Analyzing Action Representations

Y. Aloimonos and C. Fermüller

Computer Vision Laboratory, Center for Automation Research, Institute for Advanced Computer Studies, and the Department of Computer Science, University of Maryland, College Park, MD 20742-3275, USA
yiannis@cfar.umd.edu,
WWW home page: http://www.cfar.umd.edu/~yiannis/research.html

Abstract. We argue that actions represent the basic seed of intelligence underlying perception of the environment, and the representations encoding actions should be the starting point upon which further studies of cognition are built. In this paper we make a first effort in characterizing these action representations. In particular, from the study of simple actions related to 3D rigid motion interpretation, we deduce a number of principles for the possible computations responsible for the interpretation of space-time geometry. Using these principles, we then discuss possible avenues on how to proceed in analyzing the representations of more complex human actions.

1 Introduction and Motivation

During the late eighties, with the emergence of active vision, it was realized that vision should not be studied in a vacuum but in conjunction with action. Adopting a purposive viewpoint makes visual computations easier by placing them in the context of larger processes that accomplish tasks. This new framework, known by a variety of names such as active, purposive, animate or behavioral vision, has contributed a wealth of new technical results and has fueled a variety of new application areas. Several groups constructed active vision systems and studied the basics of many visual competences, but for the most part, this study of the perception/action coupling concentrated on problems related to navigation. A flurry of activity produced a veritable cornucopia of algorithmic approaches to 3D motion estimation, 3D shape recovery, tracking, and motion and scene segmentation. As a matter of fact, one can say that this is the most mature area in the field of computer vision. This was, perhaps, not surprising since all these problems are related to the most basic action of all, one that transcends all systems with vision, that of self-motion; although the perception/action cycle surrounding the basic action of self-motion is not yet fully understood (we will discuss later some of the intricacies involved), it is clear that the problem is one of geometry and statistics awaiting the right modeling for its complete resolution. But an intelligent system with perception is not just a system that understands its movement, and the shape and movement of other objects in its field of view. It has a large number of capabilities related to recognition, reasoning, planning, imagination and learning. Despite the fact that the active vision framework was embraced as the right tool in dealing with the recovery of 3D information from image sequences, that was not the case when it came to higher-level problems such as recognition, planning and reasoning. Lacking a foundation, not much progress has been made in those higher-level problems. For example, with the exception of simple objects in controlled environments, not much progress has been made in the field of object recognition. All new proposals during the past decade have followed some advance in the structure from motion problem.[1] There are, of course, many reasons for this, and some of them will become clear in the course of our exposition.

The goal of this paper is twofold, one methodological and the other technical. Our intent is to put forward the thesis that action represents the basic seed of intelligence, underlying perception of our environment, which is our major interest, recognition and other high-level processes such as communication. If this is the case, action representations are very basic components of an intelligent system. What are they? What could their nature and kind be? How can we build them? What does it mean to understand them? Most important, what techniques should we employ in order to obtain them? Answering these questions is our second goal and we approach it using only computational arguments. Before we begin, we need to discuss other efforts attempting to answer such kinds of questions.

2 Intelligence, Learning and the Stratification of Reality

Slowly but steadily, the paradigm that has dominated the study of cognition and the fields surrounding it for the past century is starting to be challenged. The dominant approach has been to view mental states as having meaning in terms of symbol systems like natural language. The emerging new approach is referred to as embodied intelligence or the sensorimotor theory of cognition which proposes that higher cognition makes use of the same structures as those involved in sensorimotor activity. This view is highly exemplified in recent efforts in AI to achieve intelligence through building robots and having them "develop" through interaction with their environment. Although there exist interesting points in the contemporary sensorimotor theories of cognition, and the projects of AI engineers have merits of their own, there exist two fundamental problems with them that we explain here.

The belief that smaller systems develop through interaction with their environment and become integrated to form new and more advanced systems owes its appeal to its intended mimicking of evolution. It makes, however, an implicit assumption, namely that evolution is evolution in degree and not evolution in kind. By evolution in kind we mean the introduction of new components having

[1] When it became clear that 3D information could be extracted from point and line correspondences, researchers concentrated on pose recovery in the polyhedral world. When view interpolation was achieved for some cases, recognition approaches considering objects as collections of 2D views appeared.

qualitatively different characteristics from the existing components. This leads to a strange kind of one-way relationship, both between the whole and its constituent parts within the system, and between higher systems and their primitive ancestors. One can define this relationship by saying that the whole is its parts, and continues to be so even if, as the result of the introduction of new subsystems (kind), it acquires a number of additional characteristics in the course of its evolution. The subsidiary systems themselves do not gain any higher characteristics and may even lose some in the process of simplification. The one-way relationship consists in the system as a whole possessing all the characteristics of its component parts, but none of these parts possesses the characteristics specific to the whole. Similarly, every system possesses most of the characteristics of its primitive ancestors, but even complete knowledge of a system's characteristics will not make it possible to predict those of its more highly developed descendants [17]. Thus, we can explain a system only if we accept the present structures in its body as our data.[2]

The second problem, related to the first, is best explained by utilizing the ideas of the philosopher Nicolai Hartmann [15]. Like all processes in life, those of acquiring and storing relevant information take place on many different levels and are interlinked at many points. The world in which we live, according to Hartmann, is built of different strata, each with its own existential categories that distinguish it from other strata. "There are in the hierarchy of existence certain phenomenal realities whose fundamental differences our minds fail to bridge .... Any true and accurate theory of categories must have as much regard for these gaps as for the existential relationships that bridge them." These relationships, however, only transcend in a unilateral manner the divisions between the four great strata of existence—the inorganic, organic, cognitive and conscious. The principles of existence and laws of nature that govern inorganic matter apply with equal validity to higher strata. But Hartmann insists that the differences between the higher and lower strata are far from restricted to the distinctions between inorganic, organic, cognitive and conscious. "The higher elements that make up the world are stratified in a similar way to the world itself," he wrote. This means that each step leading from a system of a lower order to one of a higher order is the same in nature and complexity as the coming into existence of life itself. This viewpoint about the stratification of existence can guard us against common mistakes in studying cognition. If we neglect the categories and laws that are exclusive properties of the higher strata or even deny their existence, we commit the fallacy of an upward transgression of the limits imposed upon us by the stratification of the real world. It is not possible to explain the laws and processes proper to higher levels in terms of categories derived from lower levels. Another common mistake results from transgressing the boundaries between levels of existence in the opposite direction. This is described by Hartmann

[2] Irrational Residue: The number of historical causes one would need to know to fully explain why an organism is as it is may not be infinite but it is sufficiently great to make it impossible to trace all causal chains to their end. As Polanyi pointed out [19], a higher animal cannot be reduced to its simpler ancestors.

as follows: The base of the entire world picture is then chosen on the level of conscious experience—the level on which man experiences his own subjective life—and from there the principle is extended downwards to the lower level of reality.

What does this all mean for us who wish to gain an understanding of perceptual mechanisms and, in particular, how action representations make up an intelligent system? There are several lessons we can obtain. It is a mistake, while seeking unitary explanatory principles behind the world, to try to explain lower and more primitive systems on the basis of principles applicable only to higher systems and vice versa. Similarly, the attempts of some psychologists and behavioral scientists to achieve intelligence through learning are at fault. This is true for both primitive systems that are incapable of it and systems of more advanced organisms which are not only incapable of being modified by learning but whose phylogenetic program makes them resistant to all modification.

The only chance at understanding intelligent systems seems to be to study them by accepting their present structures as our data, and work towards understanding the different components. But here is the tricky part. The components we study should be such that they do not require other components in order to be understood, or if they do, all interrelated components should be studied together. Otherwise it appears that there is no hope for understanding. And finally, our only tools should be the ones of the physical sciences, geometry, statistics and computation.

3 Understanding Actions, Objects

Our task is to increase our understanding of how to piece together the components of an intelligent system with vision. Our thesis is that action is fundamental in this regard. It is of course irresistible to attempt to arrive at a unitary world image and try to explain the diversity of the world (and intelligent systems) on the basis of ontological and phenomenal principles of one single kind. According to our previous analysis this would be a mistake after all. But our goals are much more modest. We want to uncover principles underlying the workings of the perceptual system and in particular shed light on the nature of action representations.

Our basic thesis is that the content of the mind is organized in the form of a model of the world. That model contains objects and their relationships, events and the interaction of the system with its environment. In building, maintaining, manipulating and working with that model we should distinguish three items: the world itself, the system (its body) and the system's interaction with the world. What could the best candidate be for building a foundation on the basis of which this complex model can be understood?

Recall that it is trivially true that understanding a thing is relating it to something already understood; if we continue along this chain, this means that there must be something understood in terms of itself. This entity must be something with which a system is intimately familiar in a non-symbolic way. We conclude that the best candidate for this is the agent's own activity.

There is a number of basic actions that are the basis of acquiring models of the shape and motion of the environment. These actions are related to the system's understanding of its own motion and they are universal, as all systems must have them in some form. The next sections are devoted to the computational mechanisms underlying the coupling of such actions with perception. By taking a look at these actions, on the one hand we can give an example and elaborate on what it means to understand these actions. On the other hand, since the computations involved in these basic actions form the foundation for many other computational processes we gain insight into the nature of space-time representations. That is, it allows us to deduce a number of principles on what the shape and motion models of the environment might be and how they could be computed. The computational principle of feedback loops is emerging as a foundational theme.

Next we are interested in the very large repertoire of more complex actions through which an agent interacts with his environment. Since understanding these actions should serve as a foundation for all other understanding, we should not presuppose understanding of other things. We merely assume that the agent understands its own body. Equipped with this knowledge and the foundation of space-time perception delivered from basic actions, we can provide a characterization of action as: *One understands an action if one is able to imagine performing the action with images that are sufficient for serving as a guide in actual performance.* This characterization provides two things. First, it shows that it is the ability to imagine an action which constitutes understanding it. Imagining it amounts to representing the constitutive components of the action as such, in a form that is capable of guiding the performance of the action. Because the components of the action form the content of the agent's mental state, the agent can be said to understand it. Second, this sort of understanding produces the feeling of confidence that one understands an action, because the understanding consists in a mental state of which the agent can be aware. Moving along this line of thought, understanding anything amounts to knowing the possible actions one might perform in relation to that thing and being aware of that understanding is consciously imagining those actions. Thus, we can say that one has an understanding of a perceived object if one is able to imagine incorporating the object into the performance of an action that is already understood, with an image sequence sufficient for serving as a guide in actual performance.

Thus, the ability to intentionally perform an action, manifested in the ability to imagine performing it, is a way of understanding the action that presupposes no other understanding. Similarly, an object that is intentionally incorporated into the performance of an (understood) action is itself understood in relation to that action. In moving from the understanding of actions to the understanding of objects we have crossed an important boundary. Understanding actions does not require a concept of the objective external world; that is how it can be a foundation for such a concept. But understanding objects does require such a concept.

It is not our intention to delve further into a philosophical inquiry about understanding objectivity and the mind/world distinction. Our purpose here is to achieve an understanding of action representations. Having a crisp characterization of what it means to understand an action we will be ready in Sect. 4 to investigate constraints on such representations. We will not delve into object recognition issues but the framework is in place to carry this investigation further into higher-level problems.

But remaining truthful to the principles of embodied intelligence, we cannot study in a technical sense actions unless we commit ourselves to a particular body. Of course the same is true for the basic actions and the perception of shape and motion, but in that case the dependence lies mostly in image acquisition (eye design) and computational capacity; in our days of extreme computational power these issues are not that important. But when it comes to higher-level actions and interaction with the world, it is important to choose a body. A failure to do so will turn our study completely philosophical. As such a body, we choose the human body.

4 Basic Actions

The most basic actions are the ones which can be based on the most simple model of 3D motion, and this is the model of instantaneous rigid motion. As living organisms move through the environment their eyes undergo a rigid motion. The visual system has to recognize and interpret this rigid motion from the visual input, that is, the sequence of images obtained by its eyes. This immediately gives rise to two basic actions: first, the capability of estimating one's self-motion or egomotion; second, the detection of objects in the static scene which move themselves, that is, the detection of independently moving objects. Similarly, the motion of various objects is rigid or can be closely approximated by one. Thus a third action arising from this model is the estimation of rigid object motion. This action, however, is much more difficult than the estimation of self-motion. One reason is that objects usually cover a small field of view and thus less data is available for the estimation. Another reason is that objects lie outside the system's body and thus there is less information from other sensors. In contrast, for self-motion estimation biological systems—in addition to vision—also use inertial sensors (such as those in our ears). As a result, most systems only possess the capability to partially estimate rigid object motion. Even rather simple organisms understand the approaching of the enemy, or can estimate how long it will take to intercept an object (the estimation of time to contact). A somewhat more sophisticated but still partial estimation would be a rough estimate of the object's translation or rotation.

The visual input is a sequence of images, a video. Aside from how the images are formed (that is, the properties of the eye or camera), the images depend on the spatiotemporal geometry, that is, the layout of the scene and the relative 3D motion between observer and scene, and the optics, that is, the light sources and the reflectance properties of the surfaces. By considering only changes of image patterns, an image representation of the movement of the 3D scene points is obtained. This representation, a vector field, ideally is only due to the geometry that is the relative 3D motion between the eye and the scene and the shape of the scene. If the relative 3D motion between the observer and parts of the scene is rigid, then segmentation amounts to the localization of those parts of the flow field corresponding to the same rigid motion and motion estimation amounts to decoding the 3D motion parameters. The shape of the scene determines the remaining components of the vector field and it can be decoded after motion estimation.

The interpretation of motion information, probably because of its well-defined geometric nature, has received a lot of attention in the computer vision literature. A methodology emerged under the hood of which most studies were conducted, which defined rigid motion interpretation as the "structure from motion" problem and it was considered to be carried out with three consecutive computational steps. First the exact movement of image points is estimated, either in the form of the optical flow field if images densely sampled in time are considered, or in the form of correspondences of single image points if images further apart in time are used. The optical flow field or correspondences are approximations of the projections of the 3D motion field which represents the movement of scene points. The second computational step consists in estimating the 3D rigid motion from the optical flow field, and the third step amounts to estimating the 3D structure of the scene usually from multiple flow fields using the estimates of the 3D motion.

With this framework it has been considered that exact reconstruction of space-time geometry is possible—it is just a matter of using the right sophisticated tools—and it can be addressed in a purely bottom-up manner computationally. We'll argue that this cannot be. Throughout this and the next section we'll provide computational arguments supporting our view and elaborate on what the computations should be.

First, optical flow or correspondence cannot be estimated accurately using only the image data, that is, before performing estimations about the 3D geometry. The local image data defines only one component of the image velocity. This is the so-called aperture problem illustrated in Fig. 1: Local image data usually only provides one-dimensional spatial structures. From the movement of these linear features between image frames the movement of single image points cannot be determined, but only the flow component perpendicular to the linear feature; this is the so-called normal flow. In order to estimate the two-dimensional image flow, a model of the flow field is needed; then normal measurements in different directions within local neighborhoods can be combined. Here lies the problem: In order to model the flow field, additional information is necessary. It is easy to model the parts of the image which correspond to smooth scene patches by using smoothness assumptions of different kinds, but there are discontinuities in the flow field, and their locations must be known prior to modeling the flow

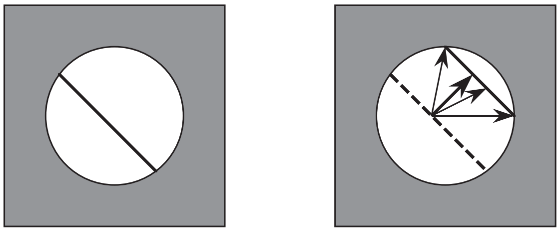

Fig. 1. Aperture problem: (a) Line feature observed through a small aperture at time t. (b) At time $t + \delta t$ the feature has moved to a new position. It is not possible to determine exactly where each point has moved to. From local measurements only the flow component perpendicular to the line feature can be computed

field. The discontinuities in the flow field are due either to independently moving objects in the scene or to scene surfaces at different depths. On the other hand, the discontinuities cannot be detected unless an estimate of the flow field is available. Thus the problems of estimating the flow field and locating the discontinuities are inherently coupled and the situation seems to present itself as a chicken and egg problem. Discontinuities provides an even more severe problem than usually considered. Most of the literature considers the idealized problem of estimating the motion for one rigidly moving observer in a static environment. Thus all the discontinuities they deal with are due to depth. Even if the locations of the discontinuities in the flow field are known, a system would need additional information about the 3D motion and scene to distinguish the motion discontinuities from the pure scene discontinuities. There also has been a lot of work in the literature on motion segmentation, but these studies consider the problem separate from 3D motion estimation: either they attempt a localization of the discontinuities on the basis of flow field information only, or they assume the 3D motion to be known. It is clear that for a real system the two problems of 3D motion estimation and independent motion detection have to be considered together.

Besides the problem due to discontinuities there are also statistical difficulties with the estimation of optical flow. As will be elaborated on later, the estimation of optical flow from noisy normal flow measurements even within neighborhoods of smooth flow is very difficult. To avoid bias and obtain very accurate flow, theoretically one would need detailed models of the flow field, and these in turn can only be obtained from 3D information; specifically, this means the discontinuities, the shape of the scene and the 3D motion.

Since accurate optical flow estimation bottom up seems to be infeasible, the question is, is it possible to perform 3D motion estimation using as input normal flow measurements? In the past we have studied this question, and we have developed two constraints that allow to extract 3D geometry directly flow normal flow. One constraint that relates image motion to 3D motion only, and a second one which also involves the scene.

5 Relating Image Measurements to 3D Geometry

Consider a system moving with rigid motion in a static environment. The motion is described by an instantaneous translational velocity $\mathbf{t}$ and a rotational velocity $\boldsymbol{\omega}$. Each scene point $\mathbf{R} = (X, Y, Z)$ measured with respect to a coordinate system $OXYZ$ fixed to the nodal point of the system's eye where Z is the optical axis has velocity $\dot{\mathbf{R}} = -\mathbf{t} - \boldsymbol{\omega} \times \mathbf{R}$.

We consider the image to be formed on a plane orthogonal to the Z axis at distance f (focal length) from the nodal point. Thus through perspective projection image points $\mathbf{r} = [x, y, f]$ are related to scene points $\mathbf{R}$ by $\mathbf{r} = \frac{\mathbf{R}f}{\mathbf{R}\cdot\hat{\mathbf{z}}}$, where $\hat{\mathbf{z}}$ is a unit vector in the direction of the Z axis. The projection of the 3D motion field on the image gives the image motion field

$$\dot{\mathbf{r}} = \frac{1}{(\mathbf{R}\cdot\hat{\mathbf{z}})}\left(\hat{\mathbf{z}} \times (\mathbf{t} \times \mathbf{r})\right) + \frac{1}{f}\hat{\mathbf{z}} \times (\mathbf{r} \times (\boldsymbol{\omega} \times \mathbf{r})) = \frac{1}{Z}\mathbf{u}_{\mathrm{tr}}(t) + \mathbf{u}_{\mathrm{rot}}(\omega) \qquad (1)$$

where Z is used to denote the scene depth $(\mathbf{R}\cdot\hat{\mathbf{z}})$ and $\mathbf{u}_{\mathrm{tr}}, \mathbf{u}_{\mathrm{rot}}$ are the direction of the component of the flow due to translation and the component of the flow due to rotation, respectively. Due to the coupling of Z and $\mathbf{u}_{\mathrm{tr}}$ in this relationship only the direction of translation (the Focus of Expansion, FOE, or focus of contraction, FOC, depending on whether the observer is approaching or moving away from the scene), scaled depth and the three rotational parameters can be obtained. The direction of the axis of rotation will be denoted as AOR. The normal flow $\mathbf{v_n}$ amounts to the projection of the flow $\dot{\mathbf{r}}$ on the local gradient direction. If $\mathbf{n}$ is a unit vector denoting the orientation of the gradient, the value of the normal flow, v_n amounts to

$$v_n = \dot{\mathbf{r}} \cdot \mathbf{n}$$

The first constraint, the "pattern constraint," uses as its only physical constraint the fact that the scene lies in front of the camera and thus all the depth values have to be positive. This allows the relation of the sign of normal flow measurements to the directions of the translation and the rotation, that is, the FOE and AOR. The basis lies in selecting groups of measurements of normal flow along pre-defined directions (orientation fields) which form patterns in the image plane, and these patterns encode the 3D motion [4], [5], [6], [7].

There are two classes of orientation fields. The first class are called copoint fields. Each copoint field parameterized by $\mathbf{s}$, is defined as the unit vectors perpendicular to a translation flow field with translation $\mathbf{s}$. These unit vectors are in the direction $\mathbf{v}_{\mathrm{cp}}(\mathbf{s}, \mathbf{r}) = \hat{\mathbf{z}} \times \mathbf{u}_{\mathrm{tr}}(\mathbf{s}) = \hat{\mathbf{z}} \times (\hat{\mathbf{z}} \times (\mathbf{s} \times \mathbf{r}))$ and they are perpendicular to the bundle of lines passing through $\mathbf{s_0}$ (the point where $\mathbf{s}$ intersects the image plane) (Fig. 2a).

Now consider a normal flow field due to rigid motion $(\mathbf{t}, \boldsymbol{\omega})$ and consider all the normal flow vectors of this field which are in direction $\mathbf{v}_{\mathrm{cp}}(\mathbf{s}, \mathbf{r})$, that is,

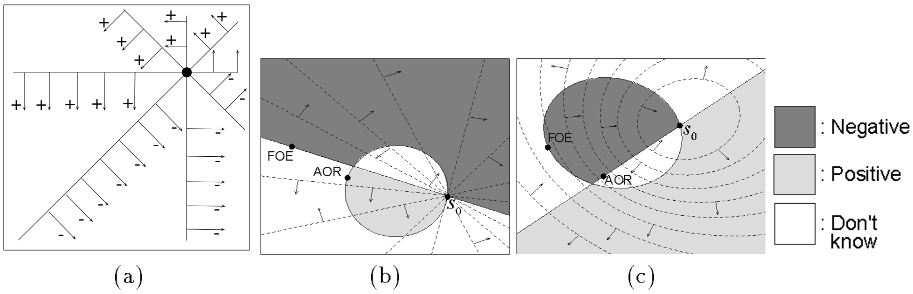

Fig. 2. (a) Copoint field filled with qualitative motion measurements. (b) copoint vectors and copoint pattern. (c) coaxis vectors and patterns

the vectors of value $\frac{\mathbf{v}_{\mathrm{cp}}}{\|\mathbf{v}_{\mathrm{cp}}\|} \cdot \dot{\mathbf{r}}$. The translational components of these vectors are separated by the conic $(\boldsymbol{\omega} \times \mathbf{r}) \cdot (\mathbf{s} \times \mathbf{r}) = \mathbf{0}$ into an area with negative values and an area with positive values. Similarly, the rotational components of the same vectors are separated by the line $(\mathbf{t} \times \mathbf{s}) \cdot \mathbf{r} = \mathbf{0}$ into areas of positive and negative values. The rigid motion field is the sum of the translational and the rotational flow field; thus the areas above need to be superimposed. Where both the translational and rotational components are positive the combined flow is positive and where both the translational and rotational components are negative the combined flow is negative. In the remaining areas the sign is undetermined. The areas of different sign form a pattern in the image plane as shown in Fig. 2b. For every direction $\mathbf{v}_{\mathrm{cp}}$ parameterized by $\mathbf{s}$ there is a different pattern and the intersection of the different patterns provides the 3D motion.

The second class of patterns is defined by the orientation field of coaxis vectors, which are the unit vectors perpendicular to a certain rotational flow field, that is, the coaxis vectors parameterized by $\mathbf{s}$ are in direction $\mathbf{v}_{\mathrm{cA}} = \hat{\mathbf{z}} \times (\hat{\mathbf{z}} \times (\mathbf{r} \times (\boldsymbol{\omega} \times \mathbf{r})))$. In this case the translational components are separated by a line and the rotational components are separated by a cone into positive and negative values and give patterns such as that shown in Fig. 2c. Using these pattern constraints the problem of self-motion estimation amounts to recognizing particular patterns in the image plane. The system would compute the normal flow and then for a number of orientation fields find the patterns which are consistent with the sign of the estimated normal flow vectors. The intersection of the possible patterns provides the solution, usually in the form of possible areas for the FOE and AOR, as most often there is not enough data to provide an exact localization of the 3D motion.

The question now is, is it possible at all to have a technique that accurately computes the 3D rigid motion parameters? The answer is no, if the data available is from a limited field of view. This statement is based on an error analysis conducted in [8], [9]. Estimating 3D motion amounts to minimizing some function. If the pattern constraints are used, this function is based on negative depth values and if the classical epipolar constraint requiring optical flow is used, this function is based on the distances of the flow vectors from the epipolar lines. The correct 3D motion is defined as the minimum in the functions. A topographic analysis of the expected values showed that for all these functions the minimum lies within a valley, and locating the exact position of the minimum within this valley under noisy conditions generally is not possible.

If the 3D motion is known, the depth of the scene can be computed from the flow field. If $\hat{\mathbf{t}}$ and $\hat{\boldsymbol{\omega}}$ are the estimates of the direction of translation and the rotation the estimated depth $\hat{Z}$ is derived from (1) as

$$\hat{Z} = \frac{\mathbf{u}_{\mathrm{tr}}(\hat{\mathbf{t}}) \cdot \mathbf{n}}{\dot{\mathbf{r}} \cdot \mathbf{n} - \mathbf{u}_{\mathrm{rot}}(\hat{\boldsymbol{\omega}}) \cdot \mathbf{n}} \tag{2}$$

If the estimates $\hat{\mathbf{t}}$ and $\hat{\boldsymbol{\omega}}$ correspond to the correct values of the actual rigid motion $\mathbf{t}$ and $\boldsymbol{\omega}$, then $\hat{Z}$ will give a correct scene depth estimate, but what if there are errors in the estimates? Then the estimated depth will be a distorted version of the actual depth [2], [3], [8], [11].

Substituting for $\dot{\mathbf{r}}$ in (2) the actual motion parameters the estimated depth can be written as

$$\hat{Z} = Z \cdot \frac{\mathbf{u}_{\mathrm{tr}}(\hat{\mathbf{t}}) \cdot \mathbf{n}}{\mathbf{u}_{\mathrm{tr}}(\mathbf{t}) \cdot \mathbf{n} - \mathbf{Z}\mathbf{u}_{\mathrm{rot}}(\delta\boldsymbol{\omega}) \cdot \mathbf{n}},$$

with $\delta\boldsymbol{\omega}$ denoting the estimation error in rotation $\delta\boldsymbol{\omega} = \hat{\boldsymbol{\omega}} - \boldsymbol{\omega}$, or to make clear the distortion as $\hat{Z} = Z \cdot D$, where $D = \frac{\mathbf{u}_{\mathrm{tr}}(\hat{\mathbf{t}}) \cdot \mathbf{n}}{\mathbf{u}_{\mathrm{tr}}(\mathbf{t}) \cdot \mathbf{n} - \mathbf{Z}\mathbf{u}_{\mathrm{rot}}(\delta\boldsymbol{\omega}) \cdot \mathbf{n}}$ is the multiplicative distortion factor to express how wrong depth estimates result from inaccurate 3D motion values. Note that D also depends on $\mathbf{n}$. For different directions, $\mathbf{n}$, the distortion factor takes on different values, ranging from $-\infty$ to $+\infty$.

This concept of space distortion forms the basis of the second constraint relating normal flow to 3D geometry, the "depth variability constraint." The idea is, instead of formulating smoothness constraints on the 2D optical flow field as is usually done, to relate the smoothness of scene patches and the 3D motion directly to normal flow. To be more specific, if we estimate the depth from the flow field corresponding to a smooth scene patch using the correct 3D rigid motion, we will compute a smooth scene patch. If we use an erroneous 3D motion estimate, then the obtained depth function, assuming that normal flow vectors in different directions are available, will be rugged. Not only do incorrect estimates of motion parameters lead to incorrect depth estimates, but the distortion is such that the worse the motion estimate, the more likely depth estimates are obtained that locally vary more than the correct ones. Thus one has to define functions which express the variability of depth values and search for the 3D motion which minimizes these functions. Locally for the flow values $\dot{\mathbf{r}}_i$ in direction $\mathbf{n}_i$ within a region R we define depth variability measures as

$$\theta_0\left(\hat{\mathbf{t}}, \hat{\boldsymbol{\omega}}, R\right) = \sum W_i \left(\dot{\mathbf{r}}_i \cdot \mathbf{n}_i - \mathbf{u}_{\mathrm{rot}}(\mathbf{u}) \cdot \mathbf{n}_i - \frac{1}{\hat{Z}}\left(\mathbf{u}_{\mathrm{tr}}\left(\hat{\mathbf{t}}\right) \cdot \mathbf{n}\right)\right)^2$$

where $1/\hat{Z}$ is the depth estimate minimizing θ_0, which is obtained by modeling the depth within the region using a parametric model, and W_i are weights that can be chosen differently. Solving for the minimum $1/\hat{Z}$ and substituting back we obtain θ_1. The sum of all θ_1 in all regions, θ_2, is the global function which we are interested in minimizing.

To algorithmically utilize these constraints one has to proceed in iterations as follows. Split the image into patches and perform a search in the space of translational directions. For every candidate translation estimate the best rotation using the depth variability measure, then perform a depth segmentation using the image data and the candidate motion parameter, and finally evaluate the depth variability measure taking into account the segmentation. The solution is found as the 3D motion minimizing the global depth variability measure.

In summary of the above discussion a number of conclusions can be drawn regarding the possible computations of space-time geometry.

- It is not possible to reliably estimate the 3D rigid motion very accurately if the field of view is limited. This is true even for the simplified case of one camera moving in a static environment. If there is less data available—as it is when one considers motion of objects covering a small field of view—the estimation is much complicated and all we can expect are bounds on the motion parameters.
- It is also not possible for a real-time system to derive accurate depth estimates. Considering the cue of motion, this follows from the fact that 3D motion cannot be estimated accurately. As a result of this one obtains a distorted depth map, but it also has been shown experimentally for other cues that human space perception is not exactly Euclidean [16].
- Since reconstruction is not possible the information describing the models must be encoded in the form of patterns in the flow field.
- The computations carrying out the space-time geometry interpretation must be implemented as feedback processes. This is true even for the simple, basic actions as will be elaborated on in the next section, and for more complicated actions the interplay between bottom-up and top-down processes becomes extremely important.

6 Feedback

The interpretation of motion fields is a difficult computational problem. From the raw images only, it is not possible to compute a very accurate 2D motion field, and from erroneous motion field measurements it is not possible to perform good 3D motion estimation, and, in the sequel, scene estimation. To estimate good flow we need information about the space-time geometry, and to estimate the geometry well we need at least the flow field to be segmented accurately.

This seemingly chicken and egg situation calls for a simultaneous estimation of image motion and 3D motion and scene geometry, and this could only be implemented through an iterative or feedback process. The whole interpretation process should take the following form: First, the system estimates approximate image velocity by combining normal flow measurements. The representation of these estimates could be in the form of qualitative descriptions of local flow field patches or bounds on the flow values, but it should allow for finding the more easily detected discontinuities in the flow field. Using the flow computed in this way or maybe even normal flow measurements within segmented areas an estimate of 3D motion is derived and a partial 3D shape model of the scene computed. Subsequently, the computed information about the space-time geometry is fed back to use image measurements from larger regions to perform better flow estimation, more accurate discontinuity localization, and in the sequel the estimates of 3D motion and structure are improved. Naturally the whole interpretation process has to be dynamic. While the system computes structure and motion from one flow field, the geometry changes and new images are taken. Thus the system also has to be able to make predictions and relate the 3D information computed earlier in time to images acquired later.

There are a number of ways these feedback processes could be implemented in a system. From a computational perspective we can ask for an optimal way, and this will depend on the computational power of the system and the accuracy of the estimates needed. There are many efforts in the biological sciences, in fields such as psychophysics, neurophysiology and anatomy, to figure out the structure and processes which exist in human and other primate brains—this structure is generally referred to as the motion pathway; but many of these studies do not consider that the sole purpose of motion processing is to interpret the 4D space-time geometry. We'll next elaborate on two visual illusions that in our opinion demonstrate that the human visual system must perform feedback processes to interpret image motion.

6.1 The Ouchi Illusion

The striking Ouchi illusion shown in Fig. 3 consists of two black and white rectangular checkerboard patterns oriented in orthogonal directions—a background orientation surrounding an inner ring. Small retinal motions, or slight movements of the paper, evince a segmentation of the inset pattern, and motion of the inset relative to the surround. Our explanation lies in the estimation of differently biased flow vector fields in the two patterns which in turn give rise to two different 3D motion estimates. When the system feeds back these motion estimates it performs a segmentation of the image using the flow fields and the static image texture, and because of the difference in 3D motion one pattern appears to move relative to the other.

The bias is easily understood by considering the most simple model of image motion estimation. Local image measurements define at a point one component of the image flow, for example, if the image measurements are the spatial derivatives E_x, E_y and the temporal derivative E_t of the image intensity function, E, at point $\mathbf{r}_i$ the flow $\dot{\mathbf{r}} = (u, v, 0)$ is constrained by the equation

$$E_{x_i} u + E_{y_i} v E_{t_i} = E_{t_i} \tag{3}$$

We consider additive zero mean independent noise in the derivatives of E. Let n be the number of indices i to which (3) applies, which for convenience we write in matrix form as

$$\mathbf{E}_s \mathbf{u} = \mathbf{E}_t \tag{4}$$

where $\mathbf{E}_s$ is the n by 2 matrix incorporating the spatial derivatives E_{x_i}, E_{y_i}, $\mathbf{E}_t$ is an n by 1 matrix incorporating the E_{t_i}, and $\mathbf{u}$ is the vector (u, v).

If (4) is solved by standard least squares estimation the solution is

$$\mathbf{u} = \left(\mathbf{E}_s{}^\top \mathbf{E}_s\right)^{-1} \mathbf{E}_t{}^\top \mathbf{u}$$

and this solution, if there are errors in the spatial derivatives, is biased. The expected value, $E(\mathbf{u})$, of $\mathbf{u}$ amounts to

$$E(\mathbf{u}) = \mathbf{u}' - n\sigma_s^2 \left(\mathbf{E}_s{}'^\top \mathbf{E}_s{}'\right)^{-1} \mathbf{u}',$$

where primes are used here to denote actual values and σ_s^2 is the variance of the noise in the spatial derivatives. This expected value of $\mathbf{u}$ is an underestimate in length and in a direction closer to the majority of gradient directions in the patch.

It has been shown in [13], [14] that it is not only gradient based techniques and least squares estimation that produce bias, but other techniques (frequency-domain, correlation techniques, and different estimation procedures) as well. Correcting for the bias would require knowledge of the noise parameters and these are difficult to obtain. In the case of the Ouchi illusion the 3D motion, which is caused by small eye movements or rapid motion of the paper, changes very quickly so the system does not have enough data available to acquire the statistics of the noise. In more natural situations, however, the system can reduce the bias problem if it performs feedback processes. This way, after some estimates of motion and structure have been obtained, image measurements from larger image regions can be considered to acquire more accurate statistical noise sampling.

6.2 *Enigma* and Variants

The static figure, *Enigma*, painted by Leviant and shown in Fig. 4, consists of radial lines emanating from the center of the image and interrupted by a set of concentric, uniformly colored rings. Upon extended viewing of *Enigma*, most humans perceive illusory movement inside the rings which keeps changing direction. Based on a study of positron emission tomography, researchers in neurophysiology [20], [21] have argued that higher level processes of motion interpretation in the human brain are responsible for the perception of this illusion. To be more specific, these researchers compared the cerebral blood flow during the viewing of *Enigma* and a similar reference image which does not give rise to an illusion, and they found the values comparable in the early motion processing area V1, but found significantly higher values for *Enigma* in V5 (or MT)—a later area in visual processing.

The spatial texture in *Enigma* is closely related to the patterns described in Sect. 5 that form constraints for 3D motion estimation. This lets us hypothesize that the illusion essentially is due to higher level activity. It is due to the

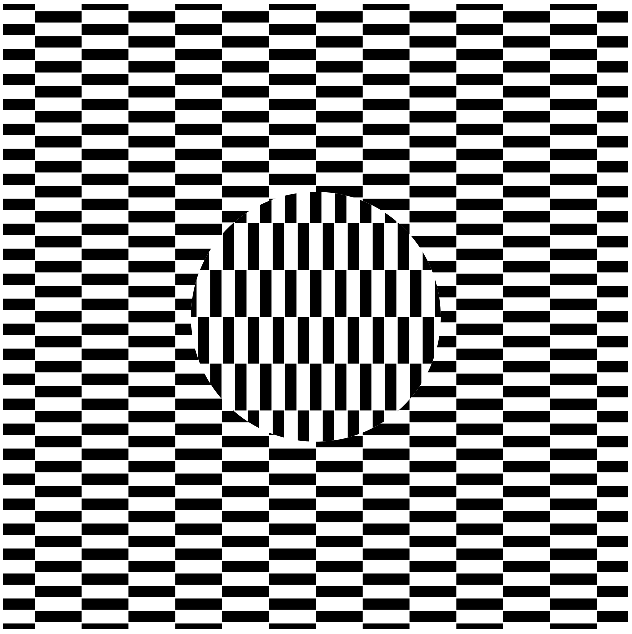

Fig. 3. A pattern similar to one by Ouchi [18]

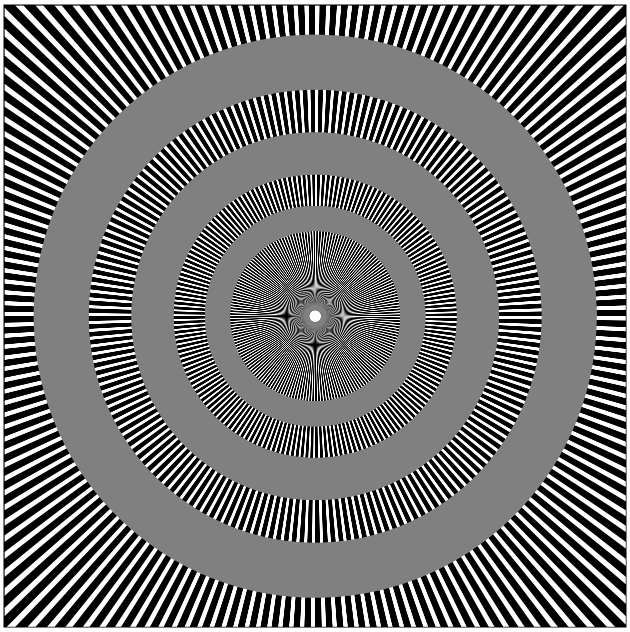

Fig. 4. *Enigma*, giving rise to the Leviant illusion: Fixation at the center results in perception of a rotary motion inside the rings

particular architecture of the visual system which solves the tasks of 3D motion estimation and segmentation through feedback processes [12]. The explanation in more detail is as follows. Small eye movements give rise to retinal motion signals. Because of *Enigma*'s particular structure motion signals occur only in the areas of radial lines; within the rings there is no texture and thus also no early motion signals can be picked up there. The motion signals, that is the normal flow vectors perpendicular to the radial lines, all belong to exactly one copoint field, the one corresponding to $\mathbf{s}$ being the axis passing through the center of the image. During the next processing stages, normal flow vectors from increasingly larger areas are combined to estimate the rigid motion and during these stages some spatial and temporal integration of smoothing takes place which causes motion signals also within the rings. After an estimate of 3D motion is found, this information is fed back to the earlier processing elements which then have to perform exact localization of independently moving objects. At this stage, temporal integration cancels out the flow information within the areas covered by the rays. It, however, does not cancel out the motion within the homogeneous regions since the responses from the processing elements there do not contradict an existing motion in these areas. Since, at this stage, the system's task is to accurately segment the scene, the edges of the homogeneous regions are perceived as motion discontinuities and motion within the homogeneous regions is reinforced. The defining feature of *Enigma* is that all the spatial gradients in its patterns excite only one class of vector fields involved in 3D motion estimation—a copoint field. To test our hypothesis we created figures based on the same principle as *Enigma*—that is with black and white rays giving rise to normal flow vectors corresponding to exactly one coaxis or copoint vector field and homogeneous areas perpendicular to these rays. For an example, see Fig. 5. We found that it is exactly these patterns giving rise to illusory motion. Other figures based on the same principle (rays interrupted by perpendicular homogeneous areas) which do not correspond to the vector fields involved in 3D motion estimation have been found not to give rise to an illusion, thus supporting the hypothesis.

7 Complex Actions

One may be able to derive a simple procedure for recognizing some action based on the statistics of local image movement. There exists a variety of such techniques in the current literature mostly applied to human movement or facial expressions, but according to our account of what it means to understand an action, representations that enable this sort of recognition are simply inadequate. What we need are representations that can visualize the action under consideration. Let us summarize a number of key points that are important in figuring out the nature of action representations:

1. Action representations should be view independent. We are able to recognize and visualize actions regardless of viewpoint.

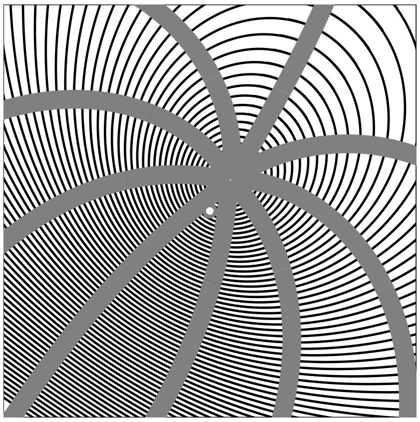

Fig. 5. Example of an illusory movement figure based on a coaxis field. Here the focal length is chosen to be about equal to the size of the image. Thus, the illusion is best experienced at a short distance

2. Action representations capture dynamic information which is manifested in a long image sequence. Put simply, it is not possible to understand an action on the basis of a small sequence of frames (viewpoints).
3. It is the combination of shape and movement that makes up an action representation.

A number of issues become clear from the impositions of the constraints listed above. *First*, to understand human action, a model of the human body should be available. *Second*, the movement of the different parts of the body relative to each other and to the environment should be recovered in a way that allows matching. *Third*, since action amounts to the movement of 3D space, action representations should contain information about both components. Let us further examine those issues in order. Acquiring a sophisticated model of the human body, although of extreme practical importance, is not essential to our discussion. We would like to caution, however, against learning approaches. It is probably an impossible task to learn such a model from visual data in a reasonable time. It is quite plausible that humans are born with such a model available to them. After all, it is a model of themselves. Later, we will describe technological solutions to this question. The second issue deserves more attention, because it introduces constraints. The previous section discussed the difficulties in acquiring 3D motion and shape, when normal motion fields were available in the whole visual field. From this we conclude that it is impossible to acquire exact 3D motion information from a very small part of the visual field, that is, it is not feasible to accurately estimate the 3D motion of the different rigid parts of the body from image sequences of it. Whatever the action representation is, it appears very hard to achieve recognition in a bottom-up fashion. The concept of feedback explained in the previous section acquires here a major importance. Recognizing action should be mostly a top-down process. Finally, the third issue reflects the need to incorporate both shape and motion in the representation. This means that any motion representation should be spatially organized. There are, however, various options here. Do we explicitly represent movement in terms of motion fields, or do we do it implicitly in terms of forces, dynamics, and inertial constraints? Both options have their pros and cons.

8 How Could Action Representations Be?

To gain insights on action representations we consider them in a hierarchy with two more kinds of representations and the properties of the possible mappings among them. *First* there is the image data, that is, videos of humans in action. Considering the cue of motion, then our image data amounts to a sequence of normal flow fields computed from the videos. The second kind of representations are intermediate descriptions encoding information about 3D space and 3D motion, estimated from the input (video). These representations consist of a whole range of descriptions of different sophistication encoding partially the space-time geometry, and they are scene dependent. Finally, we have the action representations themselves, which are scene independent.

Before we start considering the relationships in this hierarchy, let us introduce a novel and very useful tool in the study of action. It is a tool because it can help us in this study, but at the same time it amounts to a theoretical intermediate representation that combines the advantages of view independence with the capture of the action as a whole. It is a theoretical tool in the same way that the image motion field is a theoretical construct. An intermediate representation for the specific action in view is then a sequence of evolving 3D motion fields and it is the most sophisticated intermediate description that could be obtained. Acquiring this representation is no simple matter, but it can be achieved by employing a very large number of viewpoints. We have established in our laboratory a multi-camera network (sixty-four cameras, Kodak ES-310, providing images at a rate of eighty-five frames per second; the video is collected directly on disk—the cameras are connected by a high-speed network possessing sixteen dual processor Pentium 450s). By observing human action from all viewpoints we are able to create a sequence of 3D motion fields [10]. See Fig. 6 that displays a schematic description of our facility. The solution requires a calibration of the camera network [1] and proceeds by carving the rays in space using image intensity and image motion values.

Let us now consider the mapping from the second level of the hierarchy (intermediate descriptions and, in particular, 3D motion fields) to the scene independent action representations (top level). This mapping should be such that it extracts from a specific action quantities of a generic character common to all actions of the same type. These quantities most probably take the form of spatiotemporal patterns in four dimensions.

One way of obtaining such patterns is to perform statistics on a large enough sample. Considering, for example, a particular action (e.g., walking), we can obtain data in the multi-camera laboratory described before for a large number

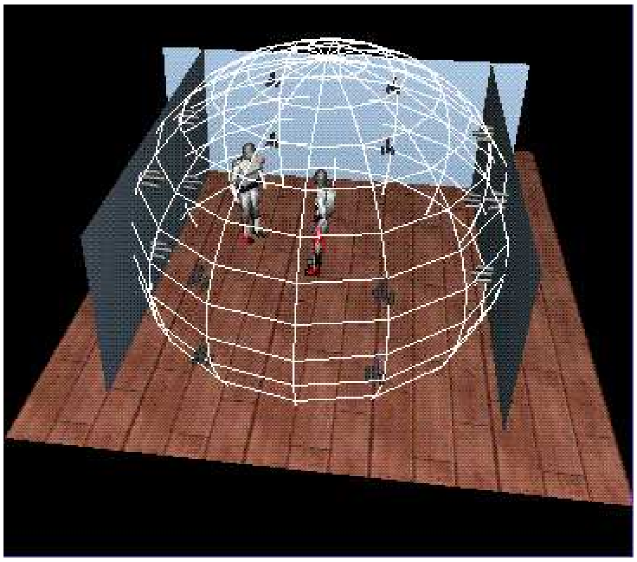

Fig. 6. A negative spherical eye

of individuals. In each case we can obtain a 3D motion field and thus have at our disposal a large number of 3D motion fields. A number of statistical techniques, such as principal component analysis, can reduce the dimensionality of this space and describe it with a small number of parameters, which parameters will in turn constitute a representation for the generic action under consideration.

Another way of obtaining these patterns would be to employ a geometric approach as well. Since a human action consists of a large number of different rigid motions, one possibility is to obtain these rigid motions from the 3D motion field, encode their relationships, and study invariances related to symmetry, and geometric quantities in space-time (angles, velocities, accelerations, periodicity, etc.).

Considering now the mappings from the top level to the image data and to the intermediate representations we can gain further insight. For the sake of recognition, action descriptions will have to be projected on the image and matched against the intermediate descriptions and the image data. But, as explained in previous sections, these intermediate descriptions are errorful and constitute distorted versions of the scene. In addition, there is variability among actions of the same type. In trying to perform this matching in space we will have to revise both the action representations and the nature of the intermediate descriptions so that the matching can be facilitated. In actual fact we may have to develop specific spatiotemporal patterns in the image data for this goal and these patterns will constitute some form of action representations for quick recognition. To conclude, the interplay among the mappings between the introduced hierarchy levels seems to be a feasible and structured way for building powerful action descriptions. This constitutes our current research efforts.

9 Conclusions

Being truthful to Hartmann's and Lorenz's approach regarding evolution and the structure of intelligent systems, we wish to study intelligent systems as they actually are, using the present structures in their bodies as our data. Considering representations for a number of important complex actions, we have two options about their origin: either the system learned them or it was born with them. We find the argument about the innate character of (at least some) action representations very appealing, because humans have retained the same basic structure and have been performing, for the most part, the same repertoire of actions for a very long time. It makes good sense to build these representations using all available knowledge and technology. This paper offered a number of ways for accomplishing this goal.

References

1. P. Baker and Y. Aloimonos. Complete calibration of a multi-camera network. In *Proc. IEEE Workshop on Omnidirectional Vision*, pages 134–141, Los Alamitos, CA, 2000. IEEE Computer Society.
2. T. Brodský, C. Fermüller, and Y. Aloimonos. Structure from motion: Beyond the epipolar constraint. *International Journal of Computer Vision*, 2000. In press. Also available as CfAR Technical Report CAR-TR-911.
3. L. Cheong, C. Fermüller, and Y. Aloimonos. Effects of errors in the viewing geometry on shape estimation. *Computer Vision and Image Understanding*, 71:356–372, 1998.
4. C. Fermüller and Y. Aloimonos. Direct perception of three-dimensional motion from patterns of visual motion. *Science*, 270:1973–1976, 1995.
5. C. Fermüller and Y. Aloimonos. Qualitative egomotion. *International Journal of Computer Vision*, 15:7–29, 1995.
6. C. Fermüller and Y. Aloimonos. Vision and action. *Image and Vision Computing*, 13:725–744, 1995.
7. C. Fermüller and Y. Aloimonos. On the geometry of visual correspondence. *International Journal of Computer Vision*, 21:223–247, 1997.
8. C. Fermüller and Y. Aloimonos. Ambiguity in structure from motion: Sphere versus plane. *International Journal of Computer Vision*, 28:137–154, 1998.
9. C. Fermüller and Y. Aloimonos. Observability of 3D motion. *International Journal of Computer Vision*, 2000. In press.
10. C. Fermüller, Y. Aloimonos, P. Baker, R. Pless, J. Neumann, and B. Stuart. Multi-camera networks: Eyes from eyes. In *Proc. IEEE Workshop on Omnidirectional Vision*, pages 11–18, Los Alamitos, CA, 2000. IEEE Computer Society.
11. C. Fermüller, L. Cheong, and Y. Aloimonos. Visual space distortion. *Biological Cybernetics*, 77:323–337, 1997.
12. C. Fermüller, R. Pless, and Y. Aloimonos. Families of stationary patterns producing illusory movement: Insights into the visual system. *Proc. Royal Society, London B*, 264:795–806, 1997.

13. C. Fermüller, R. Pless, and Y. Aloimonos. The Ouchi illusion as an artifact of biased flow estimation. *Vision Research*, 40:77–96, 2000.
14. C. Fermüller, D. Shulman, and R. Pless. The statistics of optical flow. Technical Report CAR-TR-928, Center for Automation Research, University of Maryland, 1999.
15. N. Hartmann. *Der Aufbau der Realen Welt.* de Gruyter, Berlin, 1964.
16. J. J. Koenderink and A. J. van Doorn. Relief: Pictorial and otherwise. *Image and Vision Computing*, 13:321–334, 1995.
17. K. Lorenz. *Behind the Mirror: A Search for a Natural History of Human Knowledge.* Harcourt Brace Jovanovich, New York, 1977.
18. H. Ouchi. *Japanese and Geometrical Art.* Dover, 1977.
19. M. Polanyi. *Personal Knowledge: Towards a Post-Critical Philosophy.* The University of Chicago Press, Chicago, 1958.
20. S. M. Zeki. The cortical *enigma*: A reply to Professor Gregory. *Proc. Royal Society, London B*, 257:243–245, 1994.
21. S. M. Zeki. Phenomenal motion seen through artificial intra-ocular lenses. *Proc. Royal Society, London B*, 260:165–166, 1995.