Recent Advances in Motion Understanding

Steven S. Beauchemin[1], Ruzena Bajcsy[1], and John L. Barron[2]

[1]GRASP Laboratory
School of Engineering and Applied Science
University of Pennsylvania
Philadelphia PA 19104-6228
USA

[2] Department of Computer Science
The University of Western Ontario
London Canada
N6A 5B7

Abstract: *Probably the most ambitious goal of Computer Vision is to build the universal vision machine, capable of guiding itself through arbitrary environments, recognizing objects along its path and reaching its destination, wherever that might be. In spite of this elusive goal, Computer Vision is a field of research where enormous progress has been accomplished. In particular, the paradigm of Active Vision enables the research community to better address the questions related to vision applications such as autonomous navigation, visual attention, foveal and peripheral vision and real-time issues by embedding the properties of a visual system into sets of tasks to be performed and with extended control over viewing parameters, thus greatly reducing the inherent complexities of constructing a vision machine worth bearing the name. In this keynote address, we investigate the general role of motion within the paradigm of Active Vision. In particular, we examine recent results concerning motion and visual attention, motion and autonomous navigation, motion in foveal and peripheral visual areas, both in measurement and interpretation. We conclude by outlining current research areas and promising directions in Active Vision.*

Keywords: Image motion, optical flow, structure from motion, active vision, visual attention, minimalist vision.

1 Introduction

Motion is a central aspect of vision. Its perception, interpretation and use in visual tasks is critical for navigation, obstacle avoidance, and also in the action-perception cycle, where object or feature recognition, tracking and grabbing are inherent parts of this cycle.

Up until recently, most of the research efforts were directed towards a reconstructionist approach to vision. Independently from the expected tasks or behaviors of the seeing agent, the aim has been the reconstruction of perceived surfaces from spatiotemporal images, the recovery of dense optical flow fields and depth maps and so on. Not surprisingly, the difficulties of such approaches were identified early. It was recognized that computing optical flow accurately is a daunting task, that reconstructing a scene from optical flow requires noiseless motion data, that the absence of texture in spatiotemporal signal regions confuses most algorithms for motion and stereo correspondence, *etc.*

In other words, the hypotheses that were posed in order to perform these visual tasks were restrictive

and their frequent violations within spatiotemporal visual signals limited their usefulness to very narrow domains of applications. In addition, the problems were not only limited to the kinds of hypotheses that were posed but also how and where they were applied to the image signals.

However, it has been known for some time that biological seeing agents use visual input in strong relation with their behaviors and the tasks to be accomplished. Hence, vision in the biological world is used in an active and purposive manner. Those observations strongly influenced the research directions in Computer Vision. For instance, it is recognized that given a task or set of tasks to be performed by the seeing agent, the amount of information extraction from the spatiotemporal image signal can be significantly reduced. An example of this is to understand that the house fly, although capable of autonomous flight, does not have any object cognition whatsoever and still can navigate. Hence, is it necessary to densely and completely reconstruct the 3D environment to achieve autonomous navigation? The answer is obviously no, and this is exactly what is meant by Active Vision, that is to say, given a task, such as navigation, define the relevant visual signal events and determine how to find them within the signal and how to use them for guiding the navigation.

In this contribution, we investigate the general role of motion within the paradigm of Active Vision for autonomous navigation. We examine the contributions made by classical Computer Vision and how they are integrated in the concept of Active Vision for autonomous navigation.

2 Classical Computer Vision

The reconstructionist, or classical approach to Computer Vision aims at geometrically reconstructing the visual scene with various signal properties. For instance, the Structure from Motion paradigm [20, 26] uses optical flow to reconstruct surfaces, while other approaches such as Shape from Shading use the gradient information directly or stereopsis to recover dense depth maps. One common characteristic to all of these approaches is that they are examples of Passive Vision, in which no visually guided behavior is taken into account.

In addition, these various approaches, when used in isolation, have not yielded the expected results. It appeared that real image signals are varied and complex enough to make these reconstruction methods fail in most realistic situations [4]. As Aloimonos *et al.* [1] discuss, almost every problem in passive perception is difficult to solve, because of ill-posedness. Examples of such ill-posed problems are the computation of optical flow and the correspondence problem. It comes naturally then to ask the question as to what vision processes would gain if embedded into an active visual system [1].

2.1 Passive Vision

In a seminal contribution, Aloimonos *et al.* show that the various reconstructionist approaches of Passive Vision, such as shape from shading, shape from contour, shape from texture, structure from motion and optical flow become simpler in both theory and practice when embedded in an Active Vision framework. They define Active Vision by the active control of viewing parameters, which allows to remove the actual ill-posedness and to derive numerically more robust solutions. Concurrently, the potential benefits of Active Vision have been conjectured by Bajcsy [3].

Problem	Passive Vision	Active Vision
Shape from Shading	Ill-posed. Regularization needed. Non-linear and multiple solutions.	Well posed. Unique solution Stable
Shape from Contour	Ill-posed. Solution under restrictive assumptions.	Well posed. Unique solution for monocular and binocular.
Shape from Texture	Ill-posed. Requires assumptions about texture.	Well-posed. No assumption required.
Structure from Motion	Well-posed. Numerically unstable. Nonlinearities.	Well-posed and stable. Simple solutions.
Optical Flow	Ill-posed. Requires regularization.	Well-posed. Unique solution.

Table 1: The advantages of partially known viewing parameters on various, classically ill-posed Computer Vision problems [1].

3 Active Computer Vision

An Active Vision system is defined by its capacity to dynamically control viewing and camera parameters in relation with the visual task to accomplish. Generally, the parameters are position, focus, zoom, aperture, and vergence with two-camera systems. In addition, Active Vision encompasses the concept of attention and selective sensing.

Because of the problems related with obtaining rich and accurate 3D descriptions of the visual scene, the Active Vision paradigm prescribes the use of only those aspects of the visual signal that are directly relevant to the task to perform. This principle of economy enables the design and implementation of real-time applications. In addition, the control of the viewing parameters reduces the complexities of many of the classical computer vision problems.

3.1 Attention

Achieving a visual task such as autonomous navigation in a partially or entirely arbitrary environment requires the seeing agent to exhibit certain characteristics given the constraints of the current technology. For instance, the requirement of real-time dynamic scene analysis implies the notion of computational economy, through various means that are provided by the Active Vision paradigm.

For a spatiotemporal visual signal and a certain task, only certain aspects of that signal are of potential interest. This observation pertains to the concept of *attention*, which allows to draw processing power only to those aspects of a signal that are relevant to the task [2].

- **Focal Selection:** Signal analysis is focused only on those signal regions of interest for the task. In this case the restriction applies spatially, to a relatively small region of the field of view of the sensor, which we wish to obtain at a high resolution. The selection of a region of interest may be made independently from the location of the foveal axis (covert attention).
- **Motion Selection:** Given the geometry of the sensor, its motion parameter domain and the task at hand, some restrictions about perceptual motion may be formulated. These may take the form of directional constraints or motion being more relevant in some region of the visual field (the foveal

area, for instance).

- **Compression:** It is convenient to find representations that eliminate the inherent redundancy of information contained in a signal and that are well suited for subsequent processing stages. This compression may take the form of extraction of features, which are used in later stages of processing rather than the contents of the signal itself.
- **Generic Measurements:** These measurements constitute the sum of low-level image properties that are measured prior to any signal reduction and abstraction processes. They may be correlation, optical flow, depth analysis or any other basic measurement directly performed on the output of sensors.

The concepts such as focal and motion selection, compression and generic measurements, when applied jointly and appropriately, tend to significantly reduce the computational complexity of visual tasks. However, they are only meaningful when a defined task is at hand. Further, the characteristics of these tasks must be such that they are attainable within the intrinsic capability of the visual system.

Studies of attention mechanisms in the context of Computer Vision are recent [17, 19]. In particular Tsotsos *et al.* [24] recently proposed to model visual attention with selective tuning because of both its biological plausibility and computational utility. In their approach, attention is defined as being composed of a) the selection of a region of interest within the visual stimulus, b) the selection of feature dimensions and values of interest and c) the shifting from a selected area to the next in time.

3.2 Foveal Sensing

Foveal sensing is an important characteristic of biological vision systems. Such sensing provides the observer with a multi-resolution representation, from which selection can be accomplished. Multi-resolution sensors must be embedded into an active system in order to take advantage of their high resolution foveal area. Such devices provide a wide field of view and are very well suited for obstacle avoidance in active, autonomous navigation.

In such systems, the decrease in resolution is exponential from the foveal area to the peripheral area. As a consequence, the usual projective invariants are no longer valid and the sum of image processing techniques must also be redefined in order to appropriately operate in the multiresolution space of such sensors [2, 23].

3.3 Multi-Resolution Devices for Autonomous Navigation

Calibration techniques for multi-resolution sensors, in this case spherical lenses, have now begun to appear in the literature [23, 5] and experiments in autonomous navigation have been conducted with such devices. For instance, Shah and Aggarwal have defined a system for autonomous navigation based on a stereo pair of fish-eye lenses [22]. The navigation is limited to indoor, geometrically structured environments, in which attention is given to edges that are oriented horizontally or vertically in the 3D environment. Given this restricted form of line detection, the stereo correspondence problem is greatly simplified and the environmental surfaces can easily be reconstructed as planar patches. In addition, the wide field of view of fish-eye lenses (about 180 degrees) facilitates the rotary motion planning of the robot while navigating as the lenses sense more information than traditional pinhole cameras.

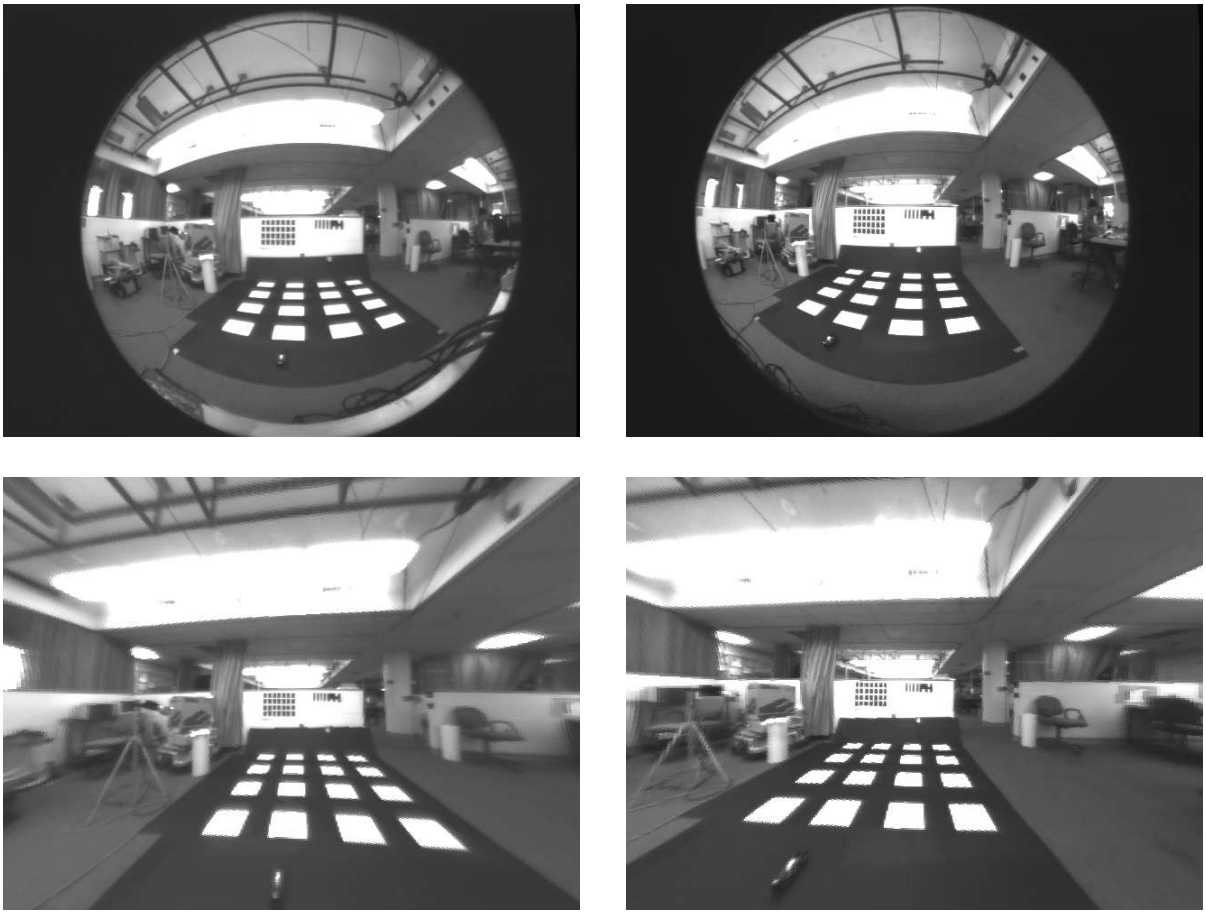

Figure 1: *A spherical stereo image pair (top) in which radial (barrel) distortion has been removed (bottom).*

However, it is pointed out that such lenses introduce a very significant amount of radial (and possibly tangential) distortion in imagery. It is thus crucial that the intrinsic and extrinsic parameters of such devices be known very accurately (see Figure 1) [23]. Their autonomous navigation experiments have been performed indoors and the autonomous system successfully guided itself through narrow passages and sharp turns.

3.4 Minimalist Vision for Autonomous Navigation

Related to Active Vision is the concept of minimalist vision put forth by Herman, Coombs and Raviv in their vision and control scheme for road following and obstacle avoidance [21, 11]. In this scheme, an explicit attempt is made at performing the minimum of processing to recover the information critical for the task or behavior to perform or exhibit. The advantages brought by the minimalist approach are

- **Simplicity:** Since only 2D information is considered in the minimalist approach to navigation, fewer assumptions have to be posed since no explicit 3D reconstruction is attempted. Reconstruction of the visual environment usually involves hypotheses about surfaces, smoothness and reflectances. In addition, calibration is kept to a minimum of parameters and devices.
- **Low error rates:** By not extracting 3D information, the approach avoids the accumulation of errors which results from the cascade of algorithms that must be applied to the signal to compute that type of information.
- **Task-dependent:** Only relevant information is being considered. For road following, the relevant information might be the road boundary data while the rest of the signal can be ignored. Obviously various tasks require different types of information. However this does not appear as a limitation, as long as only one task is carried out at a time.
- **Efficient:** With the real time issue being critical, efficiency is definitely a requirement. The minimalist approach is efficient in the sense that fewer computations need to be performed in order to generate the appropriate behavior.
- **Context** The framework uses context to drive the types of expectations, that predict where in the signal features should be found. This also reduces the required amount of computation.

In their road following experiments, the tangent point of the road edge and the associated optical flow are the relevant information coming from the signal and proved sufficient to generate the appropriate motion commands. They reported successful autonomous runs of 40 kilometers at speeds varying between 50 and 75 kilometers per hour. Also, their obstacle avoidance system uses the optical flow divergence to estimate time to collision [11].

Another example in which the principles of Active Vision have been extremely useful for autonomous navigation is the inverse perspective scheme developed by Mallot *et al.* [15] for performing obstacle detection with optical flow only. In this obstacle avoidance scheme, the navigational surface is considered as planar and the mobile visual agent is bound to navigate onto it. Hence, the agent has three degrees of freedom; two rotational and one translational. In this context, an operative definition of an obstacle is any feature rising above the navigational plane. Thus, the attentional features are those regions where elevated points onto the navigational surface are found. As Mallot *et al.* point out, variations of optical flow can be due to perspective foreshortening and the 3D structure of the scene. The key contribution is that with the hypothesis of a planar navigational surface, it is possible to define a coordinate transform that eliminates the effects of perspective in optical flow. This so-called inverse perspective transformation maps the points in the image back onto the navigational plane. Hence, any image point that is not originally located on the plane is distorted by the transformation. The detection is performed by thresholding the optical flow field for finding the distorted elevated points. This technique is a prime example of the effectiveness of the Active Vision paradigm. That is to say, the task is defined as autonomous navigation and the environmental constraint is defined as planar navigational surfaces. The pre-attentive cues for obstacle avoidance are defined as the elevated objects detected with optical flow applied to inverse-perspective transformed images, which triggers an obstacle-avoidance response in the seeing agent. In addition Coombs *et al.* have also used flow divergence and peripheral flow in defining and experimenting with obstacle avoidance techniques [12].

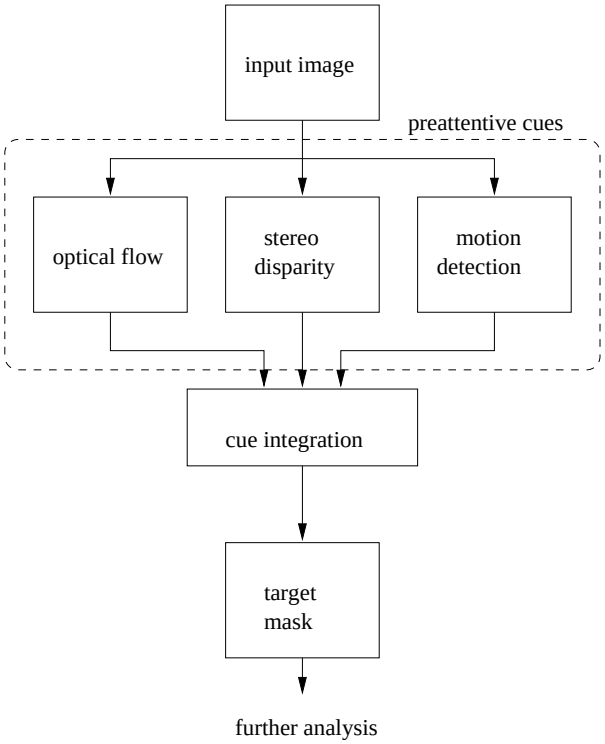

Figure 2: The preattentive cue integration scheme used by Maki [14].

3.5 Cue Integration and Selection

For an autonomous system functioning properly in a changing environment, reliance on various sources of information or cues, is essential [18]. In addition, Sensor fusion is an important concept in the field of robotics and has become known as cue intergration in the Computer Vision community. In Active Vision, the interest of fusing the outputs of various low level-mechanisms (optical flow, depth, etc) is related to the notion of pre-attentive processing of the visual signal. However, most of the research effort in cue integration and selection has been focused on reconstructing surfaces within the whole space of the visual signal rather than selecting areas of interest [10, 9, 8].

Cue selection, on the other hand, aims at utilizing only those cues that are available, reliable and of course related to the behavior to achieve. Among selected cues, fusion or integration can be performed in order to select areas of interest. There is thus a motivation for designing algorithms that can signal the availability and reliability of the computed cues. In addition, there is also a crucial need to find mechanisms that are adequate for resolving cue conflicts in a manner that allows the pursuit of the task at hand. In a normal situation where the cues do not contradict each other, some weighted averaging can take place. However when conflicts do arise, the resolution mechanism must be able to break the

averaging process in such a way that only the most reliable, non-contradicting cues will be fused.

Cues are many and will be defined differently depending on the tasks the autonomous system is to perform. For instance, the attentive stereo vision technique of Maki [14] uses fusion of binocular disparity, optical flow and motion detection to select salient areas from the input. These salient features are then located with a target mask which represents the location in the signal the system must attend to. However, cues can be defined in various ways. For example, if the attentional principle is depth-based, then cues to depth can be defined as binocular stereo, shading, and cues that inhibit perception of depth which are cues to flatness [8].

Although not yet fully understood, it has been shown that primate and human vision systems make an extensive use of feedback mechanisms [27]. In addition, it is believed that they reach the earliest stages of vision and that they are extremely important in biological vision [16]. Feedback mechanisms play an important role in cue selection and integration. The spatiotemporal evolution of cues within a region can be fed back into the various cue-computing modules in order to speed up and direct these computations. However, too strong a feedback loop will prevent the system from correcting previous estimation errors whereas too loose a coupling will not prove very useful in guiding the process [18].

3.6 Gaze Control

Gaze control plays a central role in Active Vision and is defined in the general sense as the alteration of imaging parameters to aid in the performance of visual tasks [2]. Thus, gaze control enables an Active Vision system to acquire images from different vantage points, those being selected to facilitate the accomplishment of the visual task. Generally, gaze control involves two broadly defined activities, namely gaze stabilization (or fixation) as is necessary in the visual tracking of independently moving objects and gaze change, as required for attending to attentional shifts.

In the case of gaze stabilization in fixation, the advantages are clear. The stabilization of a target or region of an image into the foveal area is essential to avoid motion blur of the image region. The motivations for gaze change are many. The agent might be seeking an advantageous parameter setting to solve some visual task or to gain computational robustness. In general, gaze control for fixation and gaze changes allow a) to stabilize image regions of interest, b) to increase the field of view and, c) to segment the independently moving objects from the scene.

3.6.1 Fixation and Visual Motion

A prime example in which gaze control reduces the complexity of classical vision problems such as egomotion is given by Fermuller and Aloimonos [13]. Their formulation of the fixation problems allowed them to reformulate the problems of 3D motion estimation, egomotion and time to collision in an efficient manner. An active agent in control of its fixation gaze can use image intensity derivatives (or normal flow and therefore no correspondence to solve for) as input to address the two perceptually distinct problems of object motion and agent motion. The use of normal flow is correctly argued to be a better choice than full optical flow as the computation of the former is well-posed whereas the latter is ill-posed because of the regularization it requires. They show that over time, the use of normal motion to maintain fixation converges to estimates of the full flow in that image region. They recover the focus of expansion, time to collision and the 3D motion parameters of the agent in motion.

4 Conclusion

To say that classical, Passive Vision has had mitigated success would be untrue. In fact, many of the early vision modules used in Active Vision directly come from successful research in Classical Vision. Many times is the paradigm of Active Vision capable of dramatically simplifying such early vision modules. However, it is essential that research be pursued outside the Active Vision framework. For instance, fundamental discoveries concerning image motion are still made [6, 7] and for certain numerous discoveries of importance will be coming forward.

It is also only fair to point out the fundamental role played by 3D and 2D motion in the Active Vision paradigm. For instance, autonomous navigation as well as independently moving objects and gaze control induce relative image motion, and as a result, image motion is probably one of the most important cues for visual tasks involving mobile agents and active stereo heads. However, it is not the only one, and cue integration and fusion still represent important challenges in Computer Vision [14, 18].

4.1 Promising Directions

Our best hope for designing visually autonomous agents resides in pursuing the conjunction of research efforts in both Active and Passive Vision. Many computational theories and methods have appeared for solving classical problems such as Structure from Motion, Shape from Shading, *etc.* and their reformulation and adaptation for Active Vision is very successful in terms of removing ill-posedness, ensuring solution uniqueness and in bringing numerical stability to several Computer Vision algorithms.

Open research areas include but are not limited to a) efficient computational schemes for attention, b) cue integration and selection methods and, in particular possibly the integration on non-visual cues such as inertia and odometry [25], c) the understanding of visual feedback mechanisms in biological vision for their inclusion in the Active Vision paradigm [27], and d) the reformulation of pattern recognition and image processing techniques for multi-resolution visual devices such as wide-angle and fish-eye lenses [2].

In closing, it is also instructive to notice that vision in primates and humans is active and represents one of the evolutionary adaptations that higher mammals and other vertebrates have followed. Biological vision most probably does not represent the only class of operative visual systems, yet they are our only proof of existence and, as pointed out in [2]: *Active Vision is the natural result of considering vision in the context of an active agent and its changing environment.*

References

[1] Y. Aloimonos, I. Weiss, and A. Bandyopadhyay. Active vision. *IJCV*, 1(1):333–356, 1988.

[2] Attendees of the NSF Active Vision Workshop. Promising directions in active vision. *IJCV*, 11(2):109–126, 1993.

[3] R. K. Bajcsy. Active perception. *Proceeding of the IEEE.*, 76(8), 1988.

[4] J. L. Barron, D. J. Fleet, and S. S. Beauchemin. Performance of optical flow techniques. *IJCV*, 12(1):43–77, 1994.

[5] A. Basu and S. Licardie. Alternative models for fish-eye lenses. *Pattern Recognition Letters*, 16(4):433–441, 1995.

[6] S. S. Beauchemin and J. L. Barron. The local frequency structure of 1d occluding image signals. submitted, 1997.

[7] S. S. Beauchemin and J. L. Barron. A theoretical framework for discontinuous optical flow. submitted, 1997.

[8] H. H. Bulthoff. *Shape from X: Psychophysics and Computation, in Computational Models of Visual Processing, M. Landy and J. A. Movshon (Eds.).* MIT Press, Cambridge, MA, 1990.

[9] H. H. Bulthoff and H. A. Mallot. Interaction of different modules in depth perception. In *Proceedings of the 1st International Conference on Computer Vision*, pages 295–305, London, England, 1987.

[10] J. J. Clark and A. L. Yuille. *Data Fusion for Sensory Information Processing Systems, The Kluwer International Series in Engineering and Computer Science.* Kluwer Academic Publishers, 1990.

[11] D. Coombs, M. Herman, T.-H. Hong, and M. Nashman. Real-time obstacle avoidance using central flow divergence. In *Proceedings of the 5th International Conference on Computer Vision*, Cambridge, MA, June 1995.

[12] D. Coombs, M. Herman, T.-H. Hong, and M. Nashman. Real-time obstacle avoidance using central flow divergence, and peripheral flow. *IEEE Trans. on Robotics and Autom.*, 14(1):49–59, 1998.

[13] C. Fermuller and Y. Aloimonos. The role of fixation in visual motion analysis. *IJCV*, 11(2):165–186, 1993.

[14] A. Maki. *Stereo Vision in Attentive Scene Analysis.* PhD thesis, Stockholms Universitet, March 1996.

[15] H. A. Mallot, H. H. Bülthoff, J. J. Little, and S. Bohrer. Inverse perspective mapping simplifies optical flow computation and obstacle detection. *Biological Cybernetics*, 44(3):177–185, 1991.

[16] J. Moran and R. Desimone. Selective attention gates visual processing in extrastriate cortex. *Science*, 229:782–784, 1985.

[17] E. Niebur, C. Kock, and C. Rosin. An oscillation-based model for the neuronal basis of attention. *Vision Res.*, 33(18):2789–2802, 1993.

[18] P. Nordlund. *Figure-Ground Segmentation Using Multiple Cues.* PhD thesis, Stockholms Universitet, May 1998.

[19] B. Olshausen, C. Anderson, and D. Van Essen. A neurological model of visual attention and invariant pattern recognition based on dynamic routing information. *J. Neurosci.*, 13(11):4700–4719, 1993.

[20] K. Prazdny. Motion and structure from optical flow. In *Proceedings of IJCAI*, pages 702–704, Tokyo, Japan, August 1979.

[21] D. Raviv and M. Herman. A new approach to vision and control for road following. In *Proceedings of the IEEE Workshop on Visual Motion*, Princeton, NJ, October 1991.

[22] S. Shah and J. K. Aggarwal. Autonomous mobile robot navigation using fish-eye lenses. *Image Analysis Applications and Computer Graphics*, 1024:9–16, 1995.

[23] S. Shah and J. K. Aggarwal. Intrinsic parameter calibration procedure for a (high distortion) fish-eye lens camera with distortion model and accuracy estimation. *Pattern Recognition*, 29(11):1775–1778, 1996.

[24] J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. Lai, N. Davis, and F. Nuflo. Modeling visual attention via selective tuning. *AI*, 78(1):507–545, 1995.

[25] T. Viéville, F. Romann, B. Hotz, H. Mathieu, M. Buffa, L. Robert, P. Facao, O. D. Faugeras, and J. T. Audren. *Autonomous Navigation of a Mobile Robot Using Inertial and Visual Cues, edited by M. Kikode, T. Sato and K. Tatsuno.* Yokohama, 1993.

[26] A. M. Waxman, B. Kamgar-Parsi, and M. Subbarao. Closed-form solutions to image flow equations for 3d structure and motion. *IJCV*, 1:239–258, 1987.

[27] S. A. Zeki. *Vision of the Brain.* Blackwell, Oxford, 1993.

Acknowledgment: The first author wishes to thank NSERC Canada for supporting this research.