Human Segmentation and Pose Recognition in Fish-eye Video for Assistive Environments

K.K. Delibasis, V.P. Plagianakos
Department of Computer Science and Biomedical Informatics
University of Thessaly, Greece
kdelibasis@yahoo.com; vpp@ucg.gr

T. Goudas, I. Maglogiannis
Department of Digital Systems
University of Piraeus, Greece
goudas@ucg.gr; imaglo@unipi.gr

***Abstract*—In this work, we present a system, which uses computer vision techniques for human silhouette segmentation from video in indoor environments and a parametric 3D human model, in order to recognize the posture of the monitored person. The video data are acquired indoors from a fixed fish-eye camera in the living environment. The implemented 3D human model collaborates with a fish-eye camera model, allowing the calculation of the real human position in the 3D-space and consequently recognizing the posture of the monitored person. The paper discusses briefly the details of the human segmentation, the camera modeling and the posture recognition methodology. Initial results are also presented for a small number of video sequences.**

Keywords: Video segmentation, background modeling, fisheye camera modeling, 3D human modeling, posture recognition.

I. INTRODUCTION

Present pervasive and sensor technology allow for extensive monitoring of all kinds of data on a person's daily lifestyle, such as exercise, steps taken, body weight, food consumed, blood pressure, cigarettes smoke and may other activities related to human routine. This type of data tracking is often referred to as the Quantified Self Task. Research has shown that tracking such values, is effective in itself as the mere activity of tracking makes the user conscious of its importance, which can already lead to lifestyle change. Fogg mentions self-monitoring and monitoring by others (surveillance) as effective strategies for behavioural change [1]. Solutions available in the state-of-the-art include mobile apps for GPS tracking of cardio activities (i.e. Runkeeper), like running and cycling, websites for tracking body weight, fat percentage and bone mass using Wifi connected body scales (Withings Body Scale), mobile apps for blood pressure monitoring, activity apps with integrated (wireless) activity monitor (Activ8, Fitbit activity monitor), etc.

A convenient way to monitor everyday activities may be achieved utilizing fixed cameras in indoor environments. This field has gained significant interest during the last years, due to the non-intrusiveness in monitoring. It finds a wide variety of applications in diverse areas, such as assistive environments, supporting the elderly or the chronic ill, surveillance and security, traffic control, industrial processes etc. The presented here work focuses on human silhouette segmentation and human pose estimation from fisheye video. The extracted pose can be used at a later stage for activity detection by a decision support system. The proposed algorithm for pose extraction is based on a parametric three-dimensional (3D) human model that can move its legs and arms, as well as on a model of the fisheye camera that allows the rendering of the parametric model. An evolutionary algorithm is used to recover the parameters of the 3D human model.

The rest of the paper is structured as follows: Section II presents the related in this field work, while Section III discuss the proposed methodology. Section IV presents some initial results proving the feasibility of the proposed ideas and concludes the paper.

II. RELATED WORK AND BACKGROUND INFORMATION

A. Video segmentation

Most of the video segmentation algorithms are based on background subtraction. The background has to be modeled, since it may change due to a number of reasons, including: motion of background objects, differences in light conditions, or video compression artifacts. Therefore, a number of techniques have been proposed for constructing a model of the background that is being gradually updated using the values of the current video frame.

In [2] the background model is, simply, defined as the previous frame. Background can be modelled by median filtering [3] of a predefined number of last frames that are hold in a buffer. The background value of each pixel in the model is independently computed as the temporal median of the pixel values along the buffer. This approach however may become slow for large frame sizes. Therefore, a class of recursive background modeling algorithms exists that use an incremental update of the background, without the need of a buffer to store previous frames, such as McFarlane, and Schofield [4]. The running average [1] is another example of this class of algorithms. An extension of the aforementioned method is the running Gaussian average that was proposed in [5]. A popular member of this class of algorithms is the Mixture of Gaussians (MoG), initially described for video sequences by Stauffer and Grimson in [6]. Another simple method is the approximated median filtering method as described by McFarlane, and Schofield [4]. In this work we have performed video segmentation using more complicate methods such as the

978-1-4799-3163-7/13/$31.00 ©2013 IEEE

illumination sensitive method (ISB) [7], as well as the Self-Organizing Background Subtraction method (SOBS) described in [8]. Furthermore we have simplified the SOBS method by removing the color information and using only the gray pixel values, since this significantly accelerates their execution, without significant effect on the results.

B. Three-dimensional human models

Articuated stick human models are quite popular for pose estimation. Volumetric human models use geometric primitives such as are spheres, cylinders or tapered super-quadrics. Surface-based models of the human body typically consist of a mesh of polygons that may be deformed [9]. Pose estimation is achieved by recovering the values of the human model parameters. A number of reported works use the "top-down estimation", by comparing the rendered 3D human model with the actual frame, using local search methods [10].

A small number of approaches use information about the camera model and setting, to assist pose recovery. In [11] the perspective information of a mono-ocular camera is used. In [12] reconstruction of 3D poses from 2D point correspondences is reported, using multiple views and known body segment lengths. Detailed description of these approaches can be found in the survey [13].

C. Fisheye camera modeling

The use of 360 field of view (FoV) hemispheric cameras, also known as fisheye cameras, is increasing both in robotic applications and in video surveillance [14], [15]. As fisheye cameras with megapixel sensors are now available, research in calibration of such cameras becomes increasingly useful. Thus, the topic of fish-eye camera calibration has attracted significant attention, on its own. In [16] and [17] the calibration of fisheye camera is reported using high degree polynomials to emulate the strong deformation effects introduced by the fisheye lens. In [18] a detailed model is presented for fisheye camera calibration, which estimates the radial and tangenial deformation, using a polynomial mapping between the radial distance from the optical axis of a real world point and its imaged point on the image plane.

III. MATERIALS AND METHODS

A. Overall Architecture

In this work we propose a system for human silhouette segmentation and pose recognition from indoor video using a single fisheye camera. Our system uses modified algorithms for human segmentation, a simple triangulated 3D human model with a small degree of freedom and an evolutionary algorithm that recovers the parameters of the 3D human model, based on an objective function that compares the rendering of the 3D model through the camera model to the segmented human. The overall architecture is shown in Fig. 1.

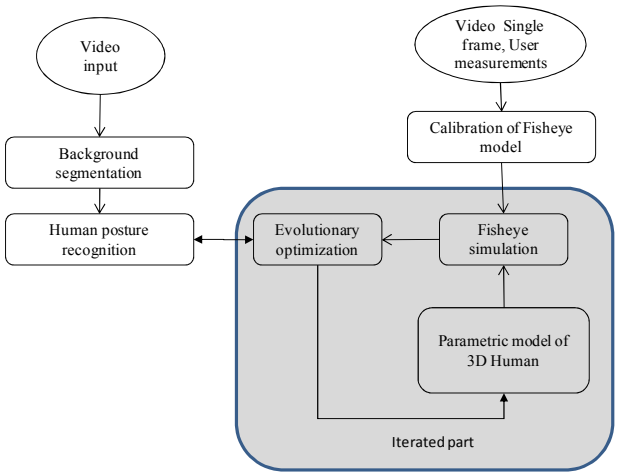


Fig. 1. The overall system architecture.

B. Human Segmentation

Illumination sensitive method (ISB) utilized in this work maintains two background models: a light and a dark one (B_{light}, B_{dark} respectively) and the currently active Background model BG. The background, B_{light} and B_{dark} are initialized as the 1st frame. The algorithm also maintains a local threshold V_t. For each incoming frame FR, the BG is compared to FR and updated accordingly. The entropy E_t of FR is also calculated. If the entropy E_t does not differ significantly from the entropy of the previous frame E_{t-1}, then V_t is used to threshold the difference between the background BG and the frame FR. The V_t is also updated accordingly. If the entropy of the current frame differs significantly from the entropy of the previous one, then the dark and light backgrounds are updated and one of them becomes the currently active background model BG. The details of the algorithm are shown in Figure 2.

Self-organizing maps (SOMs) are artificial neural networks (ANN) that are based on competitive learning and require no training set [8]. In [7], a 2D set of neurons of equal dimensions NxM with the video frame used to model the background. Each neuron has 3x3 weights that are arranged into a 2D array of dimensions 3Nx3M. This array will be called neuronal map A. Each subsequent frame is presented to the SOM. For each pixel p of the frame FR, the corresponding 3x3 weights of the neuronal map A are detected and the weight that is most similar to the value of p, FR(p), is determined. The similarity is determined in the HSV space. If the similarity is greater than a predetermined threshold, then the algorithm checks if the pixel belongs to shadow, according to a criterion implemented in the HSV color space, described in detail [2]. It has to be mentioned here that this criterion is tested for all 3x3 weights of each pixel. Our implementation has simplified the SOBS method by removing the part of the algorithm that uses the color information (HSV color space), including the similarity criterion and the shadow detection. In this way the execution is accelerated, the required data structures are simplified and the method can be comparable to the illumination sensitive method, which also operates using only the gray level pixel values. The details of the algorithm are shown in Figure 3.

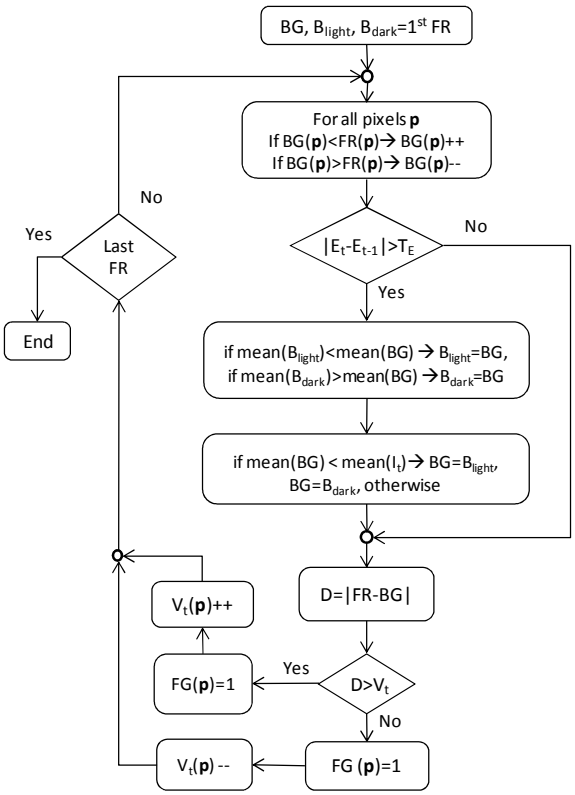


Fig. 2. The illumination sensitive (ISB) video segmentation algorithm

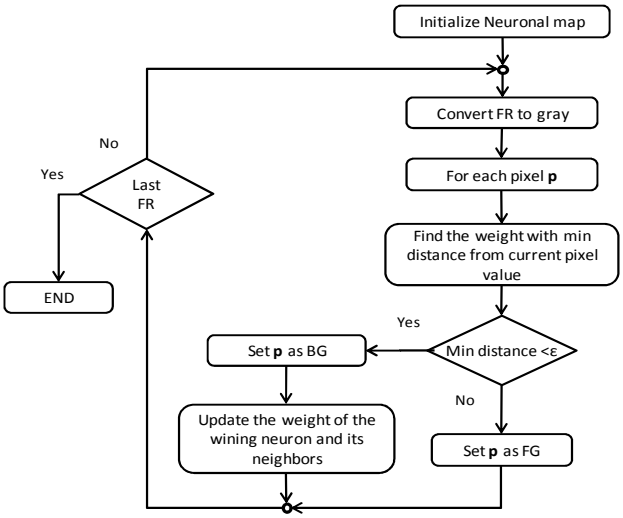


Fig. 3. The SOBS algorithm, as modified for our implementation (see text for details).

C. Fisheye Camera model

The main characteristic of the fisheye camera is that it can cover a field of view of 360 degrees. In this work we use a model to simulate the image formation using the fisheye camera, so that given the real-world position of an object (x,y,z), we may calculate the image coordinates (j,i) of its pixels. The action of the fisheye model M can be written in the general form

$$(j,i) = M(x,y,z) \quad (1)$$

We adopted a model for the fisheye camera that is based on the image formation physics described in [20], [21] and demonstrated in [22]. We consider a spherical element of arbitrary radius R_0 with its center at K (0,0,z_{sph}) (as it will become clear soon, the radius R_0 is not a parameter of the model). For any point P with real world coordinates (x,y,z), we determine the intersection Q of the line KP with the spherical optical element of the fisheye lens. The point P is imaged at the central projection (x_{im}, y_{im}) of Q on the image plane with equation z=z_{plane}, using the O (0,0,0) as center of projection, assuming that the installation of the camera is such that the imaging plane (i.e. the image sensor) is horizontal and the axis of the spherical lens is not misaligned. Thus, it becomes obvious that all real world points that lie on the KP line are imaged at the same point (x_{im}, y_{im}) in the image plane. The KP line is uniquely defined by its azimuth and elevation angles, θ, φ respectively. The concept of the fisheye geometric model is shown in Fig. 4.

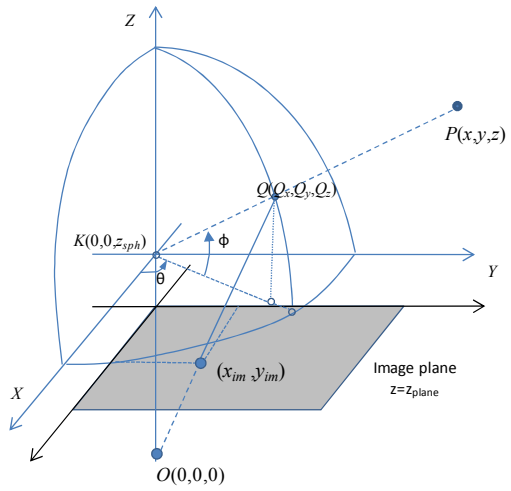


Fig. 4. The geometry of the fisheye model. See text for details.

The fisheye camera has no moving parts. Therefore, the ratio $p = z_{sph}/z_{plane}$ is the primary parameter of the fisheye model. z_{plane} is set to an arbitrary value less than R_0, thus $z_{sph} = pz_{plane}$. It is possible that small internal lens misalignments may introduce unanticipated imaging deformations [19]. In order to account for these as well, we introduce two extra model parameters, the X and Y position of the center of spherical lens x_{sph}, y_{sph} with respect to the optical axis of the camera. Thus the center of the spherical element becomes $K(x_{sph}, y_{sph}, z_{sph})$. Now the camera model parameters consist of p, x_{sph}, y_{sph}. For simplicity, Fig. 4 shows the case for $x_{sph} = 0, y_{sph} = 0$.

It can be proved that the image pixel (i,j) that corresponds to (x_{im}, y_{im}) is calculated by a simple linear transform

$$(j,i) = (x_{im}, y_{im}) \frac{R_{FoV}}{R_0} + (CoD_x, CoD_y) \quad (2)$$

where (CoDx, CoDy) is the center of distortion pixel that corresponds to elevation φ=π/2 and RFoV is the radius of the circular field of view (FoV). As suggested in [20] the CoD is located as the center of the circular field-of-view. We therefore apply the canny edge detector, using a standard deviation equal to 2 in order to detect the stronger edges in the image, which are the edges of the circular field of view. Then, we employ a simple least squares optimization to obtain the CoD and the radius of the FoV. This is done only once, during the calibration of the camera model.

D. Human model

We utilized a free triangulated model of a standing human, consisting of approx. 27.000 vertices. Since we are interested in projecting the human model through the fisheye camera in real time, we discard the triangle information of the model and we treat it as a cloud of points. We also applied a vertex decimation process to reduce the number of vertices by a factor of 8.

The vertices of the model were labeled manually using logical spatial relations, into 5 classes: right and left arm, right and left leg and the rest of the body (torso and head). The human model is used to produce a simulated (rendered) segmented video frame that is compared to the actually segmented one. Figure 5 illustrates this concept. A model of a standing man is scaled to height=1.8 m and is placed at several locations in the imaged room, touching the floor in Fig. 5, while the rendered frame using the fisheye model is shown in Fig. 6.

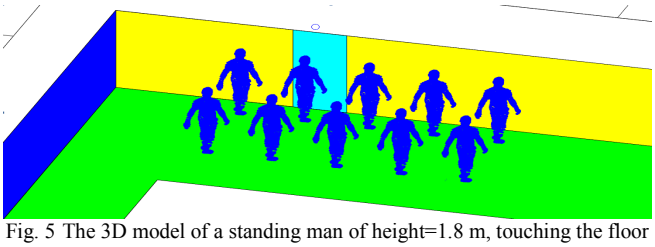


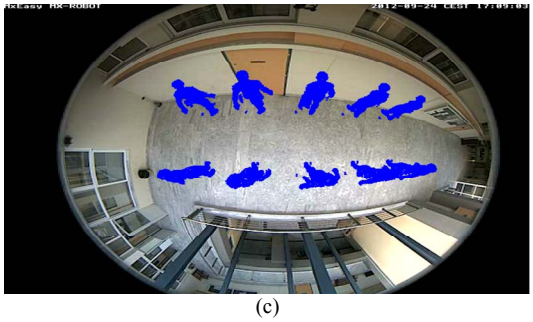
Fig. 5 The 3D model of a standing man of height=1.8 m, touching the floor and reproducing the model at several locations of the imaged room

E. Matching the human model with the segmented frame.

The human pose can be extracted by recovering the values of the parameters of the human model stored in vector pm. Let us denote IM the binary image of the parametric model generated by the fisheye model and IS the segmented image of the corresponding video frame. The objective function is defined as following:

$$f(\mathbf{p_m}) = \sum_{\substack{image \\ domain}} I_M \cap I_S - \sum \bar{I}_M \cap I_S - \sum I_M \cap \bar{I}_S \quad (3)$$

where, $\bar{I}$ denotes the boolean negative of I, $\cap$ denotes the boolean AND operator and the summation is done over the whole image domain. Thus, the objective function is defined as the number of non-zero pixels of I_M on non-zero pixels of I_S minus the number of non-zero pixels of I_M on zero pixels of I_S minus the number of zero pixels of I_M on non-zero pixels of I_S. It is evident that for the ideal set of model parameters, the value of the objective function should be maximized.

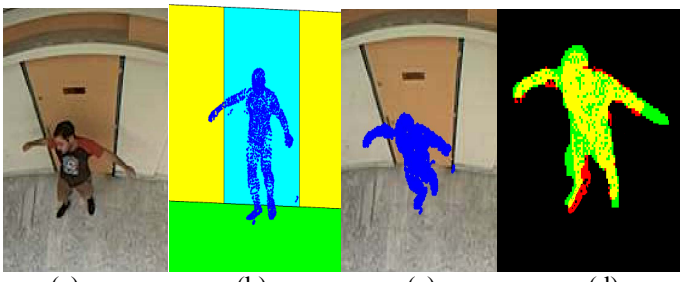


(c)

Fig. 6. The 3D model of a standing man, rendering the 3D human models through the fisheye camera model.

IV. Initial Results and Discussion

The video sequences used for experimental results were acquired using the Mobotix Q24 hemispheric camera, which was installed on the ceiling of the imaged room. The pixilation of each frame is 480x640. Figure 7 shows results from a characteristic frame (a). The recovered 3D human model is shown in (b) and the rendered 3D human model using the fisheye model is shown in (c). The fitting of the fisheye-rendered parametric 3D human model to the segmented frame is shown in Fig. 7(d), as following: segmented frame in green, the fisheye rendered 3D human in red and their intersection in yellow. It can be observed that the proposed algorithm was able to detect the specific human pose.



(a) (b) (c) (d)

Fig. 7. (a) An original video frame showing a human, (b) the 3D model with its parameters fitted to the segmented original frame and (c) the fisheye-rendered 3D model in a simulated frame and (d) the corresponding fitting

Regarding the complexity of the proposed algorithms, the video segmentation algorithm requires approximately 50 msec per frame. The generation of the 3D parametric human model is executed very fast, with approximately 1 msec including the required geometric transformations. The rendering of a 3D human model with 2.500 vertices through the modeled fisheye camera is performed in 1.5 msec. The calculation of the objective function (3) requires approximately 8 msec for a 3D model with 2.500 vertices and a frame of 480x640 pixels, although the number of pixels in the frame does not drastically affect the execution time. All timing was performed using an Intel(R) Core i5-2430 CPU @ 2.40 GHz Laptop with 4 GB Ram, under Windows 7 Home Premium, using the Matlab programming environment. No special code optimization or any kind of parallelization was

performed. It becomes clear that the optimization of (3) in order to extract the human pose has not been performed in real time. Further work will include, the adoption of a more robust statistical 3D model for the human and the refinement of the fisheye model, using more parameters to increase its accuracy. Finally more efficient implementation of the genetic algorithm based optimization will be explored for the determination of the 3D model parameters. These approaches include exploiting the converged population from the previous frames to initialize the search for the current frame and/or restricting the parameter range according to their optimal values from the previous frames.

ACKNOWLEDGMENT

The authors would like to thank the European Union (European Social Fund ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: \Thalis \ Interdisciplinary Research in Affective Computing for Biological Activity Recognition in Assistive Environments for financially supporting this work.

REFERENCES

[1] Fogg, B. J, Persuasive Technology: Using Computers to Change What We Think and Do. Morgan Kaufmann Publishers, 2003.

[2] J. Willems, G. Debard, B. Bonroy, B. Vanrumste and T. Goedemé, "How to detect human fall in video? An overview", In Proceedings of the positioning and contex-awareness international conference (Antwerp, Belgium, 28 May, 2009), POCA '09.

[3] Cucchiara, R., Grana, C., Piccardi, M., and Prati A. 2003. Detecting moving objects, ghosts, and shadows in video streams. IEEE Transactions on Pattern Analysis and Machine Intelligence 25, 10, (2003), 1337-1442.

[4] N. McFarlane and C. Schofield, "Segmentation and tracking of piglets in images", MACH VISION APPL. 8, 3, (May. 1995), 187-193.

[5] Wren, C., Azarhayejani, A., Darrell, T., and Pentland, A. P. 1997. Pfinder: real-time tracking of the human body, IEEE Transactions on Pattern Analysis and Machine Intelligence 19, 7, (October. 1997), 780-785.

[6] C. Stauffer, and W. Grimson, "Adaptive background mixture models for real-time tracking". In Proceedings of the conference on computer vision and pattern recognition (Ft. Collins, USA, June 23-25, 1999), CVPR '99. IEEE Computer Society, New York, NY, pp. 246-252.

[7] F. C. Cheng, S. C. Huang, S. J. Ruan," Illumination-Sensitive Background Modeling Approach for Accurate Moving Object Detection," IEEE trans. on broadcasting, vol. 57, no. 4, dec. 2011.

[8] Maddalena L., Petrosino A. A self organizing approach to background subtraction for visual surveillance applications, IEEE Transactions on Image Processing, Volume17, No. 7, pages 1729–1736, 2008.

[9] Q. Delamarre, O. Faugeras, "3D articulated models and multiview tracking with physical forces", Computer Vision and Image Understanding (CVIU) 81 (3) (2001) 328–357.

[10] R. Kehl, L. Van Gool, "Markerless tracking of complex human motions from multiple views", Computer Vision and Image Understanding (CVIU) 104 (2–3) (2006) 190–209.

[11] C. Barron, I. Kakadiaris, "Estimating anthropometry and pose from a single uncalibrated image", Computer Vision and Image Understanding (CVIU) 81 (3) (2001) 269–284.

[12] C. Bregler, J. Malik, K. Pullen, "Twist based acquisition and tracking of animal and human kinematics", International Journal of Computer Vision 56 (3) (2004) 179–194.

[13] C Taylor, Reconstruction of articulated objects from point correspondences in a single uncalibrated image, Computer Vision and Image Understanding (CVIU) 80 (3) (2000) 349–363.

[14] D. Liebowitz, S. Carlsson, "Uncalibrated motion capture exploiting articulated structure constraints", International Journal of Computer Vision 51 (3) (2003) 171–187.

[15] R. Poppe, "Vision-based human motion analysis: An overview", Computer Vision and Image Understanding, 108 (2007) 4–18.

[16] Z. Zhou, X. Chen, Y. Chung, Z. He, T. X. Han and J. M. Keller, "Activity Analysis, Summarization and Visualization for Indoor Human Activity Monitoring," IEEE Trans. on Circuit and systems for Video Technology, Vol. 18, No. II, pp. 1489-1498, 2008.

[17] M. Saito and K. Kitaguchi, G. Kimura and M. Hashimoto, "Human Detection from Fish-eye Image by Bayesian Combination of Probabilistic Appearance Models", IEEE International Conference on Systems Man and Cybernetics (SMC), 2010, pp243-248.

[18] H. Li and R. Hartley, "Plane-Based Calibration and Auto-calibration of a Fish-Eye" Camera, P.J. Narayanan et al. (Eds.): ACCV 2006, LNCS 3851, pp. 21–30, 2006, c Springer-Verlag Berlin Heidelberg 2006.

[19] A. Basu, S. Licardie, "Modeling fish-eye lenses", Proceedings of the 1993 IEEWSJ International Conference on Intelligent Robots and Systems Yokohama, Japan July 2630,1993.

[20] N. Greene, "Environment Mapping and Other Applications of World Projections", IEEE Computer Graphics and Applications, November 1986, vol. 6(11), pp 21.

[21] http://paulbourke.net/dome/fisheye