# 1 Visual Navigation: Flies, Bees, and UGV's*

Yiannis Aloimonos
University of Maryland

Recent developments in empirical and computational disciplines that study the problem of visual perception have advanced a great deal our understanding of the mechanisms underlying the process of visual navigation. Visual navigation amounts to the control of sensory mediated movement and encompasses a wide range of capabilities, ranging from low-level ones related to kinetic stabilization to high-level ones related to the ability of a system to acquire a memory of a place or location and recognize it (homing). This book consists of articles in several disciplines representing major developments in some area of navigation, and it is divided into three parts. The first part (chapters 2 and 3) describes advances in understanding how biological systems deal with navigation problems, and in particular how insects process images for the purpose of successfully moving around in their world. The second part (chapters 4, 5 and 6) is devoted to recent theoretical developments that will influence the design of autonomous systems and future basic research in the field. The third part (chapters 7 through 12) describes the application of the theoretical developments to the design of autonomous visual navigation systems operating in various environments. This introduction briefly examines the problem of visual navigation from several perspectives, provides a description of basic research questions, and shows how the different chapters relate to different questions and to each other. The treatment is, for the most part, of a computational nature, stressing geometry, statistics, computational techniques, and signal processing.

A system that successfully navigates using its perceptual sensors must have a number of capabilities that we can roughly separate into *local* and *global* ones. Local capabilities enable the system to interact effectively with its immediate surroundings, like understanding its own motion, recognizing obstacles and other moving objects, and obtaining a stable view of the world. Global capabilities enable the

* Unmanned Ground Vehicles. This research on navigation was funded partly by the Office of Naval Research under Grant N00014-96-1-0587, the National Science Foundation under Grant IRI-9057934, and ARPA Order #A422, through grant F49620-93-1-0576 (Integrated Active Vision for UGV RSTA), a project in which the author led a consortium including the Universities of Maryland, Rochester and Pennsylvania, as well as the National Institute of Standards and Technology, in the development of visual competences for an unmanned ground vehicle performing reconnaissance, surveillance, and target acquisition (RSTA on the move).

system to deal with larger spaces, like building a memory of an area and finding particular places in that area (homing). For the computational scientist, the fundamental problem in designing these capabilities lies in understanding and discovering the appropriate spatiotemporal representations that the system should recover from visual information. Having these representations of the outside world and algorithms for obtaining them in real time, efforts can be made to put together a broad range of navigational capabilities or competences [5]. At the same time empirical scientists can form hypotheses on the neural implementation of particular capabilities and design specific experiments for testing them. The combined effort of computational and empirical research can lead to an examination of the nature of the required spatiotemporal representations at various levels of abstraction as well as their extraction with real-time techniques.

A navigating system is in a continuous motion and at the lowest level it should extract a representation of the moving images that it receives. Traditionally this representation has been considered the idealized one of the motion field. As a system moves in some environment, every point in that environment moves with some velocity. The projection of these velocity vectors on the imaging surface constitutes the so-called motion field. This field depends on the 3D motion and the structure of the scene in view. The last fifteen years have seen a very large amount of literature on the interpretation of motion fields as regards the information they contain about the three-dimensional world. The problem is, in general, very difficult and its difficulty is compounded by the fact that the information that can be derived from the sequence of images is not the exact projection of the 3D motion field, but rather only information about the movement of light patterns. The exact movement of every point in the image is termed the optical flow field. In general, accurate values of the optical flow field are not computable; the so-called normal flow, the component perpendicular to the edges, is the only component of the optical flow that is well defined on the basis of local information. This is the well-known aperture problem. In many cases, it is possible to obtain additional flow information for areas (patches) in the image. Thus, the input that a system can use for further motion processing is some partial optical flow information. The question then is, what is the nature of subsequent motion processing?

The conventional wisdom suggests that subsequent motion processing should be directed at estimating the optical flow. It appears that this view is so deeply rooted that even the current views of most neurobiologists regarding the motion pathway in primates suggest that in area V5 or MT the aperture problem is "solved," and from this point specialized "motion processing" areas such as MST and FSP are fed with the appropriate input (see [10] for a review). Recent theoretical and empirical findings suggest that this may not be the case. Although the aperture problem may be resolved in MT, initial local motion measurements far away from an optical flow representation (and thus termed qualitative) may be sufficient to address a number of navigation tasks. The processing may be of such a nature that when it reaches the area MT the system has already at its disposal a vast amount of information about the three-dimensional world [13].

The analysis of Chapter 4, which is basically a statistical and geometrical investigation of motion fields, proves that the problem of interpreting motion fields in order to extract information about the three-dimensional world is ill-posed by investigating the stability of 3D-motion estimation processes. The chances that biological systems or successful real-time visuomotor control systems solve ill-posed problems are very slim. This means that they may possibly use a different input. Chapters 10, 11, and 12 describe the design and construction of real systems working in some environments. Such truly impressive systems demonstrate the power of computational approaches to perception, and reinforce the maturity of the discipline of computer vision, as industrial systems working in specific environments under various assumptions slowly emerge. At the same time, they demonstrate the limitations of current approaches to visual motion analysis, by showing scenarios and situations where the developed systems fail.

This calls for the development of better spatiotemporal representations at several levels of abstraction. Chapter 6 describes an approach to motion analysis based on partial optical flow information. This approach is based on a minimum of knowledge about image motion, namely, the sign of the projection of the optical flow along directions where it can be robustly computed. These measurements along a set of appropriately chosen orientations possess a rich global structure; they give rise to simple patterns in the image, and the location and form of these patterns encode three-dimensional information. This chapter could be considered an implementation of Gibson's transformational invariants [7], and demonstrates that a lot of information can be robustly extracted from a sequence of images before correspondence between the images has been established [3].

Regarding representations of visual space, as Chapters 10, 11, and 12 demonstrate, the overwhelming view has been in favor of Euclidean representations encoding distances between different features in the visible world. As research in the psychophysics of perceptual space advances, researchers in computational vision are examining a variety of space representations less complex than their Euclidean counterparts. Chapter 5 examines the usefulness of projective space representations and other qualitative representations residing between the projective and Euclidean layer, and describes how to use them to achieve a number of tasks. It is expected that the development and extraction of such qualitative representations will continue to be an exciting research area in the future. Visual memories will be studied that can be considered as structures consisting of such spatiotemporal representations [1, 4, 6].

Currently, in the absence of alternative representations, researchers consider memories consisting simply of images and test their limits using digital devices capable of storing large amounts of information and the pattern recognition and statistics technology of the past twenty years. Chapters 7 and 8 describe such an approach, and Chapter 9 examines higher level problems of planning through navigating in large spaces.

Finally, Chapters 2 and 3 describe empirical findings in insects and reinforce the view that activities simplify visual processing. The empirical findings in insects and primates confirm various issues that have been illuminated here from a computational point of view:

1. Visual motion measurements at several stages are of a qualitative nature, i.e., they are not measurements of optical flow but rather measurements of the sign of local motion along various directions [8].
2. Perception resembles a process of pattern matching a great deal. This view, which was popular a number of years ago with the emergence of the discipline of pattern recognition, was abandoned by computational theorists because it was considered extremely difficult to identify patterns encoding properties of the three-dimensional world [12]. For many years the pattern recognition approach, advocated at a philosophical level by Gibson, was severely criticized; however, advances in computational vision now make it possible to design appropriate patterns using geometry and exploiting the physics of light.
3. There is a difference between actual physical space and perceptual space, the representation of physical space in our heads. Euclidean geometry encoding distances between features in space is a good tool to model physical space. However, recent advances indicate that it is too difficult and most probably impossible to exactly match the perceptual and physical spaces, i.e., to extract from images in real-time robust Euclidean reconstructions of physical space. Psychophysics of man has provided rich evidence that humans do not perceive exact metric descriptions of space [9, 11]. What then are these descriptions of space that are robustly extracted in real time? Projective or affine descriptions are candidates having some limitations—projective structures do not carry much information and affine structures only make sense for scenes viewed from far away. The discovery of the large spectrum of such "qualitative" representations represents an exciting research goal for the immediate future [2].

Visual navigation will continue to be a rich source for basic research questions in vision. The potential for applications in autonomous systems operating in real or virtual environments will make the field flourish in the years to come. With regard to virtual environments, research into visual motion and navigational competences will revolutionize the media technology of television by providing the viewer with control of the camera angle. Virtual reality will enable the viewer to see a synthesized scene from any perspective. This can happen realistically only if an adequate model of the scene is automatically extracted. With the hope that readers will find this book useful, I wish to express my thanks to all the contributors, and to Sara Larson for her editorial and graphics skills.

## References

1. Y. Aloimonos, C. Fermüller, and A. Rosenfeld, "Seeing and understanding: Representing the visual world," *ACM Computing Surveys* **27**(3), 1995, 307–309.

2. O. Faugeras, "Stratification of 3-D vision: Projective, affine, and metric representations," *Journal of the Optical Society of America A* **12**(3), March 1995, 465–484.

3. C. Fermüller and Y. Aloimonos, "Direct perception of three-dimensional motion from patterns of visual motion," *Science* **270**, 1995, 1973–1976.

4. C. Fermüller and Y. Aloimonos, "Representations for active vision," *Proc. Int'l. Joint Conference on Artificial Intelligence*, Montreal, Canada, August 1995.

5. C. Fermüller and Y. Aloimonos, "Vision and action," *Image and Vision Computing* **13**(10), 1995, 725–744.

6. C. Fermüller and Y. Aloimonos, "Ordinal representations of visual space," *Proc. ARPA Image Understanding Workshop*, February 1996.

7. J. Gibson, *The Ecological Approach to Visual Perception*, Houghton Mifflin, Boston, 1979.

8. G.A. Horridge, "What can engineers learn from insect vision," *Proceedings of the Royal Society London B*, August 1992.

9. J.J. Koenderink and A.J. van Doorn, "Relief: Pictorial and otherwise," *Image and Vision Computing* **13**, 1995, 321–334.

10. M. Sereno, *Neural Computation of Pattern Motion*, MIT Press, Cambridge, MA, 1993.

11. J.S. Tittle, J.T. Todd, V.J. Perotti, and J.F. Norman, "Systematic distortion of perceived three-dimensional structure from motion and binocular stereopsis," *Journal of Experimental Psychology: Human Perception and Performance* **21**(3), 1995, 663–678.

12. S. Ullman, "Against direct perception," *Behavioural and Brain Sciences* **3**, 1980.

13. S. Zeki, *A Vision of the Brain*, Blackwell Scientific Publications, Cambridge, MA, 1993.