Bimodal Active Stereo Vision

Andrew Dankers[1,2], Nick Barnes[1,2], and Alex Zelinsky[3]

[1] National ICT Australia[4], Locked Bag 8001, Canberra ACT Australia 2601
[2] Australian National University, Acton ACT Australia 2601
{andrew.dankers,nick.barnes}@nicta.com.au
[3] CSIRO ICT Centre, Canberra ACT Australia 0200
alex.zelinsky@csiro.au

Summary. We present a biologically inspired active vision system that incorporates two modes of perception. A peripheral mode provides a broad and coarse perception of where mass is in the scene in the vicinity of the current fixation point, and how that mass is moving. It involves fusion of actively acquired depth data into a 3D occupancy grid. A foveal mode then ensures coordinated stereo fixation upon mass/objects in the scene, and enables extraction of the mass/object using a maximum a-posterior probability zero disparity filter. Foveal processing is limited to the vicinity of the camera optical centres. Results for each mode and both modes operating in parallel are presented. The regime operates at approximately $15Hz$ on a $3GHz$ single processor PC.

Keywords: Active Stereo Vision Road-scene Fovea Periphery

1 Introduction

The National ICT Australia (NICTA) Autonomous Systems and Sensing Technologies (ASSeT) *Smart Car* project focusses on *Driver Assistance Systems* for increased road safety. One aspect of the project involves monitoring the driver and road scene to ensure a correlation between where the driver is looking, and events occurring in the road scene [11]. The detection of objects in the road scene such as signs [19] and pedestrians [14], and the location of the road itself [2], form part of the set of observable events that the system aims to ensure the driver is aware of, or warn the driver about in the case that they have not noticeably observed such events. In this paper, we concentrate on the use of active computer vision as a scene sensing input to the driver assistance architecture. Scene awareness is useful for tracking objects, classifying them, determining their absolute position or fitting models to them.

[4] National ICT Australia is funded by the Australian Department of Communications, Information Technology and the Arts and the Australian Research Council through *Backing Australia's ability* and the ICT Centre of Excellence Program.

P. Corke and S. Sukkarieh (Eds.): Field and Service Robotics, STAR 25, pp. 79–90, 2006.

1.1 Research Platform

The *Smart Car* (Fig. 1, left), a 1999 Toyota Landcruiser, is equipped with the appropriate sensors, actuators and processing hardware to provide an environment in which desired driver assistance competencies can be developed [9]. Positioned centrally inside the front windscreen is an active stereo vision mechanism. CeDAR, the Cable-Drive Active-Vision Robot [22], incorporates a common tilt axis and two pan axes each exhibiting a range of motion of 90^o. Angles of all three axes are monitored by encoders that give an effective angular resolution of 0.01^o. An additional CeDAR unit (Fig. 1, right) identical to the unit in the Smart Car is used for initial visual experiments. Although it is stationary and cannot replicate road conditions, it is convenient for algorithm development such as that presented in this paper.

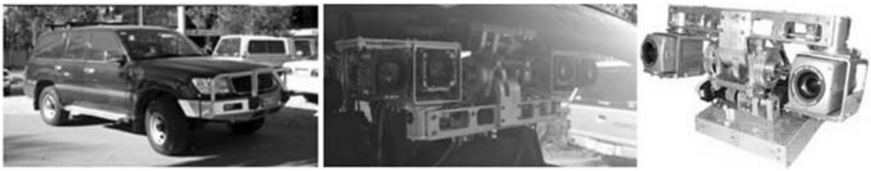

Fig. 1. Research platform. Left: *Smart Car*, and *CeDAR* mounted behind the windscreen (centre). Right: CeDAR, laboratory apparatus.

2 Active Vision for Scene Awareness

A vision system able to adjust its visual parameters to aid task-oriented behaviour – an approach labeled *active* [1] or *animate* [4] vision – can be advantageous for scene analysis in realistic environments [3]. Foveal systems must be able to align their foveas with the region of interest in the scene. Varying the camera pair geometry means foveal attention can be maintained upon a subject. It also increases the volume of the scene that may be depth-mapped. Disparity map construction using a small disparity search range that is scanned over the scene by varying the camera geometry is less computationally expensive than a large static disparity search. A configuration where fixed cameras use pixel shifting of the entire images to simulate horopter reconfiguration is more processor intensive than sending commands to a motion axis. Such *virtual* shifting also reduces the useful width of the image by the number of pixels of shift.

3 Bimodal Active Vision

We propose a biologically inspired vision system that incorporates two modes of perception. A peripheral mode first provides a broad and coarse perception of where mass is in the scene in the vicinity of the current fixation point (regardless of where that may be) and how that mass is moving. The images are processed in their entirety. It does not, however, incorporate the notion of coordinated gaze fixation or object segmentation. Once the peripheral mode has provided a rough perception of where mass is in the scene, the foveal mode allows coordinated stereo fixation upon mass/objects in the scene, and enables extraction of the object or region of mass upon which fixation occurs. We limit foveal processing resources to the region of the images immediately surrounding the image centres.

The human vision system provides the motivation for bimodal perception. Humans find it difficult to fixate on *unoccupied space*. Empty space contains little information; we are more concerned with interactions with objects or mass. Additionally, the human visual system exhibits its highest resolution around the fixation point, over a region of approximately the size of a fist at arms length. The periphery, despite being less resolute, is very sensitive to salient scene features such as colourful or moving objects [21]. For resolute processing, humans centre objects detected in the periphery within the fovea.

3.1 Peripheral Perception

We first provide an overview of the process required to rectify epipolar geometry for active stereo image pairs. Rectified pairs are then used to construct depth maps which are incorporated into an occupancy grid representation of the scene. We also describe how the flow of mass in the occupancy grid is estimated. These techniques provide a coarse 3D perception of mass in the scene.

Active Rectification and Depth Mapping

In [7] we described a rectification method used to actively enforce *parallel epipolar geometry* [15] using camera geometric relations. Though the geometric relations can be determined by visual techniques (see [20]), we use a fixed baseline and encoders to measure camera rotations. We have shown the effectiveness of the rectification process by using it to create globally epipolar rectified mosaics of the scene as the cameras were moved (Fig. 2). The mosaic process allows the use of any static stereo algorithms on an active platform by imposing a globally static image frame and parallel epipolar geometry. Here, we use the process for active depth-mapping. Depth maps are constructed using a processor economical *sum of absolute differences* (SAD) technique with *difference of Gaussians* (DOG) pre-processing[4] to reduce the effect of intensity variations [5].

[4] DOG is an approximation to the *Laplacian of Gaussian.*

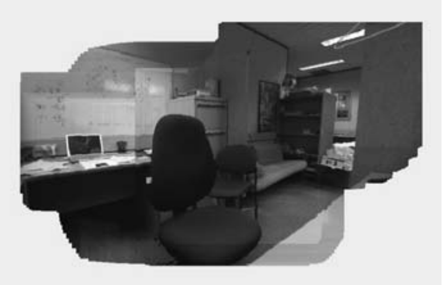

Fig. 2. Online output of the active rectification process: mosaic of rectified frames from right CeDAR camera.

A Space Variant Occupancy Grid Representation of the Scene

Occupancy grids can be used to accumulate diffuse evidence about the occupancy of a grid of small volumes of space from individual sensor readings and thereby develop increasingly confident maps [10]. Occupancy grids permit Bayesian integration of sensor data. Each pixel in a disparity map is a single measurement for which a sensor model is used to fuse data into the 3D occupancy grid. The occupancy grid is constructed such that the size of a cell at any depth corresponds to a constant amount of pixels of disparity at that depth. It is also constructed such that rays eminating from the origin pass through each layer of the occupancy grid in the depth direction at the same coordinates [7]. Fig. 3 (left) shows an example snapshot of occupancy grid construction.

As described in [8], the velocities of occupied cells in the 3D grid are calculated using an approach similar to that of [16]. This approach estimates 2D optical flow in each image and depth flow from consecutive depth maps. The mosaics remove the effect of camera rotations so that SAD based flow estimation techniques can be used to determine the vertical and lateral components of scene flow (Fig. 3, centre). We are able to assign sub-cell sized motions to the occupied cells in the occupancy grid. The occupancy grid approach was used to coarsely track the location and velocity of the ground plane and objects in the scene [8] (Fig. 3, right) at approximately $20Hz$.

3.2 Foveal Perception

We begin by assuming short baseline stereo fixation upon an object in the scene. We can ensure fixation on an object by placing it at the vergence point using saliency based attention mechanisms[5]. We want to find the boundaries of the object so we can segment it from its background, regardless of the type

[5] Gaze arbitration combines 2D visual saliency operations with the occupancy grid perception. However, visual saliency and gaze arbitration are not within the scope of this paper.

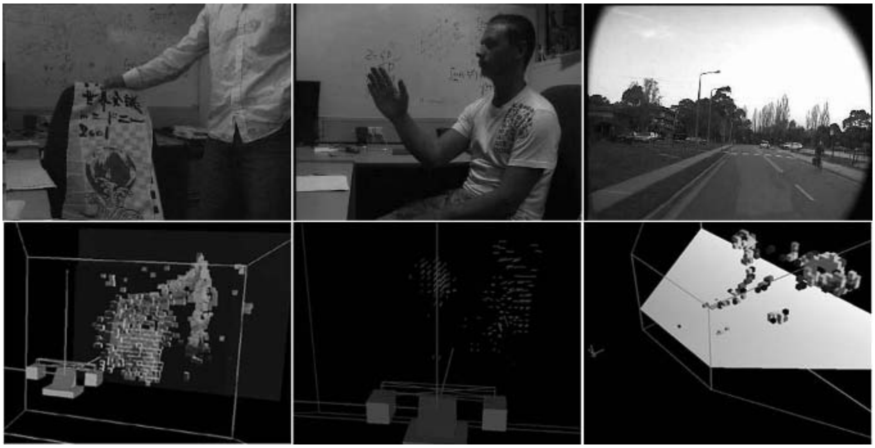

Fig. 3. Peripheral perception. Left: left camera image (top) and occupancy grid representation of mass in the scene with surface rendering (bottom). Centre: left camera image (top) and 3D mass flow vectors (bottom). Right: left camera image of road scene (top) and occupancy grid representation showing ground plane extraction (bottom).

of object or background configuration. Analogous to human vision, we define the fovea as approximately the size of a fist held a distance of 60*cm* from the camera. For our cameras, this corresponds to a region of about 60x60 pixels.

For humans, the boundaries of an object upon which we have fixated emerge effortlessly because the object is centred and appears identical in our left and right eyes, whereas the rest of the scene usually does not. For synthetic vision, the approach is the same. The object upon which fixation has occurred will appear with identical pixel coordinates in the left and right images, that is, it will be at *zero disparity.* For a pair of cameras with suitably similar intrinsic parameters, this condition does not require epipolar or barrel distortion rectification of the images. Camera calibration, intrinsic or extrinsic, is not required.

ZDF Formulation

A *zero disparity filter* (ZDF) is formulated to identify objects that map to image frame pixels at the same coordinates in the left and right fovea. Fig. 5 shows example ZDF output. Simply comparing the intensites of pixels in the left and right images at the same coordinates is not adequate due to inconsistencies in (for example) saturation, contrast and intensity gains between the two cameras, as well as focus differences and noise. A human can easily distinguish the boundaries of the object upon which fixation has occurred even if one eye looks through a tinted lens. Accordingly, the regime should be robust enough to cope with these types of inconsistencies. One approach is to

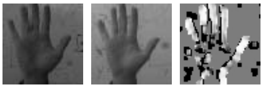

Fig. 4. NCC of 3x3 pixel regions at same coordinates in left and right images. Correlation results with higher values shown more white.

correlate a small template in one image with pixels in the same template in the other image. Fig. 4 shows the output of this approach. Bland areas in the images have been surpressed (set to 0.5) using DOG pre-processing. This is because untextured regions will always return a high NCC response whether they are at zero disparity or not. The output is sparse and noisy. The palm is positioned at zero disparity but is not categorised as such. To improve results, image context needs to be taken into account. For this reason, we adopt a Markov Random Field [13] (MRF) approach. The MRF formulation defines that the value of a random variable at the set of sites (pixel locations) P depends on the random variable configuration field f (labels at all sites) only through its neighbours $N \in P$. For a ZDF, the set of possible labels at any pixel in the configuration field is binary, that is, sites can take either the label *zero disparity* ($f(P) = l_z$) or *non-zero disparity* ($f(P) = l_{nz}$). For an observation O (in this case an image pair), Bayes law states that the a-posterior probability $P(f \mid O)$ of field configuration f is proportional to the product of the likelihood $P(O \mid f)$ of that field configuration given the observation and the prior probability $P(f)$ of realisation of that configuration:

$$P(f \mid O) \propto P(O \mid f) \cdot P(f). \tag{1}$$

The problem is thus posed as a MAP optimisation where we want to find the configuration field $f(l_z, l_{nz})$ that maximises the a-posterior probability $P(f \mid O)$. In the following two sections, we construct the terms in Eq. 1.

Prior $P(f)$

The prior encodes the properties of the MAP configuration we seek. It is intuitive that the borders of zero disparity regions co-incide with edges in the image. From the approach of [6], we use the Hammersly-Clifford theorem, a key result of MRF theory, to represent this property:

$$P(f) \propto e^{-\sum_C V_C(f)}. \tag{2}$$

Clique potential V_C describes the prior probability of a particular realisation of the elements of the clique C. For our neighbourhood system, MRF theory defines cliques as pairs of horizontally or vertically adjacent pixels. Eq. 2 reduces to:

$$P(f) \propto e^{-\sum_p \sum_{q \in N_p} V_{p,q}(f_p, f_q)}. \tag{3}$$

In accordance with [6], we assign clique potentials using the *Generalised Potts Model* where clique potentials resemble a well with depth u:

$$V_{p,q}(f_p, f_q) = u_{p,q} \cdot (1 - \delta(f_p - f_q)), \tag{4}$$

where δ is the unit impulse function. Clique potentials are isotropic ($V_{p,q} = V_{q,p}$), so $P(f)$ reduces to:

$$P(f) \propto e^{-\sum_{\{p,q\} \in \varepsilon_N} \begin{cases} 2u & \forall f_p \neq f_q, \\ 0 & otherwise. \end{cases}} \tag{5}$$

V_C can be interpreted as a cost of discontinuity between neighbouring pixels p, q. In practice, we assign the clique potentials according to how continuous the image is over the clique using the Gaussian function:

$$V_c = \frac{e^{-(\Delta I_C)^2}}{2\sigma^2}, \tag{6}$$

where ΔI_C is the change in intensity across the clique, and σ is selected such that 3σ approximates the minimum intensity variation that is considered smooth.

Note that at this stage we have looked at one image independently of the other. Stereo properties have not been considered in constructing the prior term.

Likelihood $P(O \mid f)$

This term describes how likely an observation O matches a hypothesized configuration f and involves incorporating stereo information for assessing how well the observed images fit the configuration field. It can be equivalently represented as:

$$P(O \mid f) = P(I_A \mid f, I_B), \tag{7}$$

where I_A is the primary image and I_B the secondary (chosen arbitrarily) and f is the hypothesized configuration field. In terms of image sites P (pixels), Eq. 7 becomes:

$$P(O \mid f) \propto \prod_P g(i_A, i_B, l_P), \tag{8}$$

where $g()$ is some symmetric function [6] that describes how well label l_P fits the image evidence $i_A \in I_A$ and $i_B \in I_B$ corresponding to site P (it could, for instance, be a Gaussian function of the difference in observed left and right image intensities at P; we evaluate this instance – Eq. 11 – and propose alternatives later).

Energy minimisation

We have assembled the terms in Eq. 1 necessary to define the MAP optimisation problem:

$$P(f \mid O) \propto e^{-\sum_p \sum_{q \in N_p} V_{p,q}(f_{p,f_q})} \cdot \prod_P g(i_A, i_B, l_P). \tag{9}$$

Maximising $P(f \mid O)$ is equivalent to minimising the energy function:

$$E = \sum_p \sum_{q \in N_p} V_{p,q}(f_{p,f_q}) - \sum_P ln(g(i_A, i_B, l_P)). \tag{10}$$

Optimisation

A variety of methods can be used to optimise the above energy function including, amongst others, *simulated annealing* and *graph cuts*. For active vision, high-speed performance is a priority. At present, a graph cut technique is the preferred optimisation technique, and is validated for this class of optimisation as per [18]. We adopt the method used in [17] for MAP stereo disparity optimisation (we omit their use of *α–expansion* as we consider a purely binary field). In this formulation, the problem is that of finding the *minimum cut* on a *weighted graph*:

A weighted graph G comprising of vertices V and edges E is constructed with two distinct terminals l_{zd}, l_{nzd} (the source and sink). A cut $C = V^s, V^t$ is defined as a partition of the vertices into two sets $s \in V^s$ and $t \in V^t$. Edges t, s are added such that the cost of any cut is equal to the energy of the corresponding configuration. The cost of a cut $|C|$ equals the sum of the weights of the edges between a vertex in V^s and a vertex in V^t.

The goal is to find the cut with the smallest cost, or equivalently, compute the *maximum flow* between terminals according to the Ford Fulkerson algorithm [12]. The minimum cut yields the configuration that minimises the energy function. Details of the method can be found in [17]. It has been shown to perform (as worst) in low order polynomial time, but in practice performs in near linear time for graphs with many short paths between the source and sink, such as this [18].

Robustness

We now look at the situations where the ZDF performs poorly, and provide methods to combat these weaknesses. Fig. 5a shows ZDF output for typical input images where the likelihood term has been defined using intensity comparision. Output was obtained at approximately $25Hz$ for the 60x60 pixel fovea on a standard $3GHz$ single processor PC. For this case, $g()$ in Eq. 8 has been defined as:

$$g(i_A, i_B, f) = \begin{cases} \frac{e^{-(\Delta I_C)^2}}{2\sigma^2} & \forall f = l_z \\ 1 - \frac{e^{-(\Delta I_C)^2}}{2\sigma^2} & \forall f = l_{nz} \end{cases} \tag{11}$$

The variation in intensity at corresponding pixel locations in the left and

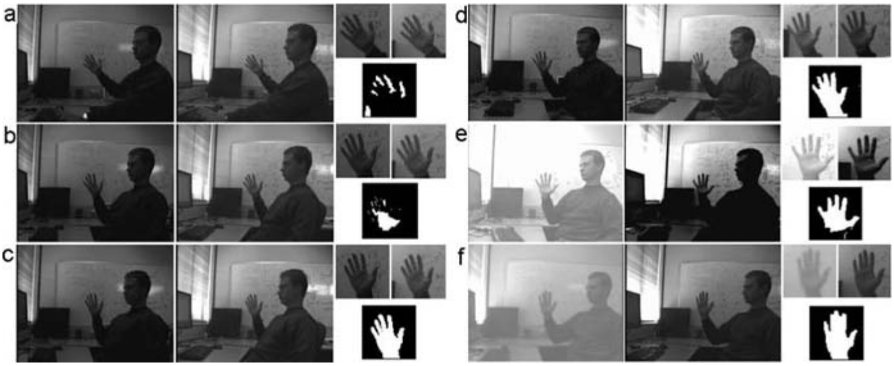

Fig. 5. Foveal perception. The left and right images and their respective foveas are shown with ZDF output (bottom right) for each case *a-f*. Result *a* involves intensity comparision, *b* involves NCC, and *c* DOG NCC for typical image pairs. Result *d-f* show NDT output for typical images *d*, and extreme conditions *e,f*.

right images is significant enough that the ZDF has not labeled all pixels on the hand as being at zero disparity. To combat such variations, NCC is instead used (Fig. 5b). Whilst the ZDF output improved slightly, processing time per frame was significantly increased ($\sim 12Hz$). As well as being slow, this approach requires much parameter tuning. Bland regions return a high correlation whether they are at zero disparity or not, and so the correlations that return the highest results cannot be trusted. A threshold must be chosen above which correlations are disregarded, which also has the consequence of disregarding the most meaningful correlations. Additionally, a histogram of correlation output results is not symmetric (Fig. 7, left). There is difficulty in converting such output to a probability distribution about a 0.5 mean, or converting it to an energy function penalty.

To combat the thresholding problem with the NCC approach, the images can be pre-processed with a DOG kernel. The output using this technique (Fig. 5c) is good, but is much slower than all previous methods ($\sim 8Hz$) and requires yet more tuning at the DOG stage. It is still susceptible to the problem of non-symmetric output.

We prefer a comparator whose output histogram resembles a symmetric distribution, so that these problems could be alleviated. For this reason we chose a simple *neighbourhood descriptor transform* (NDT) that preserves the relative intensity relations between neighbouring pixels, but is unaffected by brightness or contrast variations between image pairs.

In this approach, we assign a boolean descriptor string to each site and then compare the descriptors. The descriptor is assembled by comparing pixel intensity relations in the 3x3 neighbourhood around each site (Fig. 6). In its simplest form, for example, we first compare the central pixel at a site in the primary image to one of its four-connected neighbours, assigning a '1' to the descriptor string if the pixel intensity at the centre is greater than that of its northern neighbour and a '0' otherwise. This is done for its southern, eastern and western neighbours also. This is repeated at the same pixel site in the secondary image. The order of construction of all descriptors is necessarily the same. A more complicated descriptor would be constructed using more than merely four relations[6]. Comparison of the descriptors for a particular site is trivial, the result being equal to the sum of entries in the primary image site descriptor that match the descriptor entries at the same positions in the string for the secondary image site descriptor, divided by the length of the descriptor string.

Fig. 7 shows histograms of the output of individual neighborhood comparisions using the NCC DOG approach (left) and NDT approach (right) over a series of sequential image pairs. The histogram of NDT results is a symmetric distribution about a mean of 0.5, and hence is easily converted to a penalty for the energy function.

Fig. 5d shows NDT output for typical images. Assignment and comparision of descriptors is faster than NCC DOG, ($\sim 25Hz$) yet requires no parameter tuning. In Fig. 5e, the left camera gain was maximised, and the right camera contrast was maximised. In Fig. 5f, the left camera was defocussed and saturated. The output remained good under these artificial extremes.

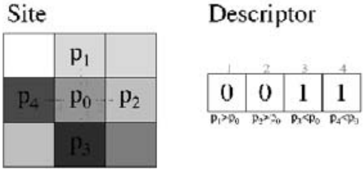

Fig. 6. NDT descriptor construction, four comparisons.

3.3 Bimodal Results

Fig. 8 shows a snapshot of output of the foveated and peripheral perception modes operating in parallel. The coarse peripheral perception detects mass near the (arbitrary) point of gaze fixation. Then the foveal response ensures gaze fixation occurs on an object or mass by zeroing disparity on peripherally detected mass closest to the gaze fixation point. By adjusting the camera geometry, the system is able to keep the object at zero disparity and centred within the foveas. Bimodal perception operates at approximately $15Hz$ without optimisation (threading and MMX/SSE improvements are expected).

[6] Experiment has shown that a four neighbour comparator compares favorably (in terms of trade-offs between performance and processing time) to larger descriptors.

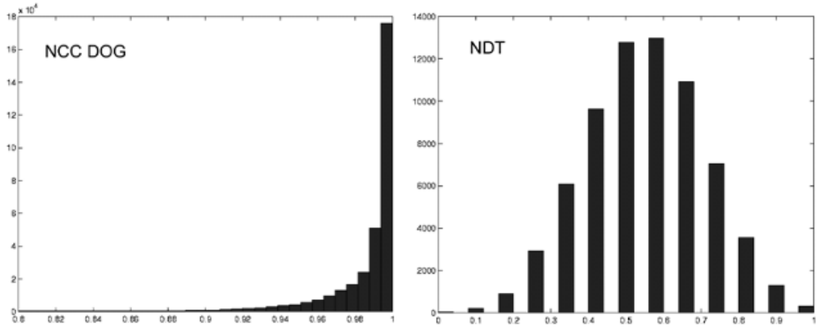

Fig. 7. Histograms of individual NCC DOG (left) and NDT (right) neighborhood comparisions for a series of observations.

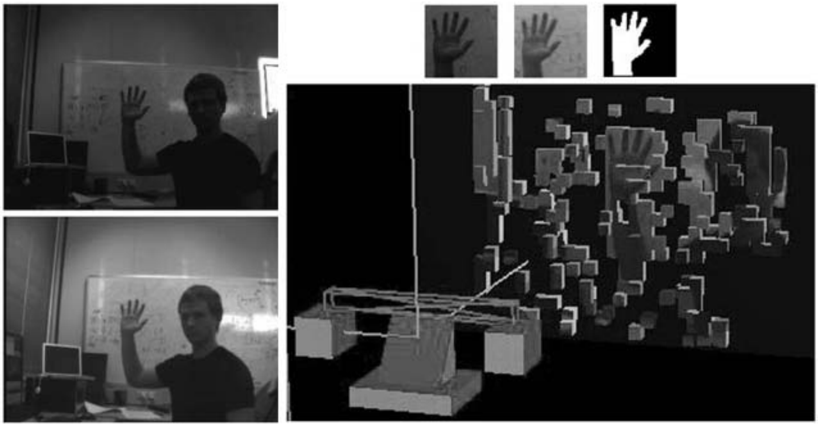

Fig. 8. Bimodal operation. Left: left (top) and right (bottom) input images. Right: Foveal perception (top) and peripheral perception (bottom). Foveal segmentation enhances the coarse perception of mass in the scene.

4 Conclusion

A bimodal active vision system has been presented. The peripheral mode fused actively acquired depth data into a 3D occupancy grid, operating at approximately $20Hz$. The foveal mode provides coordinated stereo fixation upon mass/objects in the scene. It also enables pixel-wise extraction of the object or region of mass upon which fixation occurrs using a maximum a-posterior zero disparity filter. The foveal response operates at around $25Hz$. Bimodal perception operates at approximately $15Hz$ on the $3GHz$ single processor PC.

Obtaining a peripheral awareness of the scene and extracting objects within the fovea permits experimentation in fixation and gaze arbitration. Prioritised monitoring of objects in the scene is the next step in our work towards artificial scene awareness.

References

1. J. Aloimonos, I. Weiss, and A. Bandyopadhyay, "Active vision," in *IEEE Int. Journal on Computer Vision*, 1988.
2. N. Apostoloff and A. Zelinsky, "Vision in and out of vehicles: Integrated driver and road scene monitoring," *IEEE Int. Journal of Robotics Research*, vol. 23, no. 4, 2004.
3. R. Bajczy, "Active perception," in *IEEE Int. Journal on Computer Vision*, 1988.
4. D. Ballard, "Animate vision," in *Artificial Intelligence*, 1991.
5. J. Banks and P. Corke, "Quantitative evaluation of matching methods and validity measures for stereo vision," *IEEE Int. Journal of Robotics Research*, vol. 20, no. 7, 1991.
6. Y. Boykov, O. Veksler, and R. Zabih, "Markov random fields with efficient approximations," Computer Science Department, Cornell University Ithaca, NY 14853, Tech. Rep. TR97-1658, 3 1997.
7. A. Dankers, N. Barnes, and A. Zelinsky, "Active vision - rectification and depth mapping," in *Australian Conf. on Robotics and Automation*, 2004.
8. ——, "Active vision for road scene awareness," in *IEEE Intelligent Vehicles Symposium*, 2005.
9. A. Dankers and A. Zelinsky, "Driver assistance: Contemporary road safety," in *Australian Conf. on Robotics and Automation*, 2004.
10. A. Elfes, "Using occupancy grids for mobile robot perception and navigation," *IEEE Computer Magazine*, 6 1989.
11. L. Fletcher, N. Barnes, and G. Loy, "Robot vision for driver support systems," in *IEEE Int. Conf. on Intelligent Robots and Systems*, 2004.
12. L. Ford and D. Fulkerson, *Flows in Networks.* Princeton University Press, 1962.
13. S. Geman and D. Geman, "Stochastic relaxation, gibbs distributions, and the bayesian restoration of images," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1984.
14. G. Grubb, A. Zelinsky, L. Nilsson, and M. Rilbe, "3d vision sensing for improved pedestrian safety," in *IEEE Intelligent Vehicles Symposium*, 2004.
15. R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision, Second Edition.* Cambridge University Press, 2004.
16. S. Kagami, K. Okada, M. Inaba, and H. Inoue, "Realtime 3d depth flow generation and its application to track to walking human being," in *IEEE Int. Conf. on Robotics and Automation*, 2000.
17. V. Kolmogorov and R. Zabih, "Multi-camera scene reconstruction via graph cuts," in *Europuan Conf. on Comupter Vision*, 2002.
18. ——, "What energy functions can be minimized via graph cuts?" in *Europuan Conf. on Comupter Vision*, 2002.
19. G. Loy and N. Barnes, "Fast shape-based road sign detection for a driver assistance system," in *IEEE Int. Conf. on Intelligent Robots and Systems*, 2004.
20. N. Pettersson and L. Petersson, "Online stereo calibration using fpgas," in *IEEE Intelligent Vehicles Symposium*, 2005.
21. E. Schwartz, "A quantitative model of the functional architecture of human striate cortex with application to visual illusion and cortical texture analysis," in *Biological Cybernetics*, 1980.
22. H. Truong, S. Abdallah, S. Rougeaux, and A. Zelinsky, "A novel mechanism for stereo active vision," in *Australian Conf. on Robotics and Automation*, 2000.