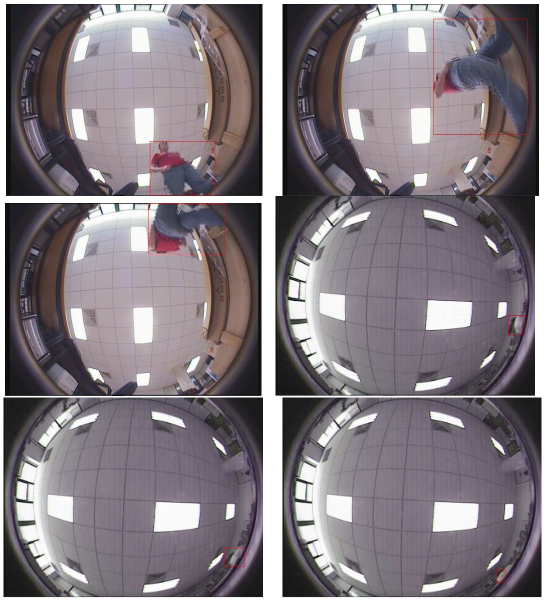


Fig. 6. Results of applying the novel adaptative background model with different subjects

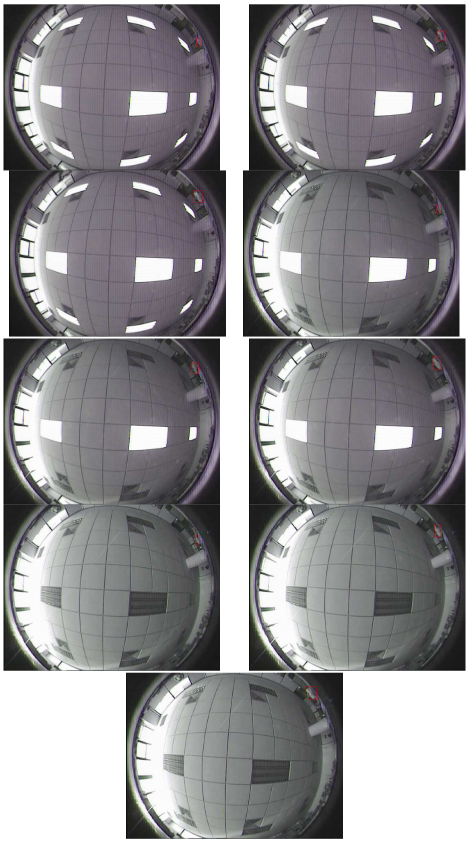


Fig. 7. Results of applying the novel adaptative background model under different illumination conditions

way, it is also capable of tracking the detected objects when they are not in movement. In addition, the designed method includes a matching process between two consecutive frames.

The next step is to estimate the distance from the detected objects to the system. For that, a stereo dioptric system with fixed baseline has been built. Therefore, it was necessary to perform a calibration process in order to obtain the fundamental matrix. Three different toolboxes were tested, but only two were used in the end. Finally, a method to estimate distance from the objects to the system was implemented. In this case, a triangulation technique is used. It is possible to perform because the cameras can see each other. It must be taken into account that epipolar geometry of the stereo dioptric systems was not used, although the combination of that with the current implementation in order to improve the accuracy in the matching process is part of our future work.

Acknowledgments

This paper describes research carried out at the Robotic Intelligence Laboratory of Universitat Jaume I. Support for this laboratory is provided in part by the European Commission under project EYESHOTS (FP7 ICT-217077), by Fundacio Caixa-Castello under project P1-1B2005-28 and by Ministerio de Ciencia under project DPI2004-01920 and FPI grant BES-2005-8860.

References

[1] T. Svoboda, T. Pajdla, and V. Hlavác, “Epipolar geometry for panoramic cameras,” in *European Conf. on Computer Vision (ECCV’98)*, Freiburt Germany, July 1998, pp. 218 – 231.

[2] S. C. Wei, Y. Yagi, and M. Yachida, “Building local floor map by use of ultrasonic and omni-directional vision sensor,” in *Int. Conf. on Robotics and Automation*, Leuven, Belgium, May 1998, pp. 2548 – 2553.

[3] S. Baker and S. K. Nayar, “A theory of single-viewpoint catadiptric image formation,” *Int. Journal of Computer Vision*, vol. 35, no. 2, pp. 175 – 196, 1999.

[4] ——, “A theory of catadioptric image formation,” in *Int. Conf. on Computer Vision (ICCV’98)*, Bombay, India, 5–8 January 1998, pp. 35 – 42.

[5] R. W. Wood, “Fish-eye views, and vision under water,” *Philosophical Magazine*, vol. 12, no. Series 6, pp. 159 – 162, 1906.

[6] C. Geyer and K. Daniilidis, “A unifying theory for central panoramic systems and practical applications,” in *European Conf. on Computer Vision (ECCV 2000)*, Dublin, Ireland, 26th June – 1st July 2000, pp. 445 – 461.

[7] S. L. G. and S. G. C, *Computer vision*, U. S. River, Ed. Prentice Hall, 2001.

[8] B. Kwolek, “Color vision based person following with a mobile robot,” in *Third Int. Workshop on Robot Motion and Control (RoMoCo’02)*, November 2002, pp. 375 – 380.

[9] M. Tarokh and P. Ferrari, “Case study: Robotic person following using fuzzy control and image segmentation,” *Robotic Systems*, vol. 20, no. 9, pp. 557 – 568, 2003.

[10] M. Kobilarov, G. Sukhatme, J. Hyams, and P. Batavia, “People tracking and following with mobile robot using an omnidirectional camera and a laser,” in *2006 IEEE Int. Conf. on Robotics and Automation (ICRA’06)*, Orlando, Florida, May 2006, pp. 557 – 562.

[11] T. Yoshimi, M. Nishiyama, T. Sonoura, H. Nakamoto, S. Tokura, H. Sato, F. Ozaki, N. Matsuhira, and H. Mizoguchi, “Development of a person following robot with vision based target detection,” in *2006 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, Beijing, China, October 2006, pp. 5286 – 5291.

[12] K. Toyama, J. Krum, B. Brumitt, and B. Meyers, “Wallflower: Principles and practice of background maintenance,” in *Seventh IEEE Int. Conf. on Computer Vision*, vol. 1, Kerkyra, Greece, 1999, pp. 255 – 261.

[13] I. Haritaoglu, D. Harwood, and L. S. Davis, “W4: Real-time surveillance of people and their activities,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 809 – 830, August 2000.

[14] H. Liu, W. Pi, and H. Zha, “Motion detection for multiple moving targets by using an omnidirectional camera,” in *IEEE Int. Conf. on Robotics, Intelligent Systems and Signal Processing*, vol. 1, Changsha, China, October 2003, pp. 422 – 426.

[15] Z. Zhu, K. D. Rajasekar, E. M. Riseman, and A. R. Hanson, “Panoramic virtual stereo vision of cooperative mobile robots for localizing 3d moving objects,” in *IEEE Workshop on Omnidirectional Vision*, 12th June 2000, pp. 29 – 36.

[16] Z. Zhu, D. R. Karuppiah, E. M. Riseman, and A. R. Hanson, “Keeping smart, omnidirectional eyes on you. adaptive panoramic stereovision for human tracking and localization with cooperative robots,” *IEEE Robotics and Automation Magazine*, pp. 69 – 78, December 2004, special Issue on Panoramic Robotics.

[17] S. C. and G. W.E.L., “Adaptive background mixture models for real-time tracking,” in *Conference on Computer Vision and Pattern Recognition (CVPR’99)*, vol. 2, 23rd – 25th June 1999, pp. 246 – 252.

[18] K. P. and B. R., “An improved adaptive background mixture model for real-time tracking with shadow detection,” in *2nd European Workshop on Advanced Video Based Surveillance Systems (AVBS 01), VIDEO BASED SURVEILLANCE SYSTEMS: Computer Vision and Distributed Processing*, K. A. Publisher, Ed., September 2001.

[19] C. Geyer and K. Daniilidis, “Properties of the catadioptric fundamental matrix,” in *The 7th European Conf. on Computer Vision (ECCV2002)*, vol. 2, LNCS 2351, Copenhagen, Denmark, 27th May – 2nd June 2002, pp. 140 – 154. [Online]. Available: http://link.springer.de/link/service/series/0558/tocs/t2351.htm

[20] Y. Negishi, J. Miura, and Y. Shirai, “Calibration of omnidirectional stereo for mobile robots,” *2004 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS 2004)*, vol. 3, pp. 2600 – 2605, 28th September – 2nd October 2004.

[21] S. Li and K. Fukumori, “Spherical stereo for the construction of immersive vr environment,” in *IEEE Virtual Reality (VR’05)*, 12th – 16th March 2005, pp. 217 – 222.

[22] http://www.infodip.com/pages/sony/camera/pdf/SSC-DC58AP.pdf.