MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY

A.I. Memo No. 1435 July, 1993

Why Stereo Vision is Not Always About 3D Reconstruction

W. Eric L. Grimson

Abstract

A common assumption of stereo vision researchers is that the goal of stereo is to compute explicit 3D information about a scene, to support activities such as navigation, hand-eye coordination and object recognition. This paper suggests reconsidering what is required of a stereo algorithm, in light of the needs of the task that uses its output. We show that very accurate camera calibration is needed to reconstruct accurate 3D distances, and argue that often it may be difficult to attain and maintain such accuracy. We further argue that for tasks such as object recognition, separating object from background is of central importance. We suggest that stereo can help with this task, without explicitly computing 3D information. We provide a demonstration of a stereo algorithm that supports separating figure from ground through attentive fixation on key features.

This report describes research done at the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology. Support for the laboratory's artificial intelligence research is provided in part by the Advanced Research Projects Agency of the Department of Defense under Office of Naval Research contract N00014-91-J-4038. The author was also supported by NSF contract IRI-8900267. The author can be reached via welg@ai.mit.edu.

1 Introduction

The title of this article is, of course, deliberately provocative, in part to capture the reader's attention, but in part also to make a point. A common assumption of researchers working in stereo vision is that the goal of stereo is to compute explicit 3D information about a scene, in order to support activities such as navigation, hand-eye coordination and object recognition. While there are applications in which such information can be accurately computed, these domains require very accurate camera calibration information. We suggest that in many applications, it may be difficult to attain and maintain such accurate information, and hence we suggest that it may be worthwhile to reconsider what is required of a stereo algorithm, in light of the needs of the task that uses stereo's output. In particular, we examine the role of stereo in object recognition, arguing that it may be more effective as a means of separating objects from background, than as a provider of 3D information to match with object models. To support this argument, we provide a demonstration of a stereo algorithm that separates figure from ground through attentive fixation on key features, without explicitly computing actual 3D information.

2 Some Stereo Puzzles

It has been common in recent years within the computer vision community to consider the stereo vision problem as consisting of three key steps [23], [27]:

- Identify a particular point in one image (say the left).
- Find the point in the other (say right) image that is a projection of the same scene point as observed in the first image.
- Measure the disparity (or difference in projection) between the left and right image points. Use knowledge of the relative orientation of the two camera systems, plus the disparity, to determine the actual distance to the imaged scene point.

These steps are repeated for a large number of points, leading to a 3D reconstruction of the scene, at those points.

There are many variations on this theme, including whether to use distinctive features such as edges or corners as the points to match, or to simply use local patches of brightness values, what constraints to apply to the search for corresponding matches (e.g. epipolar lines, similar contrast, similar orientation, etc.), and whether to restrict the relative orientation of the cameras (e.g. to parallel optic axes). Nonetheless, it has been commonly assumed for some time that the hard part of the problem is solving for the correspondence between left and right image features. Once one knows which points match, it has been assumed that measuring the disparity is trivial, and that solving for the distance simply requires using the geometry of the cameras to invert a simple trigonometric projection.

This sounds fine, but let's consider some puzzles about this approach. The first puzzle is a perceptual one, illustrated in Figure 1. This illusion is a depth variant on the standard Cornsweet illusion in brightness, and is due to Anstis et al. [2] (see also [37]). It consists of a physical object with two coplanar regions separated by a sharp discontinuity, where the regions immediately to the sides of the discontinuity are smoothly curved. These surfaces are textured with random dot paint, to make them visible to the viewer. Subjects are then asked to determine whether the two planar regions are coplanar, or separated in depth, and if it is the latter, which surface is closer and by how much. Although physically the two surfaces are in fact coplanar, subjects consistently see one of the two surfaces as closer (the left side in the case of Figure 1). The reported error is $.5cm$ and is consistent for three different view distances: $72, 145$ and $290cm$.

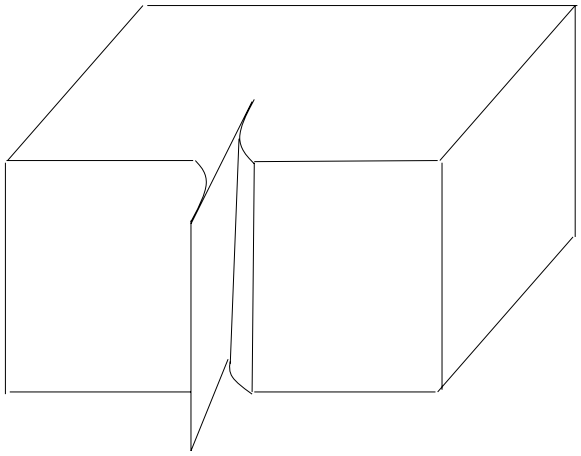


Figure 1: Cornsweet illusion in depth.

This is clearly surprising if one believes that the above description of the stereo process holds for biological as well as machine solutions. In particular, if the human system maintains a representation of reconstructed distance, and if that representation is accessible to queries, then it is difficult to see how human observers could consistently make such a mistake.

Additional stereo puzzles are provided in [40], which the authors use to argue that depth is not computed directly in humans, but is reconstructed from non-zero second differences in depth. As a consequence, they demonstrate that human stereo vision is blind to constant gradients of depth. Similar observations on the role of disparity gradients in reconstructing depth are given by [44].

It need not be the case that machine stereo systems make the same "mistakes" as human observers, but the existence of such an illusion for humans raises an interesting question about the basic assumptions of approaches that reconstruct distance.

Consider a second puzzle about the approach of matching features, then using trigonometry to convert into depth. As noted, for years stereo researchers have assumed that the correspondence problem was the hard part of the task. Once correct correspondences were found, the reconstruction was a simple matter of geometry. This is true in principle, but it relies on finding

the intrinsic parameters of the camera systems and the extrinsic parameters relating the orientation of the two cameras. While solutions exist for finding these parameters (e.g. [41]), such solutions appear to be numerically unstable [45, 43]. If one does not perform very careful calibration of the camera platform, the result will be very noisy reconstructed distances.

Of course, there are circumstances in which careful calibration can be performed, and in these cases, extremely accurate reconstructions are possible. A good example of this is automated cartographic reconstruction from satellite imagery, where commercial systems can provide maps with accuracy on the order of a few meters, from satellite photography [19]. On the other hand, if the cameras are mounted on a mobile robot that is perturbed as it moves through the environment, then it may be more difficult to attain and maintain careful calibration. Thus, we see that there are some suggestions that human observers do not reconstruct depth, and some suggestions that one needs very careful calibration (which is often hard to guarantee) in order to do this. We will explore the calibration sensitivity issue in section 3.

Given this puzzle, it is worth stepping back to ask what one needs from the output of a stereo algorithm. Aside from specialized tasks such as cartography, the two standard general application areas are navigation and recognition. Interestingly, Faugeras [8] (see also [39]) has recently argued that one can construct and maintain a representation of the scene structure around a moving robot, without a need for careful calibration. Moreover, the solution involves using relative coordinate systems to represent the scene, so that there is no metrical reconstruction of the scene.

What about object recognition? We have found it convenient to separate the recognition problem into three pieces [11]:

- **Selection:** Extract subsets of the data features likely to have come from a single object.
- **Indexing:** Look up those object models that could have given rise to one such selected subset.
- **Correspondence:** Determine if there is a way of matching model features to data features that is consistent with a legal transformation of the model into the data.

We have argued [11] that for many approaches to recognition, the first stage is the crucial one. In many cases, it reduces the expected complexity of recognition from exponential to low-order polynomial, and in many cases, it is necessary to keep the false positive rates under control. If we accept that the hard part of recognition is selection, rather than correspondence, then this has an interesting implication for stereo. If stereo were mainly oriented towards solving the correspondence problem, it is natural to expect that it needs to deliver accurate 3D data that can be compared to 3D models. But if stereo is mainly intended to help with the selection problem, then one no longer needs to extract exact 3D reconstructions, one simply needs stereo to identify data feature subsets that are roughly in the same depth range, or equivalently do not have large variations in disparity. We will examine a modified stereo algorithm in section 4 that takes advantage of this observation.

If one accepts that stereo is primarily for segmentation, not for 3D reconstruction, this leads to the further question of whether recognition of 3D objects can be done without explicit 3D input data. A number of recent techniques have shown interesting possibilities along these lines; for example, the recent development of the linear combinations method [42] suggests that one could use stored 2D images of a model to generate an hypothesized 2D image which can then be compared to the observed image. Again, one does not need to extract exact 3D data. It is also intriguing along these lines to observe that some physiological data [34, 35] may also support the idea of the human system solving 3D recognition from purely 2D views. Of course, it is possible to solve the recognition problem by matching reconstructed 3D stereo data against 3D models [27].

To summarize, we consider three main points:

- the human stereo system may not directly compute 3D depth, suggesting that humans may not need explicit depth;
- small inaccuracies in measuring camera parameters can lead to large errors in computed depth, suggesting that we may not be able to compute explicit depth accurately;
- the critical part of object recognition is figure/ground separation, which may not require explicit depth information.

We will use this to argue that stereo can contribute to the efficient solution of the object recognition problem, without the need for accurate calibration and without the need for explicit depth computation. In this case, the importance of eye movements or related control strategies is increased, causing us to reexamine the structure of stereo algorithms. Similar questions have been by systems that use actively controlled stereo eye-head systems to acquire depth information (for example, [1, 5, 6, 7, 9, 20, 30, 38, 33]).

3 Why Reconstruction is Too Sensitive

While our first point is based primarily on earlier psychophysical observations, the second point bears closer examination. Let's look in more detail at the problem of computing distance from stereo disparity. Suppose our two cameras have points of projection located at $\mathbf{b}_\ell$ and $\mathbf{b}_r$, measured in some world coordinate system. Assume that the optic axes are $\hat{\mathbf{z}}_\ell$ and $\hat{\mathbf{z}}_r$, and that both cameras have the same focal length f (though we could easily relax this to have two different focal lengths).

In this case, we can represent the left image plane by

$$\{\mathbf{v} | \langle \mathbf{v}, \hat{\mathbf{z}}_\ell \rangle = d_\ell\}$$

where $\langle ., . \rangle$ represents an inner (or dot) product. The principal point (or image center) is given by

$$\mathbf{c}_\ell = \mathbf{b}_\ell + f\hat{\mathbf{z}}_\ell$$

where we have chosen to place the image plane in front of the projection point, to avoid the inversion of the coordinate axes of the image. Since we know that this point lies on the image plane, we can deduce the constant offset, so that the left image plane is given by

$$\{\mathbf{v} | \langle \mathbf{v} - \mathbf{b}_\ell, \hat{\mathbf{z}}_\ell \rangle = f\} .$$

A similar representation holds for the right image plane.

Now an arbitrary scene point $\mathbf{p}$ maps, under perspective projection, to a point $\mathbf{p}_\ell$ on the left image plane, given by

$$\mathbf{p}_\ell = \mathbf{b}_\ell + \frac{f(\mathbf{p} - \mathbf{b}_\ell)}{\langle \mathbf{p} - \mathbf{b}_\ell, \hat{\mathbf{z}}_\ell \rangle}$$

and for convenience we write this as

$$\mathbf{p}_\ell = \mathbf{c}_\ell + \mathbf{d}_\ell$$

where $\langle \mathbf{d}_\ell, \hat{\mathbf{z}}_\ell \rangle = 0$. Here $\mathbf{d}_\ell$ is an offset vector in the image plane from the principal point:

$$\mathbf{d}_\ell = f \left(\frac{\hat{\mathbf{z}}_\ell \times ((\mathbf{p} - \mathbf{b}_\ell) \times \hat{\mathbf{z}}_\ell)}{\langle \mathbf{p} - \mathbf{b}_\ell, \hat{\mathbf{z}}_\ell \rangle} \right) .$$

Note that we haven't specified the world coordinate system yet, and we can now take advantage of that freedom. In particular, we choose the origin of the world coordinate system to be centered between the projection points, so that $\mathbf{b}_\ell = -\mathbf{b}_r = \mathbf{b}$.

By subtracting $\mathbf{d}_r$ from $\mathbf{d}_\ell$, we get the following relationship

$$\langle \mathbf{p} - \mathbf{b}_\ell, \hat{\mathbf{z}}_\ell \rangle \, \mathbf{d}_\ell - \langle \mathbf{p} - \mathbf{b}_r, \hat{\mathbf{z}}_r \rangle \, \mathbf{d}_r = f \left[-\mathbf{b}_\ell + \mathbf{b}_r - \langle \mathbf{p} - \mathbf{b}_\ell, \hat{\mathbf{z}}_\ell \rangle \, \hat{\mathbf{z}}_\ell + \langle \mathbf{p} - \mathbf{b}_r, \hat{\mathbf{z}}_r \rangle \, \hat{\mathbf{z}}_r \right] \quad (1)$$

For the special case of the origin centered between the projection points, this becomes

$$\langle \mathbf{p} - \mathbf{b}, \hat{\mathbf{z}}_\ell \rangle \, \mathbf{d}_\ell - \langle \mathbf{p} + \mathbf{b}, \hat{\mathbf{z}}_r \rangle \, \mathbf{d}_r = f \left[-2\mathbf{b} - \langle \mathbf{p} - \mathbf{b}, \hat{\mathbf{z}}_\ell \rangle \, \hat{\mathbf{z}}_\ell + \langle \mathbf{p} + \mathbf{b}, \hat{\mathbf{z}}_r \rangle \, \hat{\mathbf{z}}_r \right] . \quad (2)$$

We can isolate components of $\mathbf{p}$ with respect to each of the two optic axes, by taking the dot product of both sides of equation 1 or 2 with respect to these unit vectors. This gives us two linear equations (assuming that $\hat{\mathbf{z}}_\ell \neq \hat{\mathbf{z}}_r$), which we can solve to find these components of $\mathbf{p}$. Adding them together yields:

$$\langle \mathbf{p}, \hat{\mathbf{z}}_\ell + \hat{\mathbf{z}}_r \rangle = \frac{\left[(f^2 + \alpha\beta) \langle \mathbf{b}, \hat{\mathbf{z}}_\ell - \hat{\mathbf{z}}_r \rangle + 2f \langle \mathbf{b}, \beta \hat{\mathbf{z}}_\ell - \alpha \hat{\mathbf{z}}_r \rangle \right]}{\alpha\beta - f^2} , \quad (3)$$

where

$$\begin{aligned} \alpha &= \langle \mathbf{d}_r + f\hat{\mathbf{z}}_r, \hat{\mathbf{z}}_\ell \rangle \\ \beta &= \langle \mathbf{d}_\ell + f\hat{\mathbf{z}}_\ell, \hat{\mathbf{z}}_r \rangle . \end{aligned}$$

To explore how this computation of depth from stereo measurements depends on the accuracy of the calibrated parameters and the disparity measurements, we consider the symmetric case of:

$$\begin{aligned} \hat{\mathbf{z}}_\ell &= \cos\gamma \hat{\mathbf{z}} + \sin\gamma \hat{\mathbf{x}} \\ \hat{\mathbf{z}}_r &= \cos\gamma \hat{\mathbf{z}} - \sin\gamma \hat{\mathbf{x}} \\ \hat{\mathbf{b}} &= -b\hat{\mathbf{x}} \end{aligned}$$

where $\hat{\mathbf{x}}$ is chosen as the direction of the vector connecting the two centers of projection, and where the two cameras make a symmetric (though opposite signed) gaze angle γ with the $\hat{\mathbf{z}}$ axis, and where the offset of each camera from the origin is the same. In this case, substitution and manipulation leads to

$$\langle \mathbf{p}, \hat{\mathbf{z}} \rangle \cos\gamma = \frac{2b \left(f^2 \cos^2\gamma + d_r \sin\gamma \right) \left(f^2 \cos^2\gamma - d_\ell \sin\gamma \right)}{2\sin\gamma \left(f^2 \cos^2\gamma + d_r d_\ell \right) - f \left(\cos^2\gamma - \sin^2\gamma \right) (d_r - d_\ell)} \quad (4)$$

where we have let

$$\begin{aligned} d_r &= \langle \mathbf{d}_r, \hat{\mathbf{z}} \rangle \\ d_\ell &= \langle \mathbf{d}_\ell, \hat{\mathbf{z}} \rangle . \end{aligned}$$

Note that in the special case of parallel optic axes ($\gamma = 0$), this reduces to

$$\langle \mathbf{p}, \hat{\mathbf{z}} \rangle = \frac{2fb}{d_\ell - d_r}$$

which is exactly what one would expect, since $d_\ell - d_r$ is simply the disparity at this point.

For convenience, call $Z = \langle \mathbf{p}, \hat{\mathbf{z}} \rangle$. This equation tells us how to compute the depth Z, given measurements for the camera parameters f, b, γ and the two principal points $\mathbf{c}_\ell, \mathbf{c}_r$ as well as the individual measurements of displacement $\mathbf{d}_\ell, \mathbf{d}_r$ (or equivalently d_ℓ and d_r).

The question we want to consider is how accurately do we need to know these parameters? There has been some previous analysis of stereo error in the literature, primarily focused on the effects of pixel quantization [43, 28, 25], although some analysis of the effects of camera parameters has also been done [45, 44]. Here we are primarily interested in the effects of the camera parameters.

For sake of simplicity, we will assume that γ is small. For example, if the cameras are fixated at a target 1 meter removed, with an interocular separation of $10cm$, then $\gamma \approx .05$ radians, or if the fixation target is .5 meters off, then $\gamma \approx .1$ radians. In the second case, the small angle approximation will lead to an error in $\cos\gamma$ of at most .005 and an error in $\sin\gamma$ of at most .0002. Using the small angle approximation leads to

$$Z \approx 2b \frac{f^2 + \gamma f (d_r - d_\ell)}{2\gamma(f^2 + d_r d_\ell) - f(d_r - d_\ell)} \quad (5)$$

If we rewrite this, isolating depth in terms of interocular units ($2b$), and image offsets in terms of focal length (or equivalently in terms of angular arc), we get:

$$\frac{Z}{2b} \approx \frac{1 + \gamma \frac{d_r - d_\ell}{f}}{2\gamma - \frac{d_r - d_\ell}{f} + 2\gamma \frac{d_r}{f} \frac{d_\ell}{f}} . \quad (6)$$

In some cases it is more convenient to consider this expression in terms of relative units, that is representing depth in terms of interocular spacing, by using

$$Z' = \frac{Z}{2b}$$

and to use disparities as angular arcs by using

$$d'_r = \frac{d_r}{f} \qquad d'_\ell = \frac{d_\ell}{f} .$$

In this case, we have

$$Z' \approx \frac{1+\gamma(d'_r - d'_\ell)}{2\gamma - (d'_r - d'_\ell) + 2\gamma d'_r d'_\ell}. \quad (7)$$

By taking partial derivatives of this equation with respect to each of the parameters of interest (which we treat as independent of one another), we arrive at the following expressions for the relative change in computed depth as a function of the relative error in measuring the parameters:

$$\begin{aligned}
\left|\frac{\Delta Z}{Z}\right| &= \left|\frac{\Delta b}{b}\right| \\
\left|\frac{\Delta Z}{Z}\right| &= \left|\frac{\Delta d_r}{f}\right| \left|\frac{\gamma + Z' - 2\gamma d'_\ell Z'}{1+\gamma(d'_r - d'_\ell)}\right| \\
\left|\frac{\Delta Z}{Z}\right| &= \left|\frac{\Delta d_\ell}{f}\right| \left|\frac{\gamma + Z' + 2\gamma d'_r Z'}{1+\gamma(d'_r - d'_\ell)}\right| \\
\left|\frac{\Delta Z}{Z}\right| &= \left|\frac{\Delta f}{f}\right| \left|\frac{(\gamma + Z')(d'_r - d'_\ell) - 4\gamma d'_r d'_\ell Z'}{1+\gamma(d'_r - d'_\ell)}\right| \\
\left|\frac{\Delta Z}{Z}\right| &= |\Delta\gamma| \left|\frac{d'_r - d'_\ell - 2Z'(1 + d'_r d'_\ell)}{1+\gamma(d'_r - d'_\ell)}\right|
\end{aligned}$$

If we use standard viewing geometries (i.e. focal length much larger than individual pixel size, γ small), we can approximate these expressions as follows:

$$\left|\frac{\Delta Z}{Z}\right| \approx \left|\frac{\Delta b}{b}\right| \quad (8)$$

$$\left|\frac{\Delta Z}{Z}\right| \approx \left|\frac{\Delta d_r}{f}\right| |Z'| \quad (9)$$

$$\left|\frac{\Delta Z}{Z}\right| \approx \left|\frac{\Delta d_\ell}{f}\right| |Z'| \quad (10)$$

$$\left|\frac{\Delta Z}{Z}\right| \approx \left|\frac{\Delta f}{f}\right| |Z'(d'_r - d'_\ell)| \quad (11)$$

$$\left|\frac{\Delta Z}{Z}\right| \approx 2\,|Z'|\,|\Delta\gamma| \quad (12)$$

We note that related error expressions were obtained in [43], although the focus there was on the effects of errors in the matching of image features and the quantization of image pixels on the accuracy of recovered depth.

Our concern is how uncertainty in measuring the camera parameters impacts the computed depth. Ideally, we would like a linear relationship, so that, for example, a 1 percent error in computing a parameter would result in at most a 1 percent error in depth.

To explore this, we consider two cases: a camera system with $15mm$ focal length and $.015mm$ pixels so that a pixel subtends an angular arc of .001 radians; and the human visual system, where the fovea has a receptor packing subtending approximately .00014 radians.

By equation 8, relative errors in computed depth due to mismeasurement of the baseline separation are generally quite small. For example, a 1% relative error in measuring the baseline will result in a 1% relative error in the computed distance.

Equations 9 and 10 are essentially the same. They show a non-linear effect, in that the relative error in computing depth is a function both of the relative error in computing the position of each image point with respect to the global coordinate frame, and more importantly is a function of the distance of the object from the viewer, in units of interocular separation ($2b$). Thus, the relative error will get much worse for more distant objects. If we let the pixel error in measuring position be k, then using a standard pixel size and focal length, the relative error in depth is

$$\frac{k}{10^3}\left|\frac{Z}{2b}\right|$$

for our camera system. To see how large this can get, we need to understand what can contribute to k. Effects include:

- image based localization errors
- image based matching errors
- registration errors between the image and the world coordinates due to:
  - principal points
  - image orientation

Uncertainty and smoothing effects in the edge detector will affect the first source of error, but typically will only cause errors on the order of a few pixels. Since matching errors by definition must lead to incorrect depth reconstructions, we ignore them in our analysis. The second major source of error comes from converting the image pixel measurements to world coordinates, and here there are two main sources. One is that all of our disparity measurements in the analysis above were based on the displacement of features from the principal points. This requires that we measure those principal points accurately [21], and this is particularly important since in many cameras, the principal point can often be tens of pixels away from the center of the sensor array. For example, the CCD cameras in use in one of our stereo setups have principal points displaced from the image array center by 30 pixels in x and 1 pixel in y for the left camera and 18 pixels in x and 3 pixels in y for the right camera. Methods in the literature for locating the principal points [21] are reported to have residual errors of at most 6 pixels.

Finally, we need to know the orientation of the camera rasters with respect to the world axes. Even if we ignore the effects of gaze angle, rotation about the optic axis (cyclotorsion) can result in an error in the disparity offset with respect to the interocular baseline. Since this error goes with the cosine of the rotation, we expect the effects of such error to be small.

If we have found the principal points and the orientation of the cameras with respect to world coordinates accurately, then k will typically be on the order of a few pixels. If we have not, k can easily be on the order of tens of pixels. To see the effect of this on reconstructed depth, Figure 2 shows plots of the percentage relative error in computing depth, as a function of the distance to the object (measured in units of interocular separation), for the case of $k = 1$ and $k = 10$. For an object

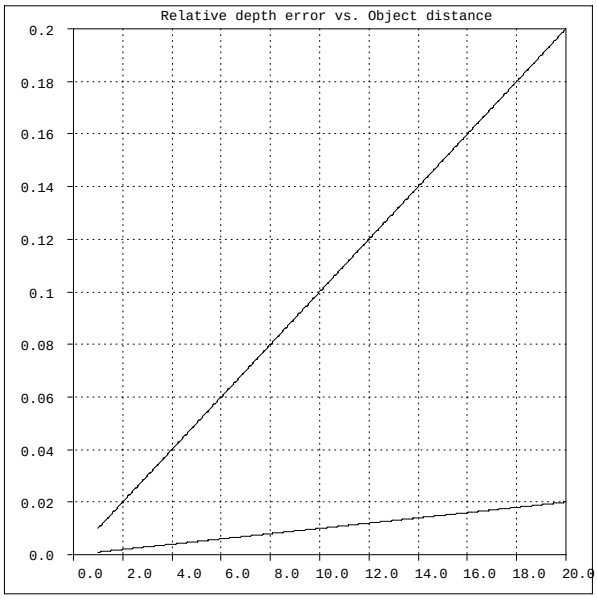


Figure 2: Vertical axis is the percentage error in computing depth, horizontal axis is the distance to the object (in units of interocular separation). Top graph is for errors in localizing image features of 10 pixels, bottom graph is for 1 pixel errors.

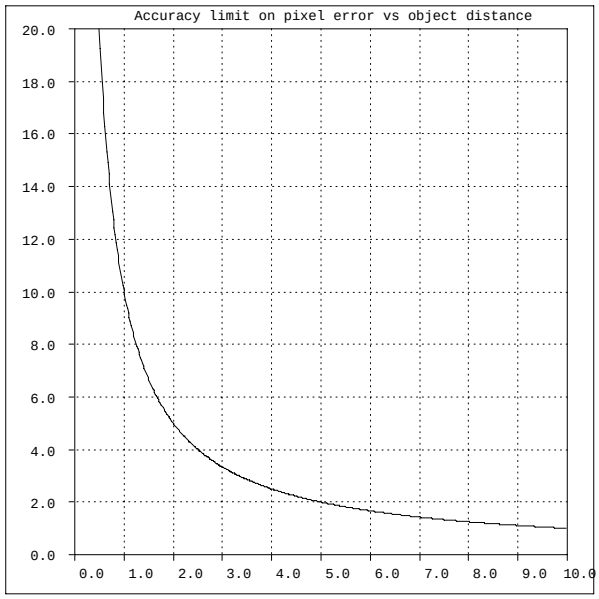


Figure 3: Vertical axis is the accuracy in pixel location needed so that the relative error in depth is less than 1%, horizontal axis is the distance to the object (in units of interocular separation).

1 meter away from our standard camera setup, $k = 10$ leads to 10% errors in computed depth. For the human system, these errors are reduced by a factor of 10. A second way of seeing this is to ask what is the accuracy on pixel location needed to keep the relative depth error less than 1%, as a function of the distance to the object. This is shown in Figure 3.

By equation 11, a 1 percent error in estimating f and disparities on the order of 10 pixels, will still only lead to 1 percent errors in relative depth for nearby objects ($Z/2b \approx 10$), which is small. Note that as the disparities get larger, the error increases. This has the interesting implication that if the object of interest is roughly fixated (i.e. the two optic axes intersect at or near the object) then disparities for features on the objects will be small, and the depth error will be small, while objects at larger disparities will have larger errors. Note that a similar observation has been made by Olson [31] who shows that much of the sensitivity of depth reconstruction to camera parameters can be isolated in the computation of the depth of the fixation point, while relative depth of other points with respect to this fixation can be computed fairly accurately.

All of this analysis is encouraging. Consider equation 12, however. Here, a 1 degree error in estimating the gaze angle will lead to 34 percent relative depth errors for nearby objects ($Z/2b \approx 10$), and even a .5 degree gaze angle error will lead to 17 percent relative depth errors. This is graphed in more detail in Figure 4. Similarly, in Figure 5, we plot the accuracy in gaze angle needed to keep the relative depth error at most 1%, as a function of distance to the object.

We note that errors due to gaze angle calibration could be a real problem. It is interesting to note that the human system appears able to measure gaze angle only up to an accuracy of roughly 1 degree [16] (page 67).

In short, we need to be certain that we have estimated the principal points accurately, and that we have very accurate measurements of the gaze angles of the cameras. If we cannot do so, then we will suffer distortion in our computed depth. More importantly, that distortion varies with actual depth, so the effect is non-linear. If we are trying to recognize an object whose extent in depth is small relative to the distance to its centroid, then the effect of this noise sensitivity is reduced. This is because the effect of the error will be systematic, and in the case of small relative depth, this uncertainty basically becomes a constant scale factor on the computed depth. On the other hand, however, if the object has noticeable relative extent in depth (even on the order of a few percent), then the uncertainty in computing depth will skew the results, causing difficulties for most recognition methods that compare computed 3D structure against stored models. Thus, the sensitivity may cause serious problems for recognition methods, both due to the large errors in depth and due to the distortions with varying depth.

4 Another Look at Stereo

Given that it may be difficult to reliably compute distance, and that distance may not be needed to handle

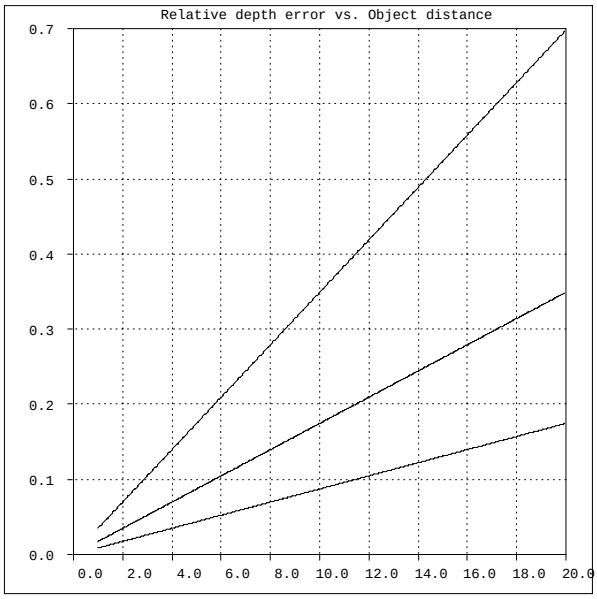


Figure 4: Vertical axis is the percentage error in computing depth, horizontal axis is the distance to the object (in units of interocular separation). Graphs are for errors in computing the gaze angle of 1, .5 and .25 degrees, from top to bottom.

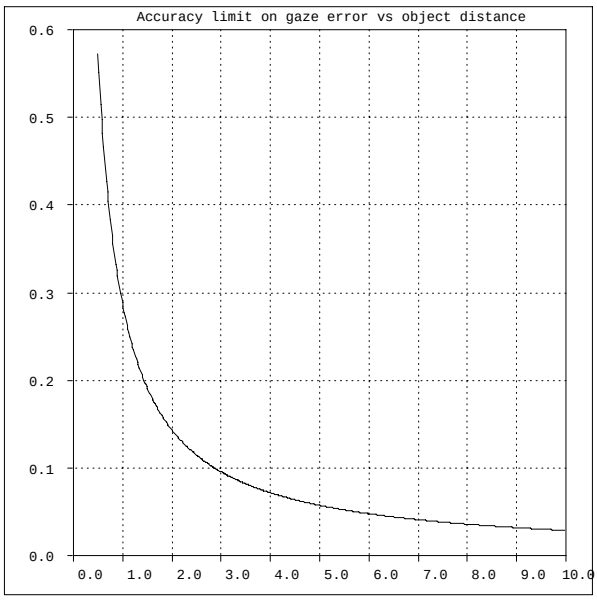


Figure 5: Vertical axis is the accuracy in gaze angle (in degrees) needed so that the relative error in depth is less than 1%, horizontal axis is the distance to the object (in units of interocular separation).

the two main uses of stereo output, we suggest that it is useful to reconsider the performance requirements that stereo should satisfy to support tasks such as object recognition. To handle figure/ground separation, a stereo algorithm should:

- be able to detect proximal (in the image) features that lie within some range of depth (i.e. find points that are near one another in 3D space, even if one does not know exactly where in 3D),
- be able to align matching distinctive features so that they are centered in the two images, to ensure that nearby parts of the corresponding object are visible in both images and can be matched,
- be able to integrate other visual cues about possible trigger features to foveate and fixate.

First, we should consider whether we can use existing stereo algorithms (e.g. [10], [4], [26], [36], [14]) to tackle the problem of figure/ground separation. We can conveniently separate stereo processing into several components:

- Choice of features to match: for our discussions, we will consider only edge based stereo matching.
- Constraints on the matching process.
- Control mechanism used to guide the matching process.

Most current stereo algorithms solve the correspondence problem as follows: Given any left image edge, search the set of right image edges for a unique match. The search is usually constrained by the (assumed known) epipolar geometry, and by a set of similarity constraints (e.g. edges should have similar orientation, similar contrast (or intensity variation), and so on). This holds both for matching individual edge points (in which case additional constraints such as figural continuity may also apply) and for extended edge fragments.

The key question is what constitutes a unique match, and this depends on the control mechanism used by the algorithm. For example, most of these algorithms attempt to find matches over a wide range of disparity, reflecting the fact that the viewed scenes may have objects ranging from close to the viewer (less than 1 meter) out to objects at the horizon. This can easily translate into disparity ranges on the order of several hundred pixels. The problem is that under these circumstances, it may be very difficult to guarantee uniqueness of match, especially when one is only considering local attributes of features, such as orientation and local contrast. One solution is to incorporate local geometric information about nearby edges [3], [29]. But an alternative is to consider changing the control mechanism.

The key problem is that previous stereo algorithms had as their goal the reconstruction of the scene, and hence they were designed to find as many correct matches as possible, over a wide range of disparities. On the other hand, if all we are interested in is separating out candidate image features that are likely to correspond to a single object, and we are willing to allow edge features to participate in several such groups,

then an alternative control method is viable. In particular, since we are interested in finding roughly contiguous 3D regions, it is attractive to envision a control method in which one fixates at some target, then searches for matching features within some range of disparity about that fixation point, collecting all such matching features as a candidate object, and continues.

Such an algorithm is similar in approach to some earlier stereo methods, notably [23, 27, 3], and it bears some similarity to evidence of the human stereo system, particular in the restriction of matching disparities only over a narrow range about the fixation point (referred to as Panum's limit in the perceptual literature) and the role of eye movements in guiding stereo [23, 27, 31]. It also clearly relates to work in active stereo head systems [1, 5, 6, 7, 9, 20, 30, 38, 33], especially work on using saliency of low level cues, or using motion information to drive stereo control loops that fixate candidate target areas [9, 6, 5, 30, 38, 33].

To demonstrate this idea, we have implemented the following stereo algorithm (influenced in part by earlier algorithms [3], [29]).

- Process both images to extract intensity edges. For convenience, process these edges to extract linear segments, using a standard split-and-merge algorithm. This latter step is mainly for reduction in computation and is not central to the demonstration.
- For each linear feature segment, record the position of the two endpoints, and the average intensity on each side of the feature. Also record the distance from each endpoint to other nearby features.
- Find a distinctive feature in one image that has a unique match in the other image, as measured over the full range of possible disparities. To begin with, we will measure distinctiveness as a combination of the length of the feature and the contrast of the feature. The idea is that such a feature can serve as a focal trigger feature. Of course many other cues could serve to focus attention [22].
- Rotate both cameras so that the distinct feature and its match are both centered in the cameras. This is a simple version of a fixation mechanism, in which the trigger feature is foveated and fixated in both cameras. Note that this will in general cause the optic axes to be non-parallel so that epipolar lines will no longer lie along horizontal rasters. A simpler version just uses a pan and tilt motion of the cameras to center the feature in one image, while leaving the optic axes parallel.
- Within a predefined range of disparity $\pm\delta$ (Panum's limit) about the zero disparity position (due to fixation), search for other features that have a unique match. Note that uniqueness here means only within this range of disparity. There may be other edges outside of this disparity range that satisfy the matching constraints, but in this case such matches are ignored. In our implementation, two edges match if their lengths are roughly the same, if a significant fraction of each edge has an epipolar overlap with the other edge, if the orientation is roughly the same, if the average intensity on at least one side of the edge is roughly the same, and if the arrangement of neighbouring edges at one of the endpoints is roughly the same.
- This set of edges now consistutes an hypothesized fragment of a single object. We can save these edges, and continue the process, looking for another unique trigger feature to align the cameras. Alternatively, we can pass these edge features on to a recognition algorithm, such as Alignment [17, 18].

We have implemented an initial version of this algorithm, and used it in conjunction with an eye-head system, which can pan and tilt as a unit, as well as change the optic axes of one or both cameras. An example of this algorithm in operation is shown in Figures 6–11. Given the images in Figure 6, we extract edges (Figure 7). From this set of edges, the most distinctive edge (measured as a combination of length and intensity contrast) with a unique match is isolated in Figure 8. This enables the cameras to fixate the edge and obtain a new set of images (Figure 9) and edges (Figure 10). Relative to this fixation, stereo matching is performed over a narrow range of disparity, isolating a set of edges likely to come from a single object (Figure 11). Notice how the tripod is extracted from the originally cluttered image, with minimal additional features.

5 Conclusions

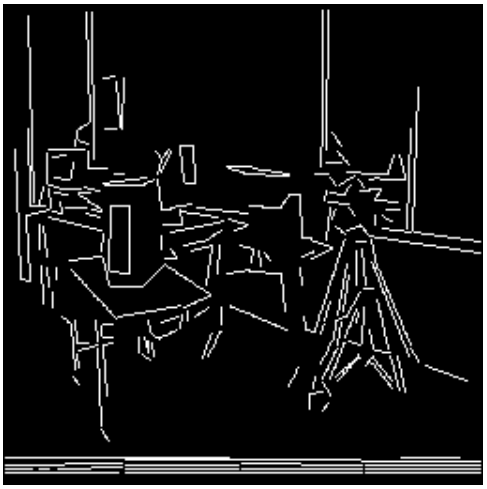
We have suggested that stereo may play a central role in object recognition, but not in the manner usually assumed in the literature. We have suggested that stereo may be most useful in supporting figure/ground separation, and that to do so it need not compute explicit 3D information. Supporting this argument were the observation that depth reconstruction is extremely sensitive to accuracy in the measured camera parameters, and the observation that the human stereo system may not compute explicit depth.

Using the idea of depth detectors tuned to a narrow range about a fixation point has been previously explored in the literature, primarily for obstacle avoidance [15], [32]. This work considers the same general idea within the context of recognition. Such an approach opens up several other avenues for investigation. For example, what is the role of other visual cues in aiding the stereo matching problem. While one option is to augment image features with attributes, such as texture or color measures, an alternative is to consider using such cues to drive vergence eye movements, helping to align the cameras on trigger features, so that the local matcher can extract image features likely to correspond to a single object. We intend to explore these and related issues in the near future.

References

[1] A.L. Abbott and N. Ahuja, "Surface reconstrucxtion by dynamic integration of focus, camera vergence, and

Figure 6: Initial test images.



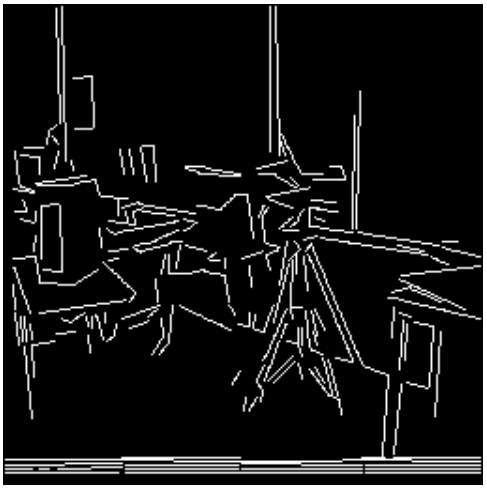


Figure 7: Initial test edges.

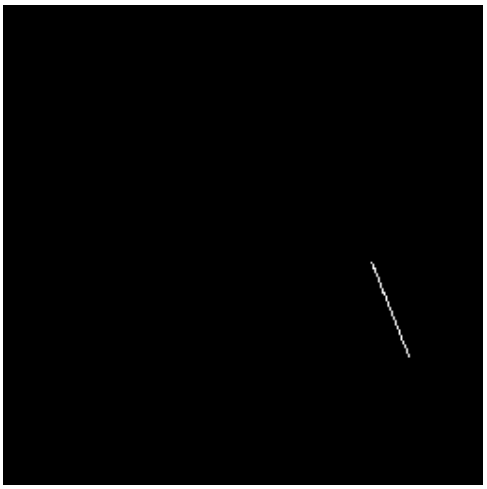
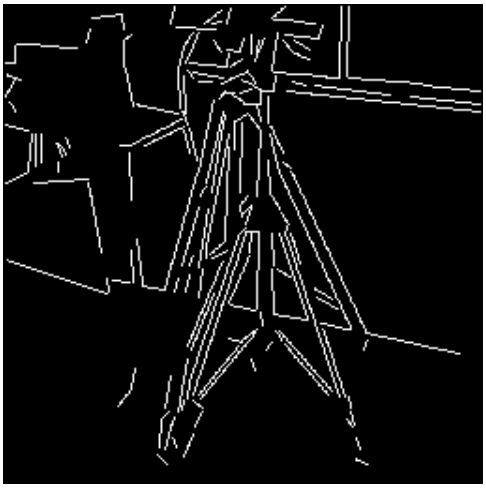


Figure 8: Initial focal edge.

Figure 9: Fixated test images. In this case, we have foveated the left image edge by a pan/tilt motion of the head.



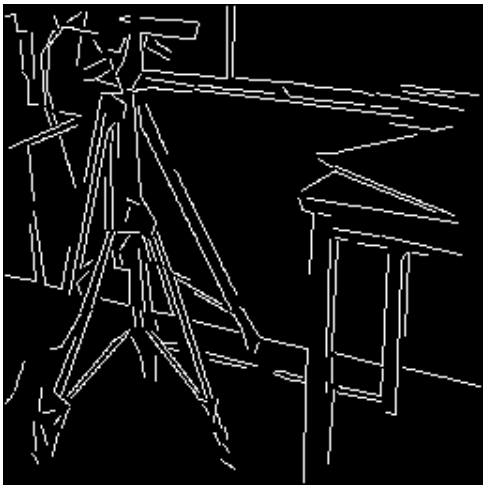


Figure 10: Fixated test edges. In this case, we have foveated the left image edge by a pan/tilt motion of the head.

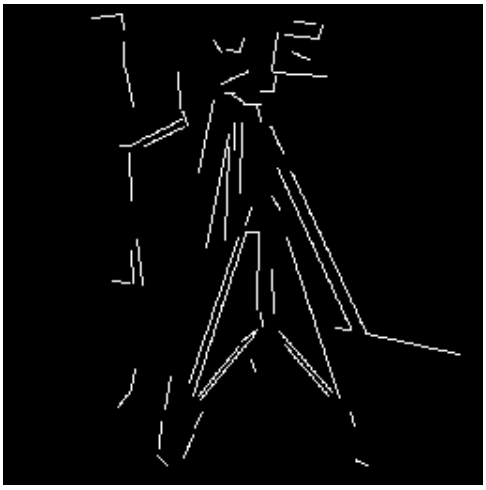


Figure 11: Matched fixated edges.

stereo", *Proc. Second Int. Conf. Comp. Vision*, pp. 532–543, 1988.

[2] S.M. Anstis, I.P. Howard, and B.J. Rogers, "A Craik-O'Brien-Cornsweet Illusion for Visual Depth", *Vision Research*, **18**:213–217, 1978.

[3] N. Ayache and B. Faverjon, "Efficient registration of stereo images by matching graph descriptions of edge segments", *International Journal of Computer Vision* **1**(2):107–131, 1987.

[4] S.T. Barnard, and M. Fischler, "Computational stereo", *ACM Computing Surveys*, **14**(4):553–572, 1982.

[5] C. Brown, D. Coombs and J. Soong, "Real-time Smooth Pursuit Tracking", in *Active Vision*, A. Blake and A. Yuille (eds), MIT Press, pp. 123–136, 1992.

[6] J. Clark and N. Ferrier, "Attentive Visual Servoing", in *Active Vision*, A. Blake and A. Yuille (eds), MIT Press, pp. 137–154, 1992.

[7] D.J. Coombs, *Real-time gaze holding in binocular robot vision*, Ph.D. Thesis, University of Rochester, 1991.

[8] O. D. Faugeras, "What can be seen in three dimensions with an uncalibrated stereo rig?", *Second European Conference on Computer Vision*, Santa Margherita Ligure, Italy, pp. 563–578, 1992.

[9] N. Ferrier, *Trajectory control of active vision systems*, Ph.D. Thesis, Division of Applied Sciences, Harvard University, 1992.

[10] W.E.L. Grimson, "Computational experiments with a feature based stereo algorithm", *IEEE Pattern Analysis and Machine Intelligence*, **7**, No. 1, 17–34, 1985.

[11] W. E. L. Grimson, *Object Recognition by Computer: The Role of Geometric Constraints*, MIT Press, 1991.

[12] W. E. L. Grimson and T. Lozano-Pérez, "Model-Based Recognition and Localization from Sparse Range or Tactile Data", *International Journal of Robotics Research*, **3**, No. 3, 3 – 35, 1984.

[13] W. E. L. Grimson and T. Lozano-Pérez, "Localizing Overlapping Parts by Searching the Interpretation Tree", *IEEE Pattern Analysis and Machine Intelligence*, **9**, No. 4, 469–482, 1987.

[14] W. Hoff and N. Ahuja, "Surfaces from stereo: Integrating feature matching, disparity estimation, and contour detection", *IEEE Trans. Patt. Anal. & Mach. Intell.* **11**(2): 121–136, 1989.

[15] I.D. Horswill, "Proximity Detection using a filter tuned in three-space", *DARPA Image Understanding Workshop*, pp. 973–978, 1992.

[16] I.P. Howard and W.B. Templeton, *Human Spatial Orientation*, John Wiley & Sons, 1966.

[17] D.P. Huttenlocher and S. Ullman, 1987, "Object Recognition Using Alignment", *Proc. First Int. Conf. Comp. Vision*, pp. 102-111.

[18] D.P. Huttenlocher and S. Ullman, 1990, "Recognizing Solid Objects by Alignment with an Image," *Inter. Journ. Comp. Vision* **5**(2):195–212.

[19] Commercial brochure, ISTAR, Sophia Antipolis, France.

[20] E. Krotkov, *Active computer vision by cooperative focus and stereo*, Springer-Verlag, 1989.

[21] R.K. Lenz and R.Y. Tsai, "Techniques for Calibration of the Scale Factor and Image Center for High Accuracy 3-D Machine Vision Metrology", *IEEE Pattern Analysis and Machine Intelligence*, **10**, No. 5, 713–720, 1988.

[22] S.T.F. Mahmood, "Attentional Selection in Object Recognition", MIT AI Technical Report 1420, 1992.

[23] D. Marr, *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*, W. H. Freeman and Company, San Francisco, 1982.

[24] D. Marr and T. Poggio, "A theory of human stereo vision", *Proc. Royal Society of London B* **204**:301–328.

[25] L. Matthies and S.A. Shafer, "Error modeling in stereo navigation", *International Journal of Robotics and Automation* **3**(3):239–248, 1987.

[26] J.E.W. Mayhew and J.P. Frisby, "Psychophysical and computational studies towards a theory of human stereopsis", *Artificial Intelligence* **17**(1–3): 379–386, 1981.

[27] J.E.W. Mayhew and J.P. Frisby, *3D Model Recognition from Stereoscopic Cues*, MIT Press, 1991.

[28] E.S. McVey and J.W. Lee, "Some accuracy and resolution aspects of computer vision distance measurements", *IEEE Pattern Analysis and Machine Intelligence* **4**(6):646–649, 1982.

[29] G. Medioni and R. Nevatia, "Segment-based stereo matching", *Computer Vision, Graphics, and Image Processing*, **31**: 2–18, 1985.

[30] D.W. Murray, F. Du, P.F. McLauchlan, I.D. Reid, P.M. Sharkey, and J.M. Brady, "Design of stereo heads" in *Active Vision*, A. Blake and A. Yuille (eds), MIT Press, pp. 155–172, 1992.

[31] T.J. Olson, "Stereopsis for Verging Systems", *IEEE Comp. Vis. Patt. Recog. Conf.*, 55–60, 1993.

[32] T.J. Olson and D.J. Coombs, "Real-time vergence control for binocular robots", *DARPA Image Understanding Workshop*, pp. 881-888, 1990.

[33] K. Pahlavan, T. Uhlin, and J.-O. Eklundh, "Dynamic Fixation"m *Proc. Fourth Int. Conf. Comp. Vision*, pp. 412–419, 1993.

[34] D.I. Perrett, P.A.J. Smith, D.D. Potter, A.J. Mistlin, A.S. Head, A.D. Milner and M.A. Jeeves, "Neurones responsive to faces in the temporal cortex: studies of functional organization, sensitivity to identity and relation to perception", *Human Neurobiology* **3**:197–208, 1984.

[35] D.I. Perrett, P.A.J. Smith, D.D. Potter, A.J. Mistlin, A.D. Milner and M.A. Jeeves, "Visual cells in the temporal cortex sensitive to face view and gaze direction", *Proc. Royal Society of London B* **223**: 293–317, 1985.

[36] S.B. Pollard, J.E.W. Mayhew, and J.P. Frisby, "PMF: A stereo correspondence algorithm using a disparity gradient limit", *Perception* **14**, 449–470, 1985.

[37] B.J. Rogers and M.E. Graham, "Anisotropies in the perception of three-dimensional surfaces", *Science* **221**, 1409–1411, 1983.

[38] P.M. Sharkey, I.D. Reid, P.F. McLauchlan, and D.W. Murray, "REal-time control of an active stereo head/eye platform", *Proc. 2nd Intern. Conf. on Automation, Robotics and Vision*, Singapore, 1992.

[39] A. Shashua, "Projective Structure from two Uncalibrated Images: Structure from Motion and Recognition", MIT AI Lab Memo 1363, 1992.

[40] K.A. Stevens and A. Brookes, "Depth reconstruction in stereopsis", *First Int. Conf. Comp. Vis.*, 682–686, 1987.

[41] R.Y. Tsai, "A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses", *IEEE Journal of Robotics and Automation* **3**(4):323–344, 1987.

[42] S. Ullman and R. Basri, "Recognition by Linear Combinations of Models", *IEEE Pattern Analysis and Machine Intelligence*, **13**(10):992–1006, 1991.

[43] A. Verri and V. Torre, "Absolute depth estimate in stereopsis", *Journal of the Optical Society of America A* **3**(3):297–299, 1986.

[44] R.P. Wildes, "Direct recovery of three-dimensional scene geometry from binocular stereo disparity", *IEEE Trans. Pattern Analysis and Machine Intelligence*, **13**(8): 721-735, 1991.

[45] L.B. Wolff, "Accurate Measurement of Orientation from Stereo using Line Correspondence", *IEEE Conf. on Computer Vision and Pattern Recognition*, 410–415, 1989.