Active Vision for Road Scene Awareness

Andrew Dankers*, Nick Barnes†, Alex Zelinsky‡
*Research School of Information Sciences and Engineering
Australian National University
Canberra ACT Australia 0200
andrew.dankers@rsise.anu.edu.au
†National ICT Australia[1]
Locked Bag 8001
Canberra ACT Australia 2601
nick.barnes@nicta.com.au
‡CSIRO ICT Centre
Canberra ACT Australia 0200
alex.zelinsky@csiro.au

***Abstract*— We present a mapping approach to road scene awareness based on active stereo vision. We generalise traditional static multi-camera rectification techniques to enable active epipolar rectification with a mosaic representation of the output. The approach is used to apply standard static depth mapping and optical flow techniques to the active case. We use the framework to extract the ground plane and segment moving objects in dynamic scenes using arbitrarily moving cameras on a moving vehicle. The approach enables an estimation of the velocity of the vehicle relative to the road, and the velocity of objects in the scene. We provide footage of preliminary results of the system operating in real-time, including dynamic object extraction and tracking, ground plane extraction, and recovery of vehicle velocity.**

I. Introduction

The concept of safety through prevention and preparedness is emerging as an automotive industry philosophy. The *Smart Car project*, a collaboration between the Australian National University, National ICT Australia, and the CSIRO, focusses its attention 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 [1]. The detection of objects on the road such as signs [2] and pedestrians [3], and the location of the road itself [4], form part of the set of observable road scene events that we would like 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 solely on the use of active computer vision as a scene sensing input to the driver assistance architecture. In particular, we focus on object and ground plane detection for subsequent use with higher level classification processes such as pedestrian, vehicle, and sign detection.

Stereo vision has become a viable sensor for obtaining three-dimensional range information [5]. Traditionally, stereo sensors have used fixed geometric configurations effective in obtaining range estimates for regions of relatively static scenes. In reducing processor expense, most depth-mapping algorithms match pixel locations in separate camera views within a small disparity range, e.g. ±32 pixels. Consequently, depth-maps obtained from static stereo configurations are often dense and well populated over portions of the scene around the fixed horopter, but they are not well suited to dynamic scenes or tasks that involve resolute depth estimation over larger scene volumes.

A visual system able to adjust its visual parameters to aid task-oriented behaviour – an approach labeled *active* [6] or *animate* [7] vision – can offer impressive computational benefits for scene analysis in realistic environments [8]. By actively varying the camera geometry it is possible to place the horopter and/or vergence point over any of the locations of interest in a scene and thereby obtain maximal depth information about those locations. Where a subject is moving, the horopter can be made to follow the subject such that information about the subject is maximised. Varying the camera geometry not only improves the resolution of range information about a particular location, but by scanning the horopter, it can also increase the volume of the scene that may be densely depth-mapped. Figure 1 shows how the horopter can be scanned over the scene by varying the camera geometry for a stereo configuration. This approach is potentially more efficient and useful than static methods because a small disparity range scanned over the scene is less computationally expensive and obtains more dense results than a single, unscannable, but large disparity range from a static configuration.

In [9], we developed a framework for using existing static multiple-camera algorithms, such as depth-mapping, on active multi-camera configurations. The approach involved the use of a three-dimensional occupancy grid for integrating range information using active stereo. We have since improved and extended this framework with the goal of detecting, segmenting and localising objects in the road-scene. We show how the framework enables 3D mass flow in the scene to be calculated from standard optical flow algorithms on an

[1] 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.

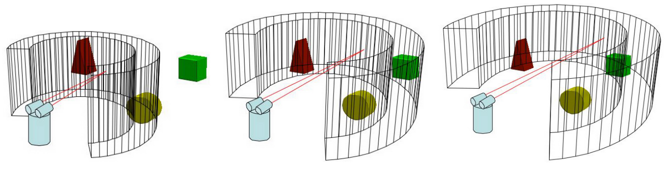

Fig. 1. Scanning the horopter over the scene: The locus of zero disparity points defines a plane known as the horopter. For a given camera geometry, searching for pixel matches between left and right stereo images over a small disparity range defines a volume about the horopter. By varying the geometry, this searchable volume can be scanned over the scene. In the first frame, only the circle lies within the searchable region. As the horopter is scanned outwards by varying the vergence point, the triangle, then the cube become detectable.

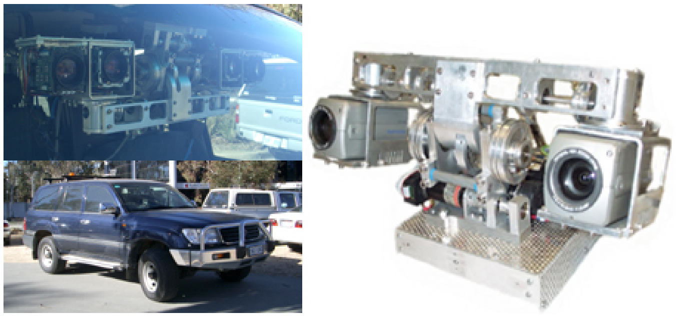

Fig. 2. The research platform. Left: The *Smart Car* (bottom) and *CeDAR* mounted behind the windscreen (top); Right: CeDAR.

active platform because the active rectification process we have developed inherently removes the effect of camera motion. We use this inherent benefit of the active rectification process to determine the velocity of objects detected in the scene relative to the vehicle and/or road at high frame rates. It is difficult to depict the achievements of an active vision system in a printed format, and for this reason, we have provided footage of the system and its components in operation (see Section X).

II. Outline

We begin by presenting the research platform (Section III) and reviewing our background work on active rectification [9] (Section IV) and occupancy grid representation of the scene (Section V).

We introduce a space variant grid representation of the scene (Section V-B), and methods to analyse the occupancy grid for ground plane and object extraction (Section VI).

We present our approach to extracting three dimensional scene motion (Section VII) and describe how the 3D flow of mass in the scene can be used to extract the motion of objects and the road while the cameras are in any geometric configuration, or even moving. We describe how we use this actively collated occupancy and velocity information to segment and track objects (Section VIII).

Finally, we conclude with a summary of our present work, and comment on future improvements and additions to the system (Section IX).

III. Research Platform

A 1999 Toyota Landcruiser 4WD is equipped with the appropriate sensors, actuators and other hardware to provide an environment in which desired driver assistance competencies can be developed [10]. Installed centrally inside the front windscreen is an advanced active stereo vision mechanism. CeDAR, the Cable-Drive Active-Vision Robot [11], 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 important kinematic property of the CeDAR mechanism is that the pan axes rotate about the optical centre of each camera, minimising kinematic translational effects to reduce complexity in the epipolar rectification process. Figure 2 shows the CeDAR platform as it is mounted in the Smart Car.

IV. Active Rectification

In [9] we described a rectification method used to actively enforce *parallel epipolar geometry* [12] using camera geometric relations, independent of the contents of the images. The camera relations may be determined by any number of methods. Visual techniques such as the SIFT algorithm [13] or Harris corner detection [14] can be used to identify features common to each camera view, and thereby infer the geometry. We use a fixed baseline and encoders to measure camera rotations. A combination of visual and encoder techniques could also be adopted to obtain the camera relationships to a more exacting degree - this is the focus of additional present work within the project [15].

The rectification process, an extension of similar work in [16], enables online epipolar rectification of the image sequences and the calculation of the shift in pixels between consecutive frames from each camera, and between the current frames from the left and right cameras (in the case of a stereo rig, though any number or configuration of cameras can be used as long as the translations and rotations between cameras are known). We have shown the effectiveness of the process by using it to create globally epipolar rectified mosaics of the scene as the cameras were moved. Figure 3 shows a snapshot of online output from the mosaic process for a single camera.

The active rectification process yields the relationship between the left and right camera view frames in terms of pixels, as well as rectifying the images. Therefore, all that needs to be done to obtain *absolute* disparities for a pair of concurrent images from any geometric configuration is to output a standard disparity map from the overlapping regions of the current rectified left and right images, and offset all disparities in the resulting disparity map by the pixel shift between the current left and right images, as calculated by the rectification process. It is then a simple matter to convert the absolute disparities to absolute depths (see [9]) to yield range data from active multi-camera configurations.

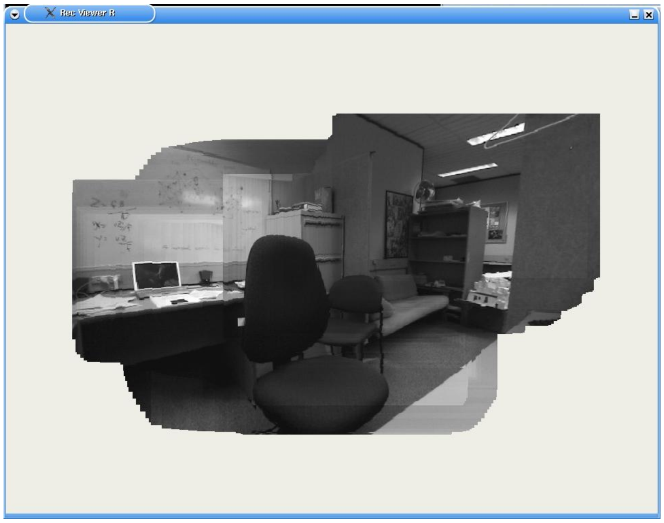

Fig. 3. Snapshot of online output of the active rectification process: mosaic of rectified frames from right CeDAR camera. The full video sequence is also available (Section X).

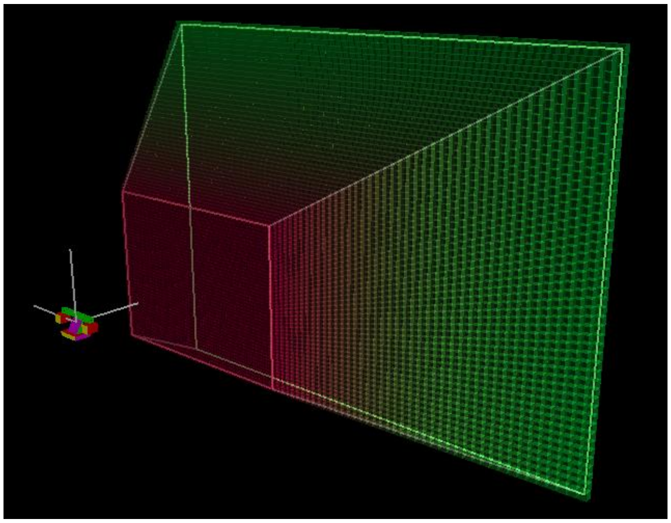

Fig. 4. Occupancy grid configuration

V. An Occupancy Grid Representation of the Scene

A. Summary of Previous Work

Traditionally, somewhat sparse and noisy stereo depth data has been used to judge the existence of mass at a location in the scene. Decisions based directly on such unfiltered data could adversely affect the sequence of future events reliant upon such a decision. In previous use of stereo range data, only a few attempts were made to strengthen or attenuate a belief in the location of mass in the scene [17]. 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 and detailed maps of a scene [18].

As well as addressing the above issues, an occupancy grid allows the integration of data according to a sensor model. Each pixel in the disparity map is considered as a single measurement for which a sensor model is used to fuse data into the occupancy grid. Not only is uncertainty in the measurements considered in the sensor model, but it is also partially absorbed by the granularity of the occupancy grid.

We integrate the active stereo depth map range data into a three-dimensional occupancy grid using a Bayesian approach [19]. Depth maps are produced using a processor economical SAD-based technique with difference of gaussian pre-filtering to reduce the effect of intensity variation [5]. Figures 5 and 6 show example occupancy grid output.

B. Space Variant Grid Representation and Ray Tracing

At farther scene depths, pixel disparities correspond to larger changes in scene depth. Accordingly, we have adopted an occupancy grid configuration that exhibits cell sizes that increase with depth. Cell cube edge lengths correspond to the effect of a specific amount of pixels of disparity at that depth. For example, 10 pixels of disparity at 1m scene depth yields a cell size much smaller than 10 pixels of disparity at 50m scene depth.

This approach significantly reduces the number of cells at larger depths where high depth resolution is not available anyway, improving processor performance. It also increases resolution in the grid at nearer depths where we are more interested in an accurate estimation of the location of objects.

The space variant occupancy grid formulation reduces ray tracing computations associated with sensor model integration of range measurements. At 1m depth, a slice through the occupancy grid contains a fixed number of cells in the horizontal direction and another fixed number of cells in the vertical direction. At any other depth, the number of vertical and horizontal cells in the slice are the same respectively, with the central cells aligned with the origin at the location of the sensor. This means that a ray emanating from the origin and passing through a cell at 1m with slice coordinates (x,y) also passes through all other slices at coordinates (x,y). This configuration means that ray-tracing through the occupancy grid for sensor integration becomes trivial. Figure 4 shows the construction of the occupancy grid and rays of cells emanating from the origin.

Updates to the grid occur at a frequency high enough for us to effectively analyze dynamic scenes through the use of a decay rate applied to all cell certainties. Propagation of uncertainties according to measured cell velocities would mean that the reliance upon high frequency updating and decay could be alleviated (uncertainty propagation is present work, see Section IX).

VI. Analysis of the Occupancy Grid

A. Ground Plane Extraction from the Occupancy Grid

Our approach to ground plane extraction is similar to that of a *v-disparity* analysis [20]. Essentially, we look at the occupancy grid from the side to produce an accumulator image

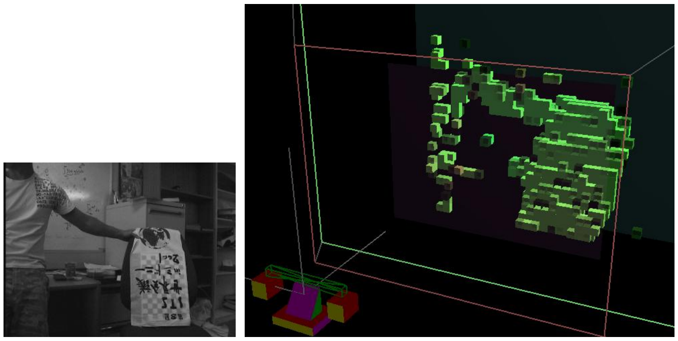

Fig. 5. Online snapshot of occupancy grid construction. Left: view from left camera. Right: occupancy grid; the two semi-transparent vertical planes show the near and far bounds on the region in which a depth response is possible, given the current camera geometry and disparity search range (±16 pixels).

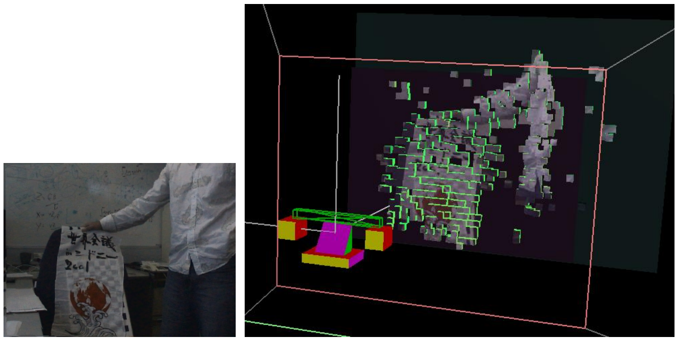

Fig. 6. Online snapshot of occupancy grid construction with colour texturing of cell front faces.

plotting the number of occupied cells at a particular height and depth in the occupancy grid. A Hough transform [21] is then used to find the most dominant line in this density side view of the occupancy grid. We search for the line within reasonable bounds of where the road is likely to be to reduce the computational expense of the Hough transform. In this manner, we are able to extract a planar approximation to the location of the ground plane in terms of altitude and attitude. The assumption is that the sensor is, under most circumstances, situated at a roll angle that is parallel to the road, and that the road is planar, so that we do not consider the roll of the road relative to the sensor. The granularity of the occupancy grid is such that small violations of this assumption are somewhat absorbed. Any systematic misalignment can be removed by calibration. Figure 8 shows an image from the online output of the occupancy grid, including location of the ground plane.

B. Object Segmentation from the Occupancy Grid

An *object* in the occupancy grid is considered to be a group of 26-connected cells located above the ground plane (see Figure 10). We use a 3D raster scan to uniquely label connected components.

VII. 3D Scene Flow

The velocities of cells in the occupancy grid are calculated using an approach similar to that of [22]. First, optical flow in each of the camera view frames is calculated to determine the x and y components of cell flow. Next, consecutive depth-maps are subtracted to provide the z component. Because we are calculating flow in image space, we are able to assign sub-cell sized motions per frame (velocities) to cells in the occupancy grid.

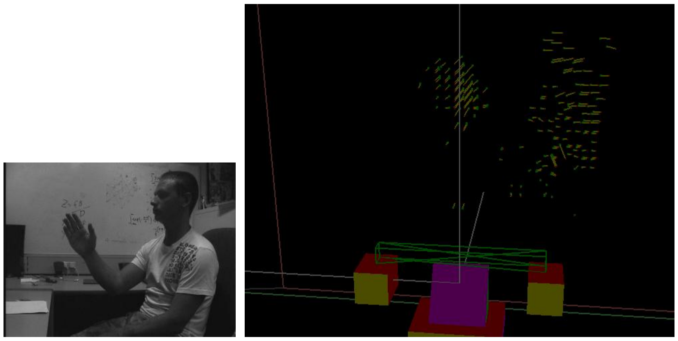

Fig. 7. A snapshot from an online sequence showing 3D vectors representing mass flow in the scene. The complete video sequence is available for viewing (Section X).

Optical flow between consecutive images from a camera is calculated using a simple, but fast flow estimation. The rectification and mosaic construction process described earlier allows the removal of camera rotations and translations so that a standard SAD-based flow estimation can be used [5]. As the location of the current and previous frame in the mosaic from a single camera is known, we calculate optical flow on the overlapping region of consecutive frames in the mosaic. In the same manner, the overlapping regions of consecutive depth-maps are subtracted to obtain the depth flow in the z direction [22].

The flow vectors are binned into bin sizes corresponding to the image frame width of a cell in the occupancy grid. The average flow for each bin is then determined. For example, if the cell sizes in the occupancy grid were chosen such that the edge length of occupancy grid cubes corresponded to 10 pixels of disparity at a particular depth (as described earlier), then adjacent bins would contain the average of flow vectors for adjacent 10 by 10 regions in the flow images.

Having determined the x,y and z components of each bin-sized region of pixels in the image, these components are then overlaid upon the occupancy grid using ray tracing. We look along a ray in the occupancy grid that corresponds to a particular bin of pixels in the image until an occupied cell is found. The velocity components of the bin are then assigned to that occupied cell. Figure 7 shows online output of 3D flow.

A. Object Segmentation from Scene Flow

After an occupancy grid – and subsequently a flow grid – has been calculated, we segment objects in the flow grid in a manner similar to that of the occupancy grid. A 3D raster scan labels adjacent 26-connected cells whose velocities are similar. We use information about the location of the ground plane from the previous step to limit the search for objects to the region above the ground plane. Essentially, if a cell has a

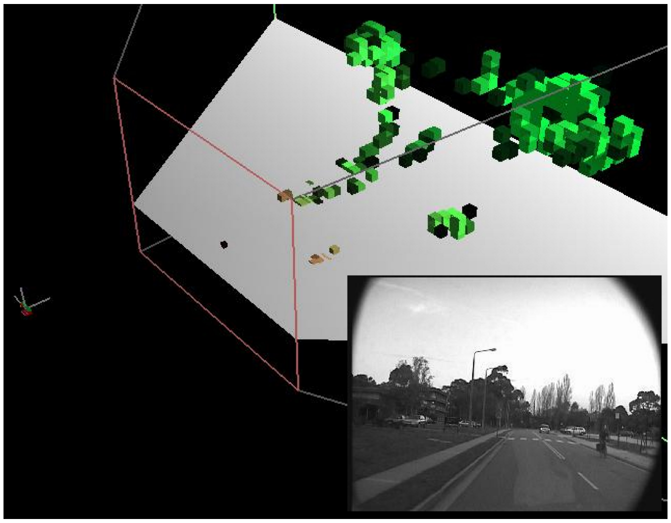

Fig. 8. Online ground plane extraction from the occupancy grid. Inset shows view from left camera. Detection of the cyclist, light pole, and trees in the background are also evident on the occupancy grid. The complete video sequence is available for viewing (Section X).

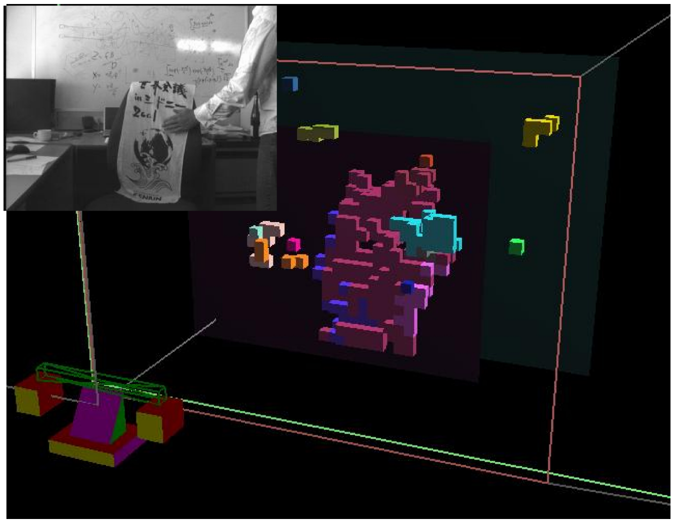

Fig. 10. Online object segmentation. Inset shows image acquired by right camera. The hand (blue) is segmented from the chair (burgundy) because it is moving. The complete sequence is available for viewing (Section X).

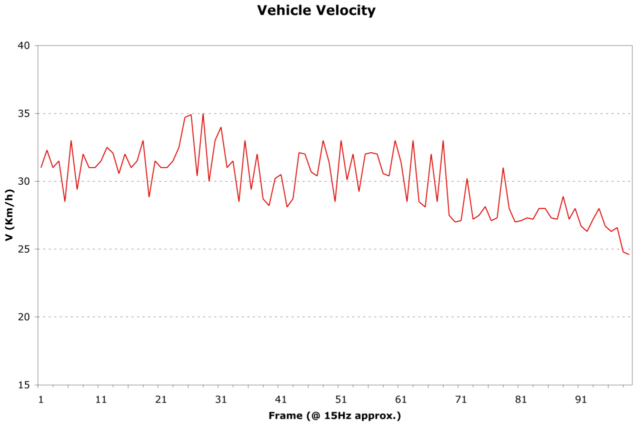

Fig. 9. Vehicle velocity according to unfiltered 3D flow data.

non-zero flow assigned to it, and its velocity is not significantly different to an adjacent cell with a non-zero velocity, it is assigned the same unique object identity as that cell.

B. *Vehicle Motion from Scene Flow*

We wish to infer the motion of the vehicle relative to the road from an analysis of the flow grid. Preferrably, the analysis would not consider regions of the scene that are likely to be moving in a manner dissimilar to that of the road. Hence, we only consider regions in the vicinity of the ground plane to extract the vehicle velocity. Histograms of the velocity components of all the cells adjacent to the previously detected ground plane are constructed. At present, we use the histogram mean and associated 95% confidence interval as a measurement of the vehicle velocity. Once the velocity of the vehicle relative to the road has been calculated, we can remove the velocity of the vehicle from calculations of the velocity of objects in the scene.

Figure 9 shows a preliminary plot of the vehicle velocity as determined by unfiltered 3D flow data. Only the flow in the z-direction (directly towards the cameras) is considered. Although the velocity of the vehicle was not logged, the fluctuation of the velocity about 30km/h fits well with the fact that the vehicle was driving in a designated 40km/h zone on the ANU campus. The data velocity from flow was determined for the same sequence of footage as that shown in Figure 8.

VIII. Tracking Objects

Tracking an object initially involves finding object correspondences in consecutive frames. Data associated with each object in the occupancy grid includes its mass and centre of gravity. Objects segmented in the flow grid also include additional velocity information. By considering each object in the current frame and comparing the data associated with it to each object in the previous frame, we are able to determine likely object correspondences over time. The velocity information enables us to distinguish, for example, a person moving in front of a parked car. Fig 10 shows a hand that has been segmented from a chair using velocity information, despite being in contact with the chair and being labelled as the same object in the occupancy grid segmentation.

IX. Conclusion

A. *Summary*

We have presented a method for active epipolar rectification that has been shown to allow static stereo algorithms – such as disparity mapping and optical flow – to operate on an active stereo platform. We have presented a framework for active depth mapping with a robot-centered occupancy grid representation for data integration and filtering. We have been able to analyze the 3D occupancy grid representation of the scene to extract objects and an estimate of the location and attitude of the road.

We have incorporated image based calculations of scene flow into the grid representation of the scene to enable sub-cell velocity estimations of mass detected in the scene. We are able to use this flow information to further segment objects in the scene, and to estimate the velocity of the vehicle relative to the road.

We are able to detect and segment moving or stationary objects in dynamic scenes (see Figure 10 and corresponding video footage) using moving cameras on a moving platform. Preliminary, unfiltered results show promise for the approach with increased accuracy and efficiency expected to come.

B. Future Work

Presently, our approach has not incorporated methods to filter parameters important to the rectification process. Noisy angular readings are to be refined by a combination of filtration and integration with image-based methods to extract the camera geometry to a more exacting degree. Already, work near completion involves the incorporation of SIFT techniques to extract angles. In addition to improved geometry calculations, the location and velocity of the ground plane and objects detected in the scene are to be filtered rather than used at face value.

Certainties are currently integrated into the occupancy grid from observations of the current scene state with the use of a decay rate to reduce these certainties over time to allow the analysis of dynamic scenes. Since we determine probability estimates of the current location of mass in the scene, as well as the likely velocity of that mass in three dimensions, we speculate that propagation of uncertainties within the occupancy grid according to their estimated velocities will yield increased certainty in the location of mass, reduce the reliance an decay rate.

In addition to improving the accuracy of the system, future work involves its use as an input for road scene understanding. At present, objects are segmented from the scene using the 3D occupancy and flow approach, and we are already able to map these segmented objects back into image space to obtain their image frame profile. Objects are thus extracted from their surroundings and background. This extraction process is expected to assist higher level tasks such as classification of objects in the scene.

The ability to obtain an awareness of the scene regardless of the geometry of the active head is the first step towards experiments in fixation and gaze arbitration. Our approach enables information about the scene to be gathered regardless of the geometry of the cameras, and for an increased, scannable volume of the scene. Actively tracking objects and retaining an awareness of other objects in the scene such that attention can be shifted between regions of the scene according to priority is the next step in our work towards artificial vision. Our work enables such gaze arbitration experimentation in realistic environments.

X. Footage

It is difficult to depict the achievements of an active vision system in a printed format. For this reason, we have published footage of the system in operation at:

http://www.rsise.anu.edu.au/∼andrew/ivs05

XI. Acknowledgments

The authors would like to thank the significant contributions made by all members of the Smart Cars project. In particular, we thank Niklas Pettersson, Lars Petersson and Luke Fletcher for their direct contributions to this paper.

References

[1] L. Fletcher, N. Barnes, and G. Loy, “Robot vision for driver support systems,” in *International Conference on Intelligent Robots and Systems*, 2004.

[2] G. Loy and N. Barnes, “Fast shape-based road sign detection for a driver assistance system,” in *International Conference on Intelligent Robots and Systems*, 2004.

[3] G. Grubb, A. Zelinsky, L. Nilsson, and M. Rilbe, “3d vision sensing for improved pedestrian safety,” in *Intelligent Vehicles Symposium*, 2004.

[4] N. Apostoloff and A. Zelinsky, “Vision in and out of vehicles: Integrated driver and road scene monitoring,” in *Intelligent Transport Systems*, 2003.

[5] J. Banks and P. Corke, “Quantitative evaluation of matching methods and validity measures for stereo vision,” in *International Journal of Robotics Research*, 1991.

[6] J. Alomoinos, I. Weiss, and A. Bandopadhay, “Active vision,” in *International Journal on Computer Vision*, 1988.

[7] D. Ballard, “Animate vision,” in *Artificial Intelligence*, 1991.

[8] R. Bajczy, “Active perception,” in *International Journal on Computer Vision*, 1988.

[9] A. Dankers, N. Barnes, and A. Zelinsky, “Active vision - rectification and depth mapping,” in *Australian Conference on Robotics and Automation*, 2004.

[10] A. Dankers and A. Zelinsky, “Driver assistance: Contemporary road safety,” in *Australian Conference on Robotics and Automation*, 2004.

[11] H. Truong, S. Abdallah, S. Rougeaux, and A. Zelinsky, “A novel mechanism for stereo active vision,” in *Australian Conference on Robotics and Automation*, 2000.

[12] R. Hartley and A. Zisserman, “Multiple view geometry in computer vision, second edition,” in *Cambridge University Press, ISBN: 0521540518*, 2004.

[13] S. Se, D. Lowe, and J. Little, “Vision-based mobile robot localization and mapping using scale-invariant features,” in *International Conference on Robotics and Automation*, 2001.

[14] C. Harris and M. Stephens, “A combined corner and edge detector,” in *Alvey Vis Conference*, 1999.

[15] N. Pettersson and L. Petersson, “Online stereo calibration using fgpas (submitted),” in *Intelligent Vehicles Symposium*, 2005.

[16] A. Fusiello, E. Trucco, and A. Verri, “A compact algorithm for rectification of stereo pairs,” in *Machine Vision and Applications*, 2000.

[17] H. Moravec, “Robot spatial perception by stereoscopic vision and 3d evidence grids,” in *Carnegie Mellon University, Pennsylvania*, 1996.

[18] A. Elfes, “Using occupancy grids for mobile robot perception and navigation,” in *Carnegie Mellon University, Pennsylvania*, 1989.

[19] H. Moravec, “Sensor fusion in certainty grids for mobile robots,” in *Carnegie Mellon University, Pennsylvania*, 1989.

[20] R. Labayrade, D. Aubert, and J. Tarel, “Real time obstacle detection on non flat road geometry through ‘v-disparity’ representation,” in *Intelligent Vehicle Symposium*, 2002.

[21] T. Tian and M. Shah, “Recovering 3d motion of multiple objects using adaptive hough transform,” in *International Conference on Pattern Analysis and Machine Intelligence*, 1997.

[22] S. Kagami, K. Okada, M. Inaba, and H. Inoue, “Realtime 3d depth flow generation and its application to track to walking human being,” in *International Conference on Robotics and Automation*, 2000.