Perceptual Surface Reconstruction

Jens Månsson

Lund University Cognitive Studies 129
2005

Printed in Lund, Sweden, by KFS i Lund AB
ISBN 91-974741-5-0
ISSN 1101-8453
ISRN LUHFDA/HFKO-1016-SE
Lund University Cognitive Studies 129

To my parents

Contents

Introduction

This thesis address various aspects of binocular depth perception, and attempts to account for how the information, contained in the left and right retinal images, is processed and transformed into a useful 3-D surface description.

The fundamental principle behind stereoscopic depth perceptions is fairly simple, and rests on the fundamental fact that the lengths of the sides of a triangle can be computed, given that at least one of the sides, and two of the angles, are known. Hence, given a distant object on which the eyes converge, the convergence angle, and the (known) distance between the eyes, can be used to compute an estimate of the distance to the object. Moreover, when an object is positioned off the horopter (i.e. plane of convergence) the projection of the object will not end up on the exact same relative position in the left and right retinas, due to the difference in perspective of the eyes, but will be more or less horizontally displaced depending on the objects distance from the horopter. This relative displacement, or *binocular disparity*, between corresponding retinal projections, together with the convergence angle of the eyes, form the basis for stereoscopic depth perception (*stereopsis)*.

While the fundamental principle behind stereopsis may be simple, there are numerous difficulties that the visual system must overcome in order to effectively, and accurately, process this information. One of the most central questions, in this respect, is how the visual system manages to identify corresponding retinal projections in the two eyes. Before the random-dot stereogram was introduced as a research tool (Julesz, 1960) this process was not well understood, and it was widely believed that objects had to be identified/recognized independently in each view before they could be binocularly matched. The fact that random-dot stereograms can be binocularly fused, despite the lack of any monocularly identifiable shapes, or cues to depth, clearly showed that stereopsis is a relatively early process, which operate on low-level stimuli. Now, most (biologically oriented) computational theories of stereopsis assume that binocular correspondence, primarily, is established between simple edge, and bar, segments, which resemble the stimuli that disparity sensitive neurones in

area V1(Hubel & Wiesel, 1962) typically respond strongly to. A difficulty with resolving the correspondence problem at such a low level, is that basically any stimuli can be broken down into edge, and line, segments, which means that the ambiguity, and likelihood of false matches, increase. To overcome this difficulty, various computational constraints have been proposed that are intended to reduce the solution space in various ways. In article I, and article III, in this thesis, the computational basis and the justification for different such constraints are discussed, and two different computational models are proposed.

A common assumption in virtually all stereo models is that surfaces, in general, are opaque and relatively smooth; i.e. that nearby points on a surface have a similar depth/disparity. This assumption have often been used to justify computational constraints that mutually reinforce potential matches that lie close to each other in the image plane as well as in depth. Obviously, this type of constraint is not justified between image primitives that are parts of different surfaces, but should only be applied within the enclosing boundaries of individual surfaces. In article II, the difficulties related with the monocular identification of occluding contours are discussed; and a computational model is proposed that, based on a few simple heuristics, enhance image discontinuities whether they are caused by a change in luminance, texture, or by line-endings, and that respond to "illusory" contours (see for example Kanizsa, 1979).

How occluding edges affect the interpretation of binocular depth is further explored in article IV. In this paper, an empirical study is described, which investigate how different types of (occluding) boundary inducers affect disparity interpolation in ambiguous image regions, and the perceived completion of sparsely defined surface.

*

In article I (*Stereovision: A model of human stereopsis*), a computational model of stereopsis is presented. As foundation for the model lies a number of ideas that has arisen from redefining the correspondence problem. Instead of establishing potential matches by the detection and matching of some set of "predefined" features, e.g. edges (zero-crossings) or bars (peaks/trough), matches are sought by comparing the overall configura-

tion of contrast within delimited regions of the two images. The main disambiguating power of the model is provided by combining the results of the matchings from a number of independent channels of different coarseness (with respect to the resolution of the contrast information). The idea is that the information in the coarser channels can be used to restrict the domain of potential matches, to be considered, within the finer channels. Important for this assumption is the concept of figural continuity. To further reduce the set of potential matches, the model relies on the constraint of uniqueness. A computer implementation of the model is presented, which from the input consisting of a stereogram, produces a representation of the binocular disparity present within the stereogram. A number of results obtained from this computer implementation are also presented and discussed.

*

In article II (*Occluding Contours: A computational model of suppressive mechanisms in human contour perception*), the fundamental problem is addressed, of how to identify the occluding contours of objects, given the ambiguity inherent in low-level visual input. A computational model is proposed for how occluding contours could be identified by making use of simple heuristics that reduce the ambiguity of individual features. In the striate cortex, a large majority of cells are selective for both contrast and orientation; i.e., they respond preferentially to simple features like contrast edges or lines. The proposed heuristics enhance or suppress the outputs of model striate-cortical cells, depending on the orientation and spatial distribution of stimuli present outside of the "classical" receptive field of these cells. In particular, the output of a cell is suppressed if the cell responds to a feature embedded in a texture, in which the "component features" are oriented in accordance with the orientation-selectivity of the cell. The model has been implemented and tested on natural as well as artificial grey-scale images. The model produces results that in several aspects are consistent with human contour perception. For example, it reproduces a number of known visual phenomena such as illusory contours, contour masking, pre-attentive pop-out (due to orientation-con-

trast), and it enhances contours that human observers often report perceiving as more salient.

*

In article III (*The Uniqueness Constraint Revisited: A symmetric Near-Far inhibitory mechanism producing ordered binocular matches*), psychophysical studies are reviewed that suggest that image features, under certain conditions, can give rise to multiple visible binocular matches. These findings are difficult to reconcile with the traditional interpretation of the uniqueness constraint. A new interpretation, a conditional uniqueness constraint, is proposed that allows multiple matching of similar primitives when a one-to-one correspondence does not exist, locally, within corresponding image regions, but prohibits it when a one-to-one correspondence does exist. A cooperative network model and an implementation are also described, where this constraint is enforced at each network node by a simple inhibitory (dual) AND-gate mechanism. The model performs with high accuracy for a wide range of stimuli, including multiple transparent surfaces, and seems able to account for several aspects of human binocular matching that previous models have not been able to account for.

*

Finally, article IV (*The Perception of Binocular Depth in Ambiguous Image Regions: Toward a computational theory of surface perception*) describes an empirical investigation of the perception of binocular depth in image regions that lack explicit disparity information. For this purpose, sparse random-dot stereograms were used. The basic stimuli-design consisted of two different depth planes; a foreground that covered the entire scene, and a background that covered only half the scene; from the center to either the left, or the right, end of the display. Despite the fact that the foreground covered the entire scene, subjects typically reported that the ambiguous image regions, in between the foreground dots, belonged to the background. When, however, a few unpaired dots were added along the center, to suggest an occluding opaque surface, subjects tended to perceive the same ambiguous region as belonging to the foreground, which suggest interaction between the binocularly paired, and unpaired, stimuli. A

number of variations, of this basic theme was investigated, including: changing the density of the dots in the two depth planes; changing the number, and positioning, of the unpaired dots along the center; making the background extend over the entire scene except for a central rectangular region; using other cues, e.g. a 2-D contour, or a pair of "Kaniza" inducers, to suggest occlusion. Our results can not be accounted for by any simple disparity interpolation scheme, but seem to require additional processing within the disparity domain, as well as interaction with processes devoted to the identification of occluding boundaries.

References

Julesz, B., 1960, Binocular depth perception of computer generated patterns. Bell System Technical Journal, 39, 1125-62

Hubel, D.H. & Wiesel, T.N., 1962, Receptive fields, binocular interaction and functional architecture in the cats visual cortex. Journal of Physiology, 160, pp 106-154

Kanizsa, G., 1979, Organization in vision: Essays on Gestalt perception. Praeger Publishers, New York.

PAPER I

Stereovision: A model of human stereopsis

Abstract – A model of the human stereopsis mechanism is presented. As foundation for the model lies a number of ideas that has arisen from redefining the *correspondence problem*. Instead of establishing potential matches by the detection and matching of some set of "predefined" features, e.g. edges (zero-crossings) or bars (peaks/trough), matches are sought by comparing the overall configuration of contrast within delimited regions of the two images. The main disambiguating power of the model is provided by combining the results of the matchings from a number of independent *channels* of different coarseness (in regard to the resolution of the contrast information). The idea is that the information in the coarser channels can be used to restrict the domain of potential matches, to be considered, within the finer channels. Important for this assumption is the concept of *figural continuity*. To further reduce the set of potential matches, the model relies on the constraint of *uniqueness*. A computer implementation of the model is presented, which from the input consisting of a stereogram, produces a representation of the binocular disparity present within the stereogram. A number of results obtained from this computer implementation are also presented and discussed.

1 Introduction

One of the major functions of the human brain is to construct a representation of the world surrounding us. For a human being, and many other animals, the perhaps most important sense for accomplishing this task is the visual sense. Without it we would be severely handicapped because it alone allows us to perceive and represent a great number of aspects of our environment. One such aspect that is of fundamental importance is that of spatial relationships. Since space is three-dimensional we have to perceive all three dimensions in order to acquire a full representation of these relationships. The problem is that the images that reaches our eyes, considered individually, only reveals the two-dimensional spatial relationships. However, taken together they contain sufficient information to allow the third dimension to be recovered. Thus, in order for the brain to reconstruct the 3-D structure of the environment, the information in the

two separate images must somehow be combined. How then is this transformation from 2-D images to a 3-D representation of the world achieved? The recovery of the third dimension is really not the result of one process, but of several more or less independent ones. The conscious awareness of depth that we perceive is therefore a product of the whole mind and can not be ascribed to one particular system. However, as we shall see there is one outstanding mechanism in the brain, referred to as *stereopsis*, that is of crucial importance for our ability to perceive depth.

Before going into the details of the stereopsis mechanism, I would first like to present some other *cues* to depth that are thought to be used by the brain.

Two *physiological cues* that are important for depth perception are the *convergence* of the eyes and the *accommodation* of the lenses. The degree to which our eyes converge depends on where we fixate our eyes. If we fix them on something near they converge more than they do if we look at something far away. The accommodation of the lens, in turn, is determined by where we focus. When focusing on something far away, the muscles around the lens are relaxed and the lens is therefore relatively thin, but in order to bring a closer scene into focus the lens has to change shape. The muscles around the lens therefore contracts to form the lens appropriately. These different types of information, about the degree of muscle contractions, are not by themselves useful to the brain, but in combination with the visual input they are essential for the ability to perceive depth.

There are several *monocular cues* to depth as well. If you have only one eye open and move your head from side to side, you will experience a sensation of depth. This phenomenon is called *motion parallax*. The shading of an object or a scene can also provide an impression of depth. Usually, we are not even aware of the existence of such cues, but there are other cues that only makes sense in combination with higher knowledge or learned relationships. For example, if one surface/object partially covers another one, it is possible to determine that the covered surface/object is furthest away. This might seem very obvious but in fact requires that, at least, a partial identification of the two objects/surfaces has taken place, so that their spatial extensions can be established. Another such cue has to do with the size of objects. If the size of an object is priorly known, it will appear far away if it produces a small image on the retina, and vice versa if it produces a large image. These are just a few examples of monocular cues, and there are several others (e.g. perspective, texture gradients, e.t.c.). As mentioned above, the extent to which higher knowledge is involved in making use of these cues varies, and sometimes it might be more appropriate to say that we are dealing with pure reasoning rather than cues.

However this might be, the by far richest source of depth-information comes from combining the information from the two eyes. Due to the fact that our eyes are horizontally separated, the image that falls on one eye will differ slightly in

perspective from that of the other. This means that the different features, making up the images, will not fall on the exact same locations in the two retinas (Fig.1). The magnitude of this horizontal displacement, or *binocular disparity*, is decided by two factors: the convergence of the eyes and the distance to the surfaces, giving raise to the features on the retinas. Now, signals about the convergence of the eyes are directly transmitted to the brain, and the binocular disparity can indirectly be measured from the combined information in the retinal images. Thus, all the necessary information is available for the brain to compute the depth of the scene. The ability, of the brain, to perform these computations is referred to as *stereopsis*.

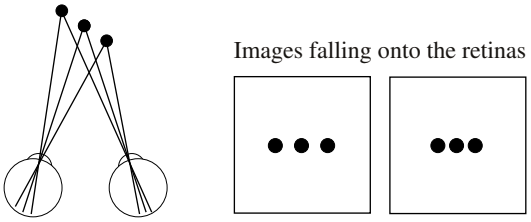

Figure 1. Due to the difference in perspective, the images of the dots will fall onto slightly different locations in the two retinas.

The first to appreciate the role binocular disparity has in seeing depth was Wheatstone, whom in 1838 invented the first stereoscope. The stereoscope became a quite popular gadget in those days, but any deeper analysis of the phenomenon was hindered due to lack of appropriate tools to investigate it with, and due to an immature general knowledge of how the brain functions. The prevalent view of stereopsis was that it depended heavily on monocular form recognition. It was thought that the image from each eye was separately analysed, and all the components of the images was identified and recognised before they could be binocularly combined. This belief placed the phenomenon of stereopsis at a relatively high level, in the cognitive chain, since it – according to these conclusions – had to occur after object recognition.

It was not until a century later that it would be proven otherwise, when Bela Julesz (1960) developed the *random-dot stereogram*. A random-dot stereogram (Fig.2) contains no information of monocular form. When viewed separately, all one can see are black dots spread out over a white surface. Only when the images are fused in a stereoscope, or by crossing ones eyes, is it possible to perceive the shape and depth of the scene. The only information available to the brain is the binocular disparity that separates the dots in one image from the corresponding dots in the other image. This clearly shows that binocular disparity

Figure 2. A random-dot stereogram contains no monocular depth-cues. The 3-D structure hidden in the stereogram can only be perceived when the images are binocularly fused in a stereoscope or by crossing the eyes.

alone is sufficient to perceive depth, and that stereopsis therefore does not have to occur after object recognition. In fact, it is now known that stereopsis occurs at an early level in the visual pathway. An important neurophysiological finding showing this was made by Barlowe, Blakemore and Pettigrew (1967) who discovered neurons in area V1 that are selective for horizontal disparity between the input from the two eyes.

The problem of stereopsis then basically boils down to the matching of corresponding features in the two images that are projected into the eyes. This is often referred to as the *correspondence problem*. Conceptually, it can be clarifying to consider the matching process as being divided into, using Julesz terminology, a "local" and a "global" matching process. In the local matching process, possible candidates to which a feature may match are sought. If each feature could be uniquely described there would be only one possible match in the opposite image, and thus would there be no correspondence problem. Naturally, this demand for uniqueness is not very realistic (I will return to the reasons for this in the following section). In fact, the result of the local matching is often highly ambiguous. The mechanism that resolves this ambiguity, and sorts the correct matches from the "ghosts", is in this framework referred to as the global matching process.

I will in this paper present a model of human stereopsis, which in a number of aspects simulates the behaviour of the human stereopsis mechanism. In the following sections, I will first discuss what primitives could be used as input to such a mechanism? I will then go on to discuss how different constraints could be imposed on the matching process in order to dissolve ambiguous matches. Fi-

nally, will I present the model and the outlines of a computer implementation that from the input, consisting of a stereogram, reconstructs the 3-D structure of the scene.

Anyone trying to model human stereopsis, or any other information-processing system, has to face a number of decisions about what is to be calculated, what information and representation is to be used, what transformations should be performed and why they should be performed. Marr (1982) has thoroughly analysed which questions, like those above, are relevant to ask for such a task, and also what has to be known about any information-processing system before it could be said to be fully understood. His main idea is that any information-processing system can be explained at different levels of abstraction, and he emphasises the importance of understanding each of these levels separately, before the whole system can be understood. Marr has chosen to divide this analysis into three different levels: the level of computational theory, the algorithm- and representational-level, and the level of implementation. At the first level, one has to make clear what the goal of the computation is and how this goal can be accomplished? What strategy is to be used and what makes it justified? Applied to the analysis of stereopsis, an important part of this involves finding constraints, imposed by the physical world, that can be used to justify the global matching processes. At the second level, the type of information and representation has to be considered. What is the input and output, and what algorithm could perform this transformation? The final level is concerned with the details of the physical implementation of the algorithm.

One can only agree that this is a most reasonable approach and it has therefore been somewhat of a guideline to my thoughts during my attempt to model human stereopsis. I have also had as an aim, with this paper, to cover most of these different aspects of the stereopsis problem.

2 Matching primitives

From a philosophical or computational point of view, one could say that there is a trade-off to be made between the representational capacity, and the amount of processing, needed to solve the correspondence problem that depends on the complexity of the features used in the matching process.

On the one extreme, using low-level features (e.g. like the intensity value in each point of the image) would require little representational capacity, but also make it quite impossible to establish the correct set of matches simply by comparing features, since such a procedure – in the general case – would cause a large amount of ambiguous matches. An extensive amount of (global) processing would therefore be needed to sort the correct matches from the “ghosts” – if at all possible.

On the other extreme, if one could divide the image into a number of more complex features (e.g. objects or sub-regions containing a particular texture e.t.c.) that allowed each feature to be "uniquely" described, practically no matching-process would be necessary since the "uniqueness" would assure a one-to-one correspondence between features. This strategy would however put high demands on the representational capacity, since it would have to be able to represent, very accurately, an enormous number of different features in order to allow for discrimination among these. In fact, the later of these strategies is not plausible, in its extreme form, even if we had an infinite representational capacity. The reason for this is that the demand for uniqueness is not realistic in the general case. The answer in turn to why uniqueness is not realistic depends somewhat on how one chooses to interpret the complexity of a feature and is not straightforward to answer completely, but I will give two simple examples that gives a general idea. The first is simply that two, or several, features that give raise to the exact same projection on the retina, obviously will have to be represented exactly the same way too. Thus, will they be impossible to discriminate from each other by comparison alone, no matter how elaborate and exhaustive the representation of them are. Second, since the disparity we are seeking has the effect of producing different images in our eyes, the corresponding features will often appear slightly differently, and this makes the one-to-one correspondence based on uniqueness impossible.

As seen above, both strategies have their benefits and shortcomings concerning the need for representational capacity and processing power. Neither of them, in their extreme form, seems likely to be used by the human brain. Instead, what one should look for is some kind of compromise in which the best properties could be combined. I will at the end of this section suggest a way in which this might be accomplished.

Philosophical or computational considerations alone will not tell us what matching primitives are used by the brain, but they can guide the search in the right direction. In order to tie these ideas to reality, one has to know something about the neuronal machinery and the information it feeds on. In the light of discussing this next I will present some of the various matching primitives that have been suggested to be used by the brain, and I will also present some evidence in favour and against these.

It was early proposed that a point-by-point matching of brightness values could be conducted, but for various reasons this idea has now little support. In most types of images the intensity changes smoothly over surfaces and is often constant within relatively large regions. The probability of establishing a one-to-one correspondence between all points in the images, simply by comparing brightness values, would therefore seem to be low due to the large number of potential matches. It would also be difficult to defend such a strategy in the light of findings made by Julesz (1971), who showed that images with different degree

of contrast could easily be fused. Another important reason why this seems unlikely is that the information of the absolute light intensity, measured by the receptors in the retina, is not directly transmitted to the cortex where fusion occurs. The information leaving the eye, the output of the retinal ganglion cells, in fact represents something quite different from the raw light intensity values reaching the retina.

There are two major kinds of retinal ganglion cells: *on-centre* and *off-centre* cells (Fig. 3). The on-centre cells responds most strongly when light hits the central part of their receptive field. If diffuse light covers both the excitatory centre and the inhibitory periphery the response is weakened, and if only the peripheral parts are exposed the response will be suppressed. The off-centre cells have a reversed response pattern since their central parts are inhibited by light and the surround is excited. There are many different sizes of these receptive fields and they could roughly be said to grow with the distance from the fovea, but there are large ones in the central parts as well. Also important is that neighbouring cells' receptive fields overlap almost completely, so that they together cover the whole visual field (Hubel 1988). Considering the compositions of these receptive fields, it is clear that these cells does not respond to the absolute amount of light hitting the retina, but rather to the difference between the light falling on the central and the surrounding parts of their receptive fields. In other words, the output of the eye basically contains information about the relationship of contrast within the retinal image.

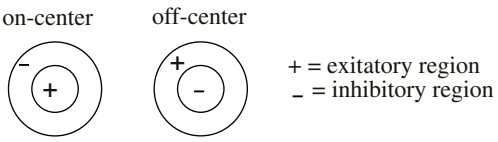

Figure 3. Receptive field mapping of the retinal ganglion cells.

Still this information is not directly used by the stereopsis mechanism, but as we shall see it is used by other cortical cells which output, in turn, is used as input to the stereopsis mechanism. Before discussing stereopsis in more detail, I will therefore first describe some of these "other" cells and explain to what type of stimuli they react.

Hubel and Wiesel were the first to make successful recordings from cells in the cortex of cats (Hubel & Wiesel 1959) and later monkeys. They found a number of cells, which they divided into two major groups, *simple* and *complex* cells, depending on their response to different types of stimuli. Simple cells all have in common that they respond most strongly when a particular configuration of light fall within their receptive field. A typical simple cell gives a strong response if a

rectangularly shaped area of light, with a particular orientation, falls within its receptive field (Fig. 4a). If the light falls too much outside of the central part of the receptive field, the response will be low or suppressed. There are many variations of simple cells and some respond best to a border, between light and darkness, of a certain orientation (Fig. 4b). The sizes and distribution of the simple cells' receptive fields coincide fairly well with those of the retinal ganglion cells'.

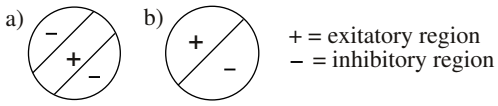

Figure 4. Receptive fields of two typical simple cells.

Complex cells have slightly larger receptive fields than simple cells. These cells also give a strong response for border- and "bar"-shaped stimuli of a certain orientation, but there are other factors determining their response as well. Some complex cells respond equally well to a particular stimulus, with the right orientation, no matter where it falls within its receptive field. Others only respond if the stimulus, except from being of a certain kind and orientation, moves across the receptive field as well.

A special group of complex cells, called *hypercomplex* or *end-stopped* cells, have receptive fields similar to the complex cells' described above, but for one exception. For instance, if the stimulus is a bar-shaped light with the right orientation, the cell will respond equally strong no matter where the light falls within the receptive field, as long as the bar does not extend over a certain border. If it does the response will be weakened or suppressed (Hubel 1988).

The simple and complex cells above were all described as taking their input from only one eye, but both simple and complex cells with binocular receptive fields have been found as well. Even more important considering stereopsis is that cells have been discovered in area V1 that respond optimally to stimuli with a certain horizontal disparity between the eyes (Barlowe, Blakemore & Pettigrew 1967). Studies of cells in macaque monkeys, an animal which has a capacity to perceive depth very similar to that of humans, found that as many as 60–70% of the cells in striate cortex, and an even larger number in prestriate cortex, were sensitive to horizontal disparity, and that many of these showed properties like those of simple and complex cells (Poggio & Poggio 1984). As we can see the necessary input for the stereopsis mechanism seems to be available, and the interesting question therefore becomes how this information is used? Are the simple and complex cells actual "feature-detectors" or is the information they provide used to produce some more elaborate description?

a) b) c) d)

Figure 5. (a) Showing an image (128x128 pixels) and the results after having convoluted the image with the ∇^2G-operator. The space constant σ has the values of 1, 2 and 4 pixels in (b), (c) and (d) respectively.

Marr and Hildreth (1980) have argued that an important result of early vision is the construction of a "raw primal sketch". In short this is a symbolic description of the different primitives making up the image (e.g. edges, bars, and blobs) that contains information about their size, orientation and position within the image. In order to discover such primitives in an image a first step is to detect changes in the light intensity values. A number of different derivatives, or "filters", could be used for this purpose. Marr and Hildreth (1980) have for various computational reasons argued that the operator most suitable to detect such changes is the filter ∇^2G, where ∇^2 is the Laplacian operator($\delta^2/\delta x^2+\delta^2/\delta y^2$) and G the two-dimensional Gaussian distribution

$$G(x,y)=e^{-\frac{x^2+y^2}{2\pi\sigma^2}}$$

with standard deviation σ. The Gaussian part of this function has the effect of blurring the image by whiping out all details smaller than the space constant σ (Fig. 5). Since contrast is a relative concept and occurs at different scales within an image, one must use several different values for the space constant in order to get a complete description of the light intensity changes. The next step in the construction of the raw primal sketch is to detect *zero-crossings* (a change in light intensity along a certain dimension will give rise to a peak or through in the first derivative and to a zero-crossing in the second derivative, Fig. 6) in the filtered image from which in turn the different primitives can be detected. What is interesting in the context of stereopsis is not so much the raw primal sketch itself, but the zero-crossings used to construct it. Marr and Poggio (1979) has suggested that zero-crossings are the most important, but not the only, input to the stereopsis mechanism. The idea of using zero-crossings seems to be, at least somewhat, supported by the neurophysiological findings described above. The

output of the retinal ganglion cells is probably quite similar to that of an image convoluted with a number of ∇^2G-operators with different σ-values. And the purpose of the simple and complex cells, responding to borders between brighter and darker areas, could possibly be to detect such zero-crossings within different spatial frequencies.

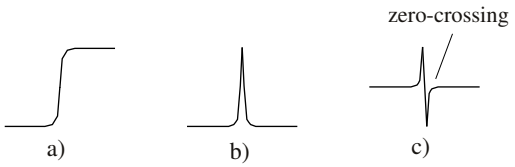

Figure 6. A change in light intensity (a) will rise to a peak (b) in its first derivative, and to a zero-crossing (c) in its second derivative.

However other primitives have been suggested to be important as well. Mayhew and Frisby (1981) showed in an experiment (using stereograms of saw tooth luminance gratings of the same period but with slightly different shapes) that the experienced percept could not be satisfactorily explained simply by considering zero-crossings. They therefore suggested that the "peaks" and "troughs" in the convoluted images should be matched as well. In this context, peaks and troughs refers to the maximum and minimum values in the convoluted image (Fig. 6c).

No doubt, the information corresponding to peaks/troughs and zero-crossings is of essential value to the matching process, but I believe that human stereopsis might be better described by a rather different framework than in terms of the detection and isolated matching of such features. I also believe that stating that the exclusive purposes of the simple and complex cells are to detect such features is a somewhat hasty, or at least too narrow, conclusion. To shed some light on my proposed alternative framework, I will describe two subtly, but yet fundamentally, different ways of interpreting the correspondence problem which are important to the context.

The most common interpretation of the correspondence problem is that the matching is conducted by first identifying some set of *predefined features* (e.g. bars or edges) in one image, and then finding the corresponding features in the other image. Theories relying on peaks/troughs, zero-crossings or other similar measurements for this purpose could therefore be said to be *feature-oriented* approaches.

Another way of looking at the correspondence problem is that a sub region (a delimited area) of one image is compared to other, similarly composed, sub regions in the other image (kind of like laying a jigsaw puzzle). A strategy like this would not be dependent of any particular set of predefined features, but would

instead rely on the similarity of the overall configuration of light within different regions. In contrast to being *feature-oriented*, this approach could be said to be *region-oriented*, since the descriptive element to be matched is a delimited region of the image.

With this alternative interpretation of the correspondence problem as a foundation, I will suggest a strategy in which the matching is conducted by comparing the configuration of contrast within elements/regions of different but fixed sizes. That the information of contrast is preferred rather than raw light intensity values should be evident from the discussion earlier in this section. Now, in order to fairly well describe an image in terms of contrast (remember that contrast is a relative measure), this information has to be gathered from within a number of different spatial frequencies. To efficiently make use of this information and to make the matching meaningful, only elements containing contrast information of the same spatial resolution should be matched.

Finally, for reasons that I will return to, I suggest that the sizes of these elements should be proportional to the spatial wavelength from within the information of contrast was detected. Thus, the larger elements will contain low-resolution contrast information and the smaller ones will contain high-resolution information.

What I believe is an advantage of this region-oriented strategy is that the matching can be carried out on a lower, "non-symbolic", level that is richer in information contents, since the matching is performed directly on the contrast values. In feature-oriented strategies, relying on the matching of a set of predefined features, these features would first have to be extracted from the information of contrast, and would thus be of a more symbolic nature since part of the information has been lost in the process of extracting them. I am therefore convinced that the suggested region-oriented approach would provide the matching process with a greater power of discrimination (allowing a greater reduction of false matches), than would any feature-oriented strategy relying on more "symbolic"/predefined features as matching primitives.

Since my ambition is to model the human stereopsis process, the suggested strategy would be of little value if the neurophysiological findings described earlier could not be accounted for by my model. I will therefore try to show, by interpreting these findings slightly differently, how they could be explained within the suggested model.

At first reflection the requirement that the matching should be conducted directly on the contrast values, corresponding to the output of the retinal ganglion cells, seems to lack any support in the neurophysiological findings. No cells with binocular receptive fields have been found that responds to the information at such a low level. What have been found are the simple and complex cells, which each responds optimally when a particular configuration of light is present, and thus only indirectly to "raw" contrast. These cells have therefore often been in-

terpreted as being “feature-detectors”. However, from the fact that these cells respond optimally to certain configurations of light does not necessarily follow that their purpose simply are to detect such isolated features in the image. I believe that the functionality of the simple and complex cells should not be explained, in isolation from each other, as feature-detectors. Instead I believe that the combined response from a group of such cells, sharing the same receptive field, could be seen as just another way of representing the information of contrast within their common receptive field.

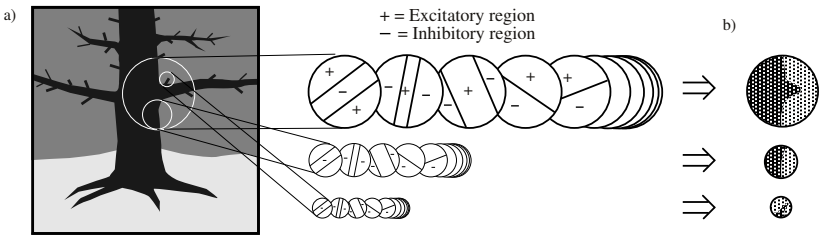

Figure 7. (a) Schematic organisation of the suggested groups of simple and complex cells, showing the sizes and compositions of these cells' receptive fields. 7 (b) is supposed to illustrate how the overall configuration of contrast, within the receptive fields, could be reconstructed from the “superimposed” response of the cells in the group.

To better see why such an interpretation makes sense, it is important to recall that there is a great variety of simple and complex cells. Both concerning the sizes of their receptive fields and concerning the configurations of light they are tuned to detect. Also important is that for any part of the visual field, there is a great number of such different cells that have common receptive fields. Now imagine how these various types of cells could be organised into groups, or columns, so that all cells belonging to a particular group would have the same receptive field, both in matter of size and location within the visual field (Fig 7a). These groups in turn could then be organised according to the sizes of their receptive fields into different layers, so that each separate layer only consisted of groups of cells with similar sized receptive fields. Now suppose that the matching does not rely on the individual responses from these different types of cells, but on the combined response from all the cells within such a group/column. In that case a more appropriate description of the purpose of the simple and complex cells might be that they could function as a form of *tuned detectors*. By tuned detectors I mean that these cells on a more continuous scale could measure, or “sample”, to what degree their tuned configuration of contrast is present within their receptive field, rather than just detect the presence, or non-presence, of a particular feature. With this view, the individual responses from these cells

would be of subordinate importance to the matching process, and instead it would be the summed, or "superimposed", response from all the cells within a group that mattered (as a mathematical metaphor this could be compared to how different wave functions can be superimposed to form a new wave function that is different from any of its individual parts but still contains the same information). With such an organisation in the back of the mind – not just literally speaking – it is possible to imagine how the various types of simple and complex cells, each and one, would contribute to register different aspects of the contrast-relationships, but that they together would represent the overall contrast-configuration within their common receptive field (Fig. 7b). Naturally would the resolution of the contrast, measured by any such group, be determined by the size of the common receptive field, or rather by the exact shapes and spatial extensions of the light configurations to which the individual cells are tuned to detect. However, by having several different layers of such groups, were each layer only contains groups of cells with similar sized receptive fields, this problem can be avoided and the contrast can be measured/"sampled" within several different spatial frequencies.

I believe this account shows how the activity in the simple and complex cells possibly could be interpreted as being just another form of representing contrast, and that this interpretation is as likely, or perhaps even closer to the truth, than an interpretation where these cells are described as "feature-detectors". There is thus a possibility that the human stereopsis mechanism relies on the correspondence of contrast values in the matching process.

Finally, one might wonder – if the "raw"contrast information really is matched – why would the brain do it in such an indirect way? One would imagine that the most straightforward way to conduct a matching of contrast values, would be to perform some kind of cross-correlation of matrices containing these values. One reason why no evidence of such an organisation is to be found is probably because such operations would be badly suited for a neural implementation. A point-by-point correlation of contrast values would require a much larger number of comparisons, that to be effective would demand a very high, almost "digital", precision. It might just be that by implementing this through the simple and complex cells, the same thing could be achieved in a more "analogue" way better adapted to the neural machinery. It is also possible that the information represented by the simple and complex cells are used by other systems within the visual pathway, and that this "design" therefore would be a form of "neural compromise" to simultaneously satisfy different requirements.

3 Constraints

No matter what matching primitives are used, false matches can not completely be avoided. There will always be ambiguous matches and in most images there

are areas that are impossible to match because they are visible from one only eye. Further processing is therefore needed to sort out the correct matches from the false ones. Exactly what then is this further processing? How can the right matches be separated from many possible "ghosts"? Without any knowledge about how the world behaves, this would be an impossible feat since any match would be as likely to be the correct one as the next. Fortunately, the world is bound by the laws of nature which imposes certain constrains on the behaviour of matter and energy. This makes some aspects of the behaviour of matter and energy predictable (e.g. solid matter is usually not transparent, a photon follows a straight line after being emitted, e.t.c.). If some of this knowledge was available to the brain, or rather the stereopsis mechanism, it could be used to constrain the search for the correct matches to certain sub-domains within the total domain of all possible matches. This would be possible since matches that were not in congruence with this "knowledge" – and thus not with the laws of nature – would be less likely to be correct. Of course this knowledge is not of an intellectual or conscious sort, but should rather be seen as built into the visual system by millions of years of evolution. The problem is to discover which of all potential physical constraints that could be important for the stereopsis mechanism. Many such constraints have been suggested and some seems to be more useful than others. Also, the suggested constraints are not always clear cut so there is room for different interpretations. For these reasons I will only discuss those constraints, which I believe are most important and relevant to my model.

The most important – and maybe most obvious – physical constraint is the fact that the search for the correct matches roughly can be restricted to a one-dimensional horizontal search. This is possible since our eyes are separated only horizontally, and the difference in perspective will therefore not affect the vertical positions of the features in the left/right images. Naturally, this alignment is not perfect but in practice correct enough to allow the search problem to be reduced from a 2-D one to a 1-D search problem.

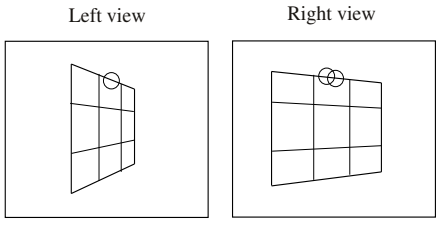

Figure 8. Due to the orientation of the surface, and the difference in perspective, the light from the marked edge will be projected onto regions of different sizes in the two retinas.

Marr and Poggio (1976) have formulated a constraint of *uniqueness*, stating that any given point on a surface can occupy only one location in space at a time. In a strict mathematical sense this formulation is true, but when applying this constraint to images caution has to be taken. To interpret this constraint correctly one must realise that the definition of a *point* can be ambiguous. In mathematical terms a point has no extension in space. When referring to a point in an image, the usual meaning is that of a small area of the image (however tiny the point might be it is still occupying a certain area). Now since the images that reaches our eyes are 2-D projections of 3-D structures, and due to the difference in perspective, there is no guarantee that any particular surface will be projected onto areas of equal sizes in the two retinas (Fig. 8). It would therefore be wrong to state that any particular point in one image should be matched with only one other point in the other image. I believe this observation is important and it shows that this constraint should not be implemented in a too strict sense (not in an exclusive/or manner), but in a way that allow for some "overlap". In fact, *Panum's limiting case* (Fig. 9) seems to indicate that the human stereopsis mechanism makes use of a more relaxed form of this constraint. In *Panum's limiting case*, a feature in one image can be matched with either of two identical, horizontally separated, ones in the other image, and the resulting perception is that of two identical features hovering at different depths.

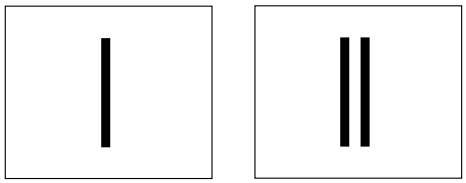

Figure 9. Illustrating *Panum's limiting case*. The bar in the left image can be matched with either of the two bars in the right image. When fused the experienced percept is that of two separate bars, hovering at different depths.

The main part of all light that reaches our eyes is reflected from surfaces of solid matter. Solid matter is per definition continuous. The atoms are closely and strongly tied together into larger units (e.g. crystals, rocks, cells, plants). The surfaces of solid matter will therefore be more or less continuous or smooth. This physical fact has been exploited in a number of suggested constraints.

Marr and Poggio (1976) has formulated a constraint of *continuity* stating that the disparity of matches should vary smoothly over the image, except at the boundaries of objects, because the distance to neighbouring points on a surfaces generally varies continuously.

Pollard, Mayhew and Frisby (1985) has for similar reasons justified the use of a *disparity-gradient* limit to constrain the search for matches. The disparity gradient is a relative measure of the change of disparity between two neighbouring points in an image. In a number of psychophysical studies they found that the human stereopsis system seems to favour matches that are within a disparity gradient value of 1.

Mayhew and Frisby (1981) have also suggested a constraint of *figural continuity*, which is a bit more interesting in the context of the model to be presented. Due to the continuity of matter and the generally smooth changes of depth in an image, the relative spatial relationships between features will usually be preserved in the left/right image. A match will thus more likely be correct if the features in its near vicinity are similar to the ones in the image from which the matching was initiated. This constraint of *figural continuity* has a central role in the model I will present, since it is inherent in the choice of matching primitive.

4 Spatial frequency channels

There is a great deal of evidence suggesting that the visual system relies upon a set of independent *channels*, of different coarseness, in the monocular analysis of the image, probably corresponding to receptive fields of different sizes (Poggio & Poggio, 1984). It therefore seems likely that such channels also could be important for stereopsis. In fact, there are evidence indicating that the matching, at least to a certain degree, is conducted independently within such channels. For instance has it been known since long that images with high frequency noise added to them (resulting in rivalry within the higher resolutions) still can be binocularly fused if the noise leaves the lower frequency information unaffected, which thus still can be correlated (Julez & Hill, 1978). One assumption about these channels, supported by psychophysical observations (Felton, 1972; Kulikowski, 1978; Levinson & Blake, 1979), is that the coarser channels detect large disparities while the finer channels can match only small disparities.

However, the purpose of, and activity within, these channels should probably not be described as being completely isolated and independent of each other. Although the initial part of the matching procedure could be performed within independent channels, there is still the possibility that the output, from this initial matching, is combined at a later processing level, at the level where ambiguous and false matches are dissolved. Evidence in this direction has been found by Mayhew and Frisby (1981) (with the “missing fundamental” experiment and with spatial frequency filtered stereograms portraying corrugated surfaces). The important question then is how the information from such independent channels could be combined to reduce the set of false matches.

Before giving my own account for how I believe this could be done, I will briefly describe a model of stereopsis devised by Marr and Poggio (1979) that

has inspired me. The matching primitives used in this algorithm were zero-crossings, derived from different spatial resolutions. The main idea is that within the lower resolutions the number of zero-crossings will be relatively few, and not too close, and the matching will therefore result in few false matches. Once the set of potential matches has been established from the lowest spatial resolution, this information is written down into a memory buffer. The disparity information in this buffer is then used as starting point for the matching of zero-crossings of a higher resolution, within a smaller range of disparity. When this procedure has been repeated for all the successively finer resolutions, the resulting set of matches can, with a high probability, be considered to be the correct set, since most of the false matches simply have been avoided (see Marr & Poggio, 1979, for a mathematical analysis of these conclusions).

Although my model is similar to the Marr-Poggio model, there are still a number of important differences, and my arguments for how the information from different spatial channels are used are not directly built upon any mathematical analysis, but instead closely tied to the concept of figural continuity.

To see how the information, from different spatial channels, could be combined in my suggested model, it is important to understand some of the physical properties of the proposed matching primitive – or rather matching unit (delimited regions containing arbitrary contrast configurations). These properties in turn are determined by factors, inherent in the correspondence problem, which has to do with the fact that the world is made up of 3-D objects, while the images that hits our eyes are 2-D projections of the surfaces of these objects. The important thing to realise is that within an image, the larger the considered region is, the greater is the probability that the different features are projections of surfaces at different depths. Now, since the suggested matching procedure relies on the similarity of the contrast configuration, within different regions of the images, it becomes evident that the sizes of the regions in consideration will affect how the within-channel-matching results should be interpreted. And since the matching is performed independently on elements of different sizes, containing contrast information of different resolution, the conclusions that can be drawn from the results of this matching will be quite different from channel to channel. Roughly speaking, it is a matter of trade-off between the accuracy of the measured disparity and the probability that a match is correct.

Considering the larger matching elements, which contain lower frequency spatial information, each of these cover a relatively large region of the image and will thus be more likely to contain information from surfaces with larger variation in depth. This fact has two important implications. First, the slight distortion between the two images, due to the larger variation in disparity, will have the effect that certain parts within two correctly matched elements might be uncorrelated or even negatively correlated. However, due to the lower resolution, which has the effect of blurring the contrast information, and the fact that the rel-

ative spatial relationships almost always are preserved, the total correlation of two correctly matched elements will be positive. Second, due to the mixture of the disparity information within these elements, the result of two correctly matched elements will only give a rough estimate, or average, of the actual disparity within that region. To resume, the negative aspect of using larger elements is that the result from the matching will not be very specific, but will instead give an estimate of a sub-domain in which the correct disparity is to be found. The positive aspect is, because a larger region of the image is considered, that it is unlikely that any region outside of this sub-domain will show the same figural continuity. In other words will the result of the matching not be very precise, but it will with a high probability indicate within which range, or sub-domain, the correct disparity lies.

Turning to the smaller elements, by simply inverting the arguments, these will be shown to display the opposite properties. Since these elements are used to match the higher resolution information, within smaller regions of the image, the different features within these elements are more likely to correspond to surfaces lying at similar depths. Thus will the distortion between two correctly matched elements be quite low. This means that the disparity measure, for two correctly matched elements, will be quite specific, and also that the resulting correlation will be relatively strong. The negative side of the coin is that the high resolution, and the small sizes of the considered regions, means that there will be a greater number of regions that exhibit similar configurations of contrast. Thus, due to the high resolution but lack of reliance on figural continuity, the matching within the finer channels will result in quite specific disparity measurements, but also give raise to a considerably higher amount of false matches.

Considering the conclusions above, it would clearly be desirable if one could combine the best properties of the information provided by these different channels. Preferably, this would be done by somehow letting the coarser channels, corresponding to the larger elements, guide the matching of the smaller elements, similar to the idea described earlier in the model of Marr and Poggio. Before describing the whole of my model and putting the parts together in the next section, I will close this section by briefly commenting on some of the main differences compared to the model of Marr and Poggio.

Apart from the different choices of matching primitives, the major difference is the reliance of figural continuity in my model, while this is not considered in the Marr-Poggio model. No matter what mathematical arguments they use to justify that the false matches simply can be avoided (by considering the channels one at the time and in order from coarser to finer), this still requires that the zero-crossing used to initiate the matching is the correct one from the beginning. In my opinion, this problem (of finding the correct “starting-point”) can not be solved without considering figural continuity. Further, in Marr and Poggio’s algorithm the matching is performed in steps of successively finer resolutions,

where at the end of each step the result is written down into a memory buffer, which then is used as the starting point for the next level. In the model I am suggesting, the matching is performed simultaneously within the different channels, and the activation in the larger channels are directly affecting the activation in the finer channels. There will thus be no unnecessary delay caused by the waiting for input from the coarser channels, nor is there any need for an extra memory buffer storing intermediate results.

5 The Model

In the following two sections I will present a model of human stereopsis that is built upon the different ideas discussed in the earlier sections. For pedagogical reasons I have chosen to divide this presentation into two levels. In this section I will give only a general account for how the main ideas could be implemented, and present an overview of how the different processing levels are structured and how the information is passed between these different levels. In the following section a computer implementation of the model is presented which better describes some of the details. However, before starting this presentation I would like to jump ahead for a minute and discuss an exception in the model that deserves special attention. This exception concerns a simplification in the implementation of the matching process.

One aim I have had with this paper is to show that the correspondence problem can be solved more efficiently if the matching is conducted by a direct comparison of contrast values, rather than by comparing a set of more “symbolic” features. I have also tried to show, by interpreting the functionality of the simple and complex cells slightly differently, how these cells possibly could represent the information of contrast. An important assumption for the validity of the model is therefore that the proposed groups of simple and complex cells actually are capable of representing the contrast information, with a precision equal to that of the output of the retinal ganglion cells. In order to support this assumption it would be desirable if such a model could simulate the individual responses from each and one of these cells. Unfortunately, the algorithm in question is not designed to model the stereopsis process in such an elaborate way. In short there are two major reasons why it would be difficult to implement such a model. First of all, the physiological knowledge of the visual system is not complete enough to allow for the construction of such an exact model. Not only is it uncertain exactly to what kind of stimuli many of these cells respond optimally to, nor is it known exactly how they are distributed over the visual field. The second reason is of more practical nature and concerns the fact that such an implementation would require a considerable amount of memory and processing capacity. Unfortunately, due to limitations in computer power, such an explicit implementation has been out of the question, and instead I have been forced to implement a

somewhat simplified matching process that relies on a form of cross-correlation of contrast values.

Thus, the validity of this model relies on that the above assumption, about the functionality of the simple and complex cells, holds. However, to my defence I would like to say that although my model relies on a critical assumption, I believe this assumption is not more daring than the assumptions of most other models, and it should therefore be judged with this in mind. With all this said I will now return to the presentation of the model.

5.1 Input and convolution

Starting with the input, consisting of the raw intensity values of the two images (Fig. 10, level A), the first step is to extract the contrast information within the images. To detect the contrast information within different spatial frequencies, each image is convoluted with the 2-dimensional operator $\nabla^2 G$, with three different values for the space constant σ (Fig. 10, level 1). Apart from computation-

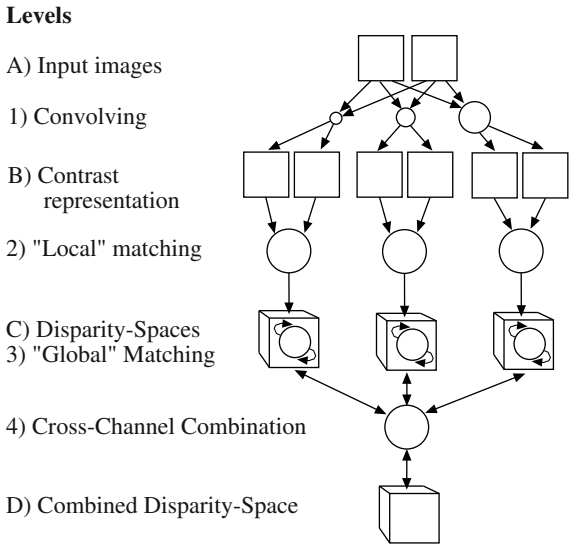

Figure 10. Schematic overview of the different levels of representation and processing. Representational states are shown as squares/cubes and are labelled with letters (A–D). Processing stages are displayed as circles and are labelled with numbers (1–4). (A) The input stereogram. (1) Each image is convoluted with three different $\nabla^2 G$-operators. (B) Contrast representations. (2) Initial, or “local”, matching. (C) Disparity-spaces. (3) “Global” matching. The constraints of uniqueness and continuity are implemented by the inhibition and excitation of nodes/cells within the disparity-spaces. (4) Cross-channel combination. (D) Combined disparity-space (“result”).

al reasons presented by Marr and Hildreth (1980), I have chosen this operator because the result of an image convoluted with this filter seems to resemble that of the output of the retinal ganglion cells. I will save the exact details about the sizes of these filters for the next section, but here it will be enough to say that the radius of the central part of the filter is doubled for each successively larger filter. After the images have been convoluted we thus have six sets, or three pairs, of separate contrast representations (level B), where the spatial resolution of the contrast information for each pair is determined by the size of the filter (the space constant σ) used to produce it.

5.2 Matching procedure

The next step is to perform the initial, or "local", matching procedure (Fig. 10, level 2) to establish the set of all potential matches. This matching is conducted independently, and in parallel, on the three pairs of contrast representations, thus resulting in three different sets of potential matches (Fig. 10, level C). As suggested earlier the general idea is that each contrast representation is divided into a large number of partly overlapping regions, corresponding to the receptive fields of the suggested groups/columns of simple and complex cells, and that the contrast values within these regions are then cross-correlated with the contrast values within such regions in the other image. An important matter that remains to be considered is how large these regions should be in relation to the spatial resolution of the contrast information.

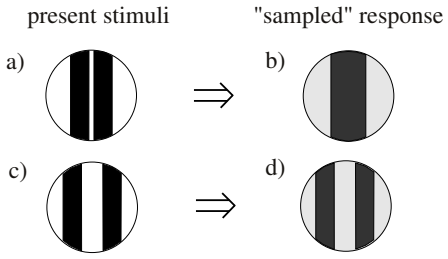

Figure 11. (a & c) Present stimuli within the receptive field of a group of simple and complex cells. (b & d) Showing the (assumed) "sampled" response.

The problem is to establish some kind of relationship between these two factors, that could reflect the relationship between the resolution of the contrast, "sampled" by a group of simple and complex cells, and the size of their common receptive field. Naturally, it is hard to justify any such relationship in a strict mathematical sense. However, if one considers what type of stimuli these individual cells responds optimally to, it is clear that there must be a limit to how

high this resolution, or to how complex the overall configuration of contrast within this receptive field, can be.

As an example, consider two parallel "bar"-like features that are present within the receptive field of such a group (Fig. 11a). If these were too close to each other, the resulting "sampled" response would probably be more similar to that of one thicker bar (Fig. 11b). On the other hand if they were further apart (Fig. 11c), they would more likely be detected as two separate bars (Fig. 11d). My point with this example is to show that the resolution, of the contrast information that such a group of cells could measure, probably would depend very much on how close the changes in light-intensity are. In more mathematical terms, if one considers the second-derivative of the light-intensity values along any dimension within the receptive field, one could say that there should not be too many such changes (zero-crossings) of the same sign, and that they should not be too close, if the present configuration of contrast is to be measured/sampled "correctly". To relate this observation to my algorithm and formulate a more concrete relationship, I have decided to restrict the size of the regions, to be cross-correlated, to the size roughly corresponding with the central part of the filter that was used for the convolution. It can be shown that within such a region, of a filtered image, there in the general case (or with randomly produced light-intensity values), with a high probability, will be only one zero-crossing with a particular sign and orientation along any dimension within the region (see Marr, 1982, for a full mathematical analysis).

Having divided each contrast-representation into partly overlapping regions, of sizes determined by the sizes of the filters used for the convolutions, the matching within each "channel" is performed as follows.

To establish the degree of correspondence between two regions, a point-by-point cross-correlation is performed on the contrast values within these regions. A problem with performing an "ordinary" correlation is that two (equal) low-contrast values will result in an as good correlation as will two (equal) high-contrast values. Two regions containing no contrast would thus be considered as perfectly matched. This would go badly with the fact that the individual simple and complex cells only responds to stimuli where there is change in the light intensity. To reflect this in the matching procedure, each point-by-point correlation is weighted with a factor that is proportional to the strength of the weakest of the two contrast values. The result of these correlations are then added up and divided with the total number of correlations within the region in order to receive a normalised value. These normalised values will then all lie in the range between –1.0 and 1.0. A high such value indicates that the two regions correspond fairly well, and that they contain a high amount of contrast. A low value indicates either low contrast, and will thus be of little interest, or that there within the region are different sub-regions that considered individually are positively and negatively correlated, but when taken together will cancel out the value for the

whole region. Finally, a high negative value simply indicates that the two regions do not match well at all. Now this particular algorithm is only concerned with the degree of similarity of two regions, and therefore will only the positive values be of interest. All negative values are therefore set to zero and will consequently be considered as bad matches.

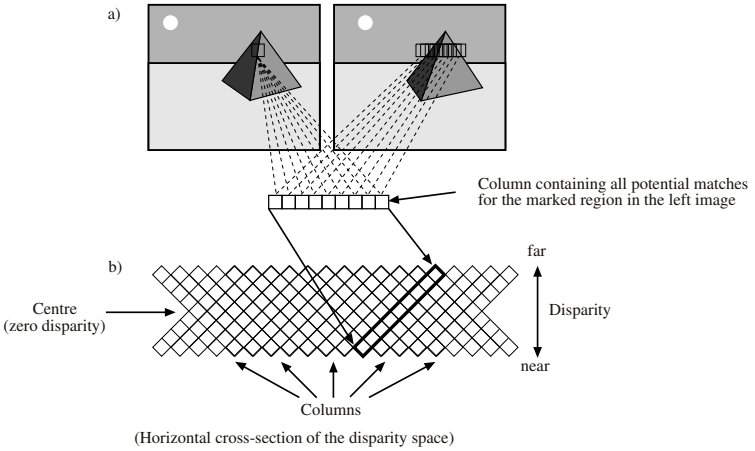

Figure 12 (a) The search for potential matches is restricted to consider only regions, in the opposite image, that are horizontally shifted, and which lies within a certain range from the same relative position as the region from which the matching was initiated. (b) The result of each of these comparisons are then mapped into the corresponding column in the disparity-space.

Since the purpose of the matching procedure is to establish the disparity between two corresponding regions, each region has to be matched with a number of different regions in the contrast representation of the opposite image (Fig. 12a). As described earlier this search can basically be restricted to consider only regions that are horizontally shifted, but since it is (practically) hard to perfectly align two images, the search is performed within a small vertical range as well. The area delimiting this search can be seen as the equivalent of *Panum's fusional area*. In human stereopsis, *Panum's fusional area* refers to the binocular region in which two features must lie in order to be correctly fused (Poggio & Poggio, 1984). The results of these individual comparisons are then mapped into a 3-dimensional, topologically ordered, *disparity-space* (DS). A horizontal cross-section of such a disparity-space is shown in figure 12b. This structure consists of a large number of nodes, or "cells", where the degree of activation in each cell represents the result of a comparison of two regions. Each column in a

disparity-space thus corresponds to a particular region of the image, and each node within these columns represents a particular disparity, with zero-disparity at the centre node. After each region has been matched and mapped into the disparity-space, the "local" matching procedure is completed and the result is that of three separate disparity-spaces (schematically portrayed as cubes in fig. 10, level C), produced from the three different pairs of convolutions. The rest of the algorithm is basically concerned with one problem, and that is to determine which nodes, of all the active ones, that indicate correct disparity values, and which have been activated due to false matches.

5.3 Implementation of the constraints

To solve the problem of false matches some of the constraints that were discussed in the two earlier sections have been incorporated into the algorithm. Of particular interest are the constraints of *uniqueness* and *continuity*, but also how the information from the different channels can be combined to further reduce the set of potential matches. The way I have chosen to implement the first two of these constraints have been greatly inspired by an early cooperative model of Marr and Poggio (1976), in which these constraints were implemented by the inhibition and excitation of interconnected "neurones", in a structure similar to the disparity-space described above.

To recapitulate, the constraint of *uniqueness* suggests that any point on a physical surface can have only one 3-D location in space, and thus any feature in an image should be matched with only one feature in the other image. Apart from the objections presented earlier this conclusion is fairly correct, and since a feature per definition is bound to have a 2-D spatial extension in the image, the same basic argument holds when matching regions. Considering the disparity-spaces described above, this means that only one of the active nodes, in each column, can represent the correct disparity.

The constraint of *continuity* in turn is motivated by the fact that surfaces generally are smooth and continuous, except at their boundaries, and the measured disparity should therefore also vary smoothly over the image. For the same reason the relative ordering of the features, in the two images, should also be preserved. This latter aspect is often referred to as *figural continuity* or as the *ordering* constraint. Thus considering the disparity-spaces, active neighbouring cells representing similar disparities should be preferred instead of isolated active cells.

To see how these constraints can be implemented, consider a horizontal cross-section of the disparity-spaces (Fig. 13). Now the constraint of uniqueness is implemented simply by letting all the cells in a column inhibit the activity of each other, where the strength of the inhibition is proportional to the total activity of the cells in the column. Since each cell in the disparity-space is a member of two

columns, one corresponding to a region in the left image and vice versa, each cell will be inhibited by the activity in two columns.

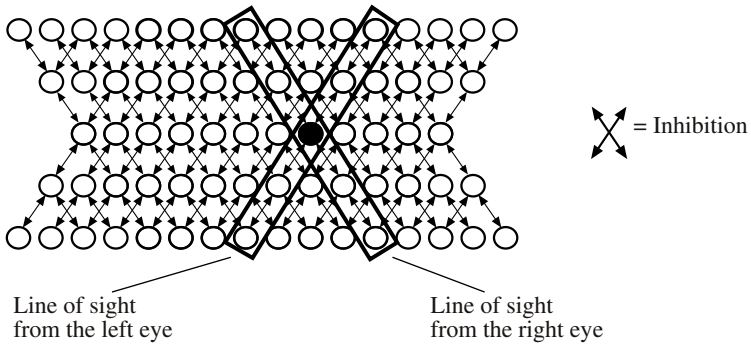

Figure 13. Horizontal cross-section of a disparity-space. The constraint of uniqueness is implemented by letting all cells, along the two lines of sight, inhibit each other.

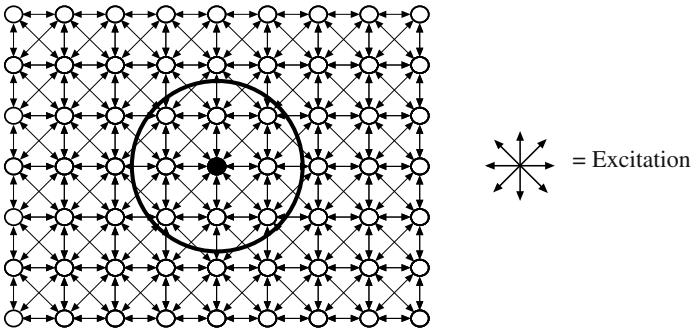

Figure 14. Vertical cross-section of a disparity-space. The constraint of continuity is implemented by letting all active cells excite the cells, in neighbouring columns, that representing similar binocular disparity.

The constraint of continuity is implemented in a similar, but opposite way, by letting the activity in each cell positively influence neighbouring cells in surrounding columns, which represents matched regions of the same binocular disparity (Fig. 14). Each cell is thus exciting their neighbours within a disc-shaped region of the disparity-space, in the horizontal-vertical plane and with the centre at the exciting cell.

5.4 Cross-channel-combination

This mutual inhibition and excitation of cells is performed independently within each of the three disparity-spaces, thus leading to somewhat different results. As argued in the previous section the matching process could benefit from combining these different results by letting the activity in the coarser disparity-spaces guide the activity in the finer channels. To implement this idea a fourth disparity-space is introduced (Fig. 10, level D), in which each cell is excited by the combined activity of the three cells, with the same relative 3-D position within the three original disparity-spaces. Thus, cells in this *combined disparity-space* (CDS) that are excited by all three channels will be more activated than those only receiving activation from one or two channels. Now to recall the discussion in the previous section, the activity in the coarser channels will be more diffuse, but also more concentrated to certain sub-regions, within the disparity-space, that are more likely to hold the correct matches. Thus could cells in the CDS that lie within such sub-regions, and that also are excited by cells from the finer channels, be considered as more likely to indicate the correct disparity than those that lie outside of these sub-regions.

Finally, to favour the correctly activated cells in each of the three original channels, the activity in the CDS is feed back to the corresponding cells in each of these, and the process is repeated until the activity of all the cells has been stabilised.

6 Implementation

The program code of this implementation was written in the C-language and is about 750 lines long. In order to save some space and to make the program available to readers not familiar with C, I will only present the more important features of the implementation, and instead of the original C-code I will use a more general form of notation that hopefully could be understood by a majority of readers.

Input: Each image is represented as a 128 x 128 byte matrix, where each byte represents a light intensity value ranging from 0–255 (0 = black, 255 = white).

Step 1 (*Convolution*): To detect the contrast relationship within each image, the 2-dimensional ∇^2G-operator (described earlier) is used with three different values for the space constant (σ=1, 2 and 4 pixels). To normalize all contrast values, (Cx,y) {0<x<128, 0<y<128}, they are divided with the value of the absolute product of the light intensity value and the value given by the ∇^2G-operator, summed over the region covered by the filter centred at (x,y). More formally, the normalized convoluted value, C^N, at point (x,y) are given by the following equation:

$$C^{N}_{x,y}=\frac{\sum_{s=-r}^{r}\sum_{t=-r}^{r}\nabla^2 G(s,t)I_{x+s,y+t}}{\sum_{s=-r}^{r}\sum_{t=-r}^{r}\left|\nabla^2 G(s,t)I_{x+s,y+t}\right|},$$

where r=4σ and *I* is a matrix containing the light intensity values.

Step 2 (***Matching***): Each contrast representation is then divided into a number of regions that is equal to the number of pixels in the original images. Thus, two neighbouring regions will almost completely overlap each other since they are shifted by only one pixel. To establish all potential matches and construct the disparity-spaces, each region is matched with 21 different regions in the other image. For example, to establish the set of potential matches for a region in the left image, centred at pixel (x_L,y_L), the region in question is matched with all regions in the right image that are centred within a 10 pixel range of the pixel (x_R,y_R) in the right image, which has the same relative position as the centre of the left region (x_R=x_R, y_R=y_R). Each such set of comparisons corresponds to one column in one of the disparity-spaces (see above).

The matching, or cross-correlation, of a region in the left image centred at (x_R,y_R) with a region in the right image centred at (x_R,y_R) is formally described by the following equation:

$$C=\frac{1}{(2r+1)^2}\sum_{x=-r}^{r}\sum_{y=-r}^{r} sign(L_{x,y}\cdot R_{x,y})\cdot min\left(\frac{|L_{x,y}|}{|R_{x,y}|},\frac{|R_{x,y}|}{|L_{x,y}|}\right)\cdot W\left(min(|L_{x,y}|,|R_{x,y}|)\right),$$

where **r** is the radius of the matched region, which is equal to the space constant (σ) used for the particular convolution. *L* and *R* are matrices containing the normalized contrast values for the left region centred around (x_L,y_L), and the right region centred around (x_R,y_R) respectively. The function

$$W(x)=1-e^{-c|x|},\ c\approx 6$$

returns a value between 0.0 and 1.0 that is proportional to the strength of the weakest of the two contrast values. As explained earlier, the purpose of this component is to avoid high correlation values when there is low, or no, contrast within a region. The result of the whole matching (*C*) will be in the interval [–1.0, 1.0], but since only the positive values are of interest all negative values are set to zero.

Step 3 (***Constraints***): After the matching procedure have been completed, every node, "or cell", in the disparity-spaces will have a value, or activation, between 0.0 and 1.0. These values are now used as input for the next layer of

processing. The new value each node will receive is determined by three factors: the current degree of activation, the strength of inhibiting "cells" lying along the same two lines of sight, and the strength of exciting neighbouring "cells" representing similar disparity. The following functions describes how the new activation value (A_N) is computed for a node in a disparity space:

$$A_N(A_C, P, N) = A_C + Excitation(P) - Inhibition(N) ,$$

where A_C is the current activation. P is the positive contribution given from surrounding cells, with similar disparity, that lies within a radius equal to the radius of the regions that where matched to produce the particular disparity-space. The contribution each of these cells give is directly proportional to the activity in the contributing cell, and proportional to the inverse of the squared distance to the receiving cell. In other words, more distant cells will contribute less to the excitation. The purpose of the function

$$Excitation(P) = \frac{P}{P+c}$$

is to moderate the positive contribution to the cell so that the change from the current value to the new one will be smooth, and also to avoid that the new value becomes larger than 1.0. The constant c is used to normalise the value of P and is equal to the sum of the squared inverse of each of the distances from the receiving cell to the contributing cells. N is the negative contribution (the summed activity of all cells lying along the same two lines of sight). The purpose of the function

$$Inhibition(N) = 1 - \frac{1}{(1+N)^c}$$

is (the same as for the function Excitation(P)) to avoid too rapid changes of the activity in the cell. The c constant (c=0.18) determines the strength of the inhibition. This value was empirically found to balance the average positive and negative contributions.

Step 4 (*Cross-Channel Combination*): The combined disparity-space (CDS) is produced by simply multiplying the values of all cells, that has the same relative 3-D location, and then raise the product to one third, so that the new value will be unchanged if all three values are the same. A reason for multiplying the values rather than just add them is that by doing so, only matches that are present within all three channels will survive.

Step 5 (***Feedback***): Before repeating the whole sequence from step 3, each cell in the three original disparity-spaces will receive a new value that is determined by three factors: the result of the initial matching, the current activity in the cell and the activity of the cell in the CDS that has the same relative 3-D location. These new values are produced in the same manner as in the cross-channel combination (step 4), by raising the product of these three values to one third.

7 Results

The results presented in the following pages were all produced by the computer implementation described above. The results show the processing of five different stereograms. The first three stereograms are made up of artificially produced images. These stereograms were partly designed to be as simple as possible, but also to illustrate some of the different effects imposed by the constraints. The last two stereograms are made up of natural images, and therefore better shows how the model behaves with more "natural" input.

Before going into the details of each processing a few words about the form of the presentations are in place. For each stereogram below the activity within the CDS will be presented in three different ways. The first type of result shows the activity within the CDS directly after the initial matching procedure has been completed (step 2 in the algorithm above). The activity within the CDS is displayed "slice-by-slice" (vertical cross-sections), with increasing depth from left to right, and from top to bottom. Further, the activity within the "cells" in each layer is displayed in a gray-scale, where brighter regions indicate high activity and darker regions indicate low, or no, activity. In the second type of results, the activity is shown after a number of iterations (corresponding to the loop of step 3, 4 and 5 in the algorithm), after that the activity has stabilised within the network of nodes. Here too the activity is displayed in a "slice-by-slice" manner, but instead of using a gray-scale, the original (left) image has been mapped onto the regions that still are active (activity > 0.2), so that the reader better can see to which part of the stereogram the active regions correspond. The last type of result also shows the activity within the CDS after a number of iterations, but here the maximally activated nodes (within each column of the CDS) have been tied together to form a wire-diagram.

7.1 Trial 1

Starting simple, figure 15 shows a stereogram with three groups of thin vertical lines. For those readers not capable of fusing stereoimages, the three groups form a triangle (in the horizontal-depth plane) where the middle group is closest and the rightmost group lays furthest away. Although simple this example clearly demonstrates how efficiently the cross-channel combination resolves false matches. Figure 16 shows the results of the convolutions with the three dif-

ferent filters. If only the information within the finest channel (fig. 16a) was considered, it would be difficult to establish the correct set of matches since each line could be matched with several other lines in the other image. However, due to the facts that the resolution in the coarser channel is lower, and the size of the matched regions are larger, there will be no ambiguity in the coarser channels since the thinner separate lines, within the three groups, will not be present (fig. 16c).

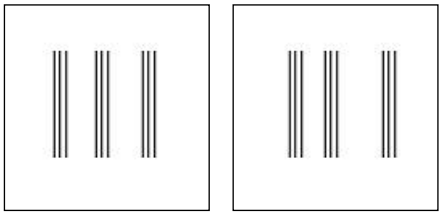

Figure 15. Input stereogram.

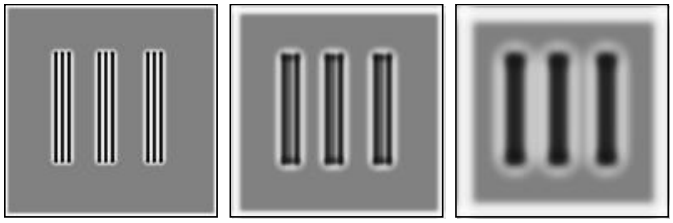

Figure 16. Result of the convolutions. If only the information within the finest channel (a) was considered, each of the thinner lines could be matched with any of the three thin lines, in the corresponding group, in the other image. However, in the coarser channels (b and particularly c) there is no such ambiguity, and the information within these channels will therefore "guide" the activity within the finest channel, so that the false matches can be dissolved.

As the results shows, the correct set of matches is considerably more activated than the false ones, even directly after the initial matching procedure (fig. 17). And after only three iterations the false matches have been dissolved almost completely (fig. 18 and19).

Unfortunately, the implementation of the cross-channel combination also seems to cause a few side effects. One of these can be noticed, in figure 18, in that the established disparity extends a bit outwards from each group of lines. Largely, this "filling in" (or in this case "floating out") effect could be ascribed

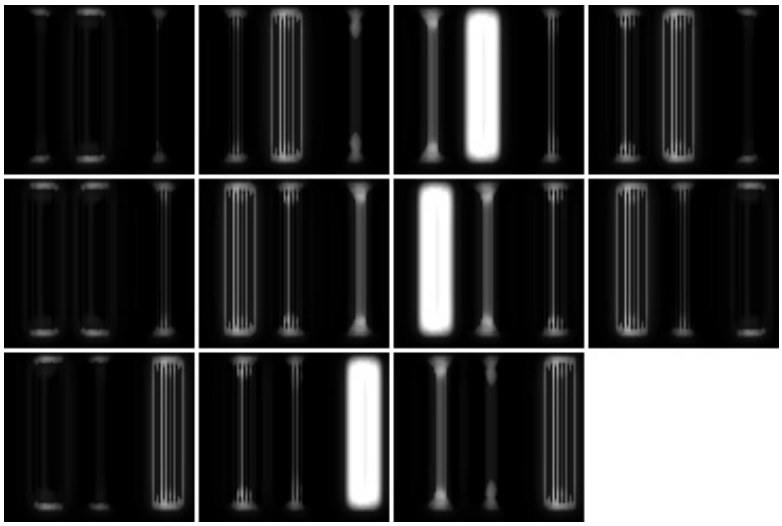

Figure 17. Each image above shows the activity within a vertical cross-section of the CDS. The brighter areas indicate high activity (potential matches). Depth increases from left to right, and from top to bottom.

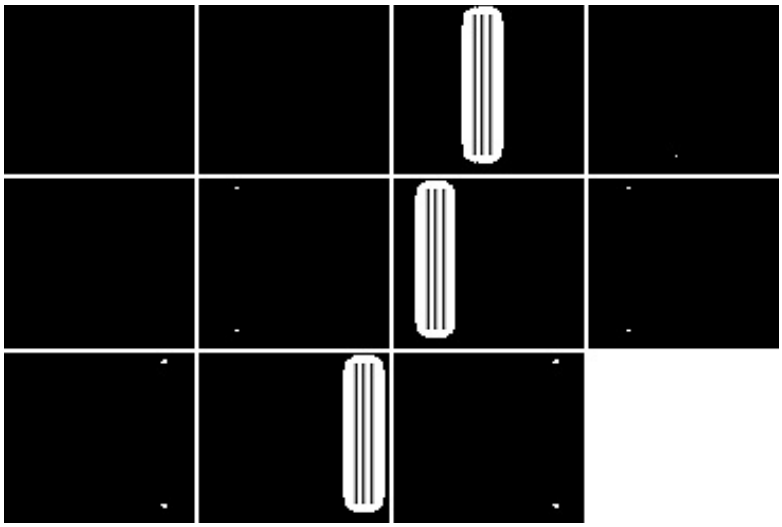

Figure 18. Activity within the CDS after 3 iterations. The original (left) image of the stereogram has been mapped on top of areas that still are active.

to how the constraint of continuity has been implemented (by the excitation of neighbouring nodes in the disparity-spaces), but in part also to the implementation of the cross-channel combination. Since highly activated nodes in the coarser channels spread their activity over relatively larger regions in the finer channels. And thus might activate nodes in the finer channels that were not active initially. However, seen from a purely technical view, it is difficult to definitely state that this behaviour is incorrect, since it is impossible to establish any depth information about the white background. If the background had some kind of texture (which often is the case in natural images), its depth could be established and thus would the correct matches (for the background) "override" the activation caused by the side effect.

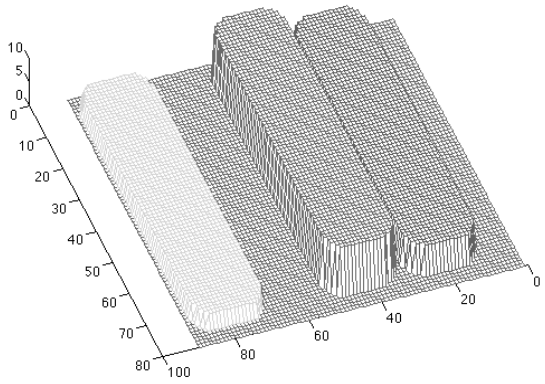

Figure 19. Wire-diagram of the disparity (activity) within the CDS after 3 iterations.7

7.2 Trial 2

The next motif is a bit more complex. Figure 20 shows a random-dot stereogram with a 25% density of black dots. When fused three different planes can be perceived. The closest plane frames the scene and has a rectangular opening at its centre. The next plane lies further away and also has a rectangular opening at its centre. The third plane is located furthest away and can be seen through the "hole" that is formed by the openings of the two other planes.

This example too shows how efficiently the false matches are dissolved, by combining the information within the three different channels. Again, if only the information within the finest channel was considered it is easily seen that any dot could be matched with numerous other dots in the opposite image. However, due

to the greater reliance on figural continuity, within the coarser channels, it is less likely that any two incorrectly matched regions within these channels will be highly correlated. And thus by combining the rough estimate of disparity, from the coarser channels, with the more precise information within the finer channels a large amount of false matches can be ruled out.

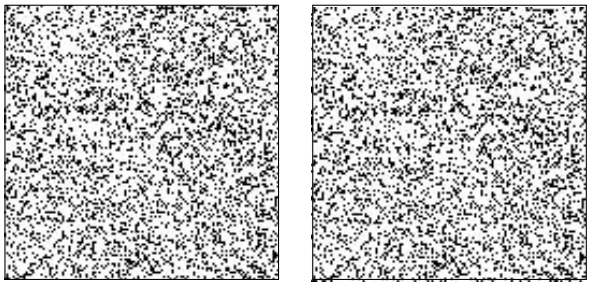

Figure 20. Random-dot stereogram with a density of 25% black dots.

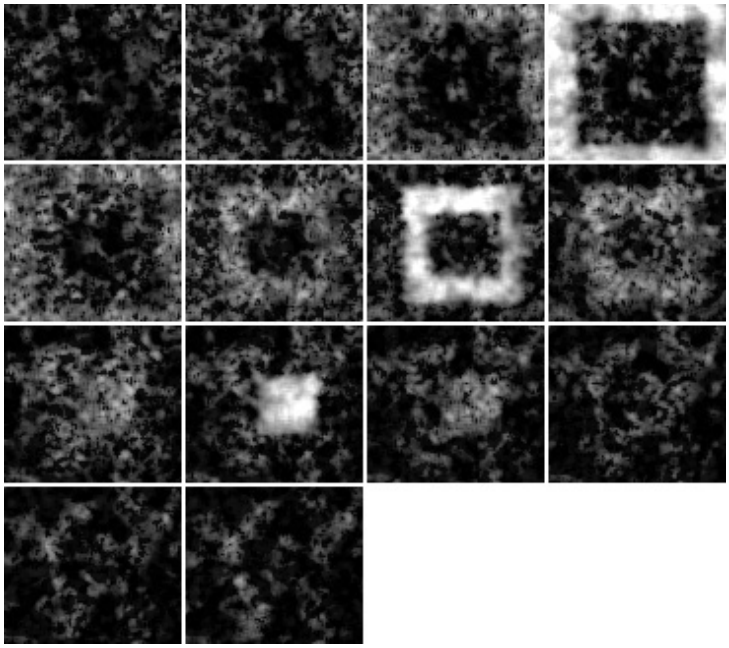

Figure 21. Activity within the CDS after the initial matching procedure.

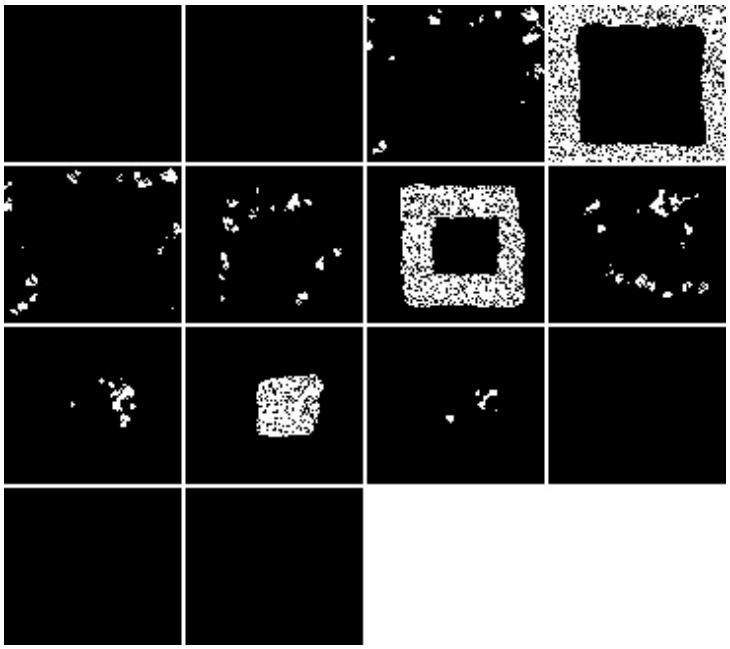

Figure 22. Remaining activity after 5 iterations (with the original left image mapped on top).

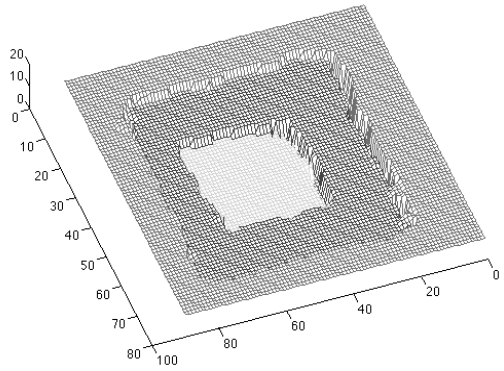

Figure 23. Wire-diagram of disparity (activity) within the CDS after 5 iterations.

As the results of the initial matching procedure (fig. 21) shows, one can distinguish the three different planes even before the constraints of uniqueness and continuity has been applied. And after only 5 iterations (fig. 22 and 23), only a few false matches remain active.

In the previous example (trial 1) I pointed to one of the side effects, caused by how the cross-channel combination was implemented, that for some motifs can cause questionable results. In this example however, the same side effect could be seen to have a positive influence on the result. Since the activity within the coarser channels is spread over to intermediate regions of nodes in the finer disparity-spaces which were not initially active, the resulting disparity-map will be more continuous (i.e. the points in each plane will be tied together).

7.3 Trial 3

One strength of the model is that it seems to be quite robust, in the sense that it performs satisfiable even if a substantial amount of "noise" (uncorrelated information) is added to the stereogram, or if the individual images are slightly shifted vertically.

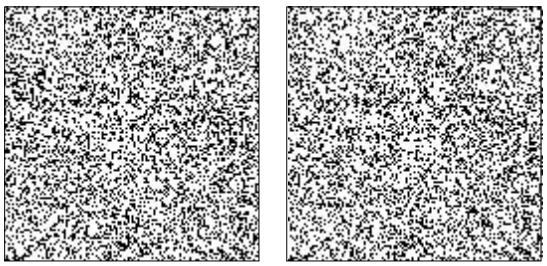

Figure 24. Random-dot stereogram where only 75% of the dots are correlated.

An example of the insensitivity to "noise" can be seen above. The stereogram in figure 24 is the same as in trial 2, except that an additional number of dots have been introduced so that only a total of 75% of the dots are correlated (i.e. 25% of the dots, in each image, have no corresponding match in the other image).

Due to the added noise the resulting activity after the initial matching (fig. 25) is much less pronounced than what were the case in the in the two earlier examples. Nevertheless, after 7 iterations (fig. 26 and 27), roughly the same three planes have been produced. Naturally there is a larger number of false matches still active, and the planes are not as distinctly shaped as in the previous example, but they can clearly be distinguished (particularly in fig. 27 that shows the maximally activated node within each column of the CDS).

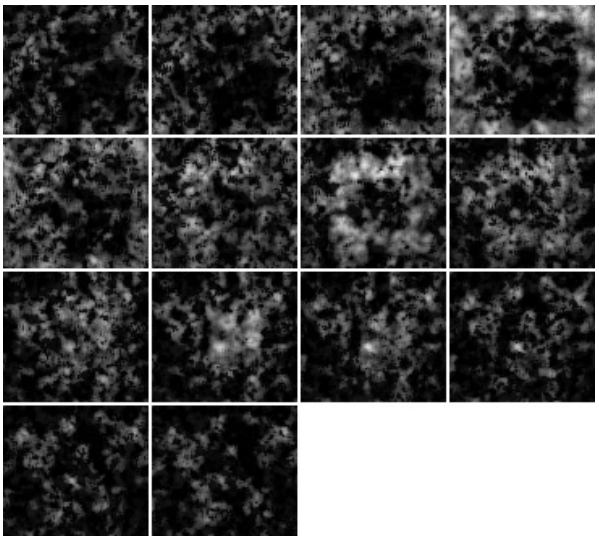

Figure 25. Initial matching.

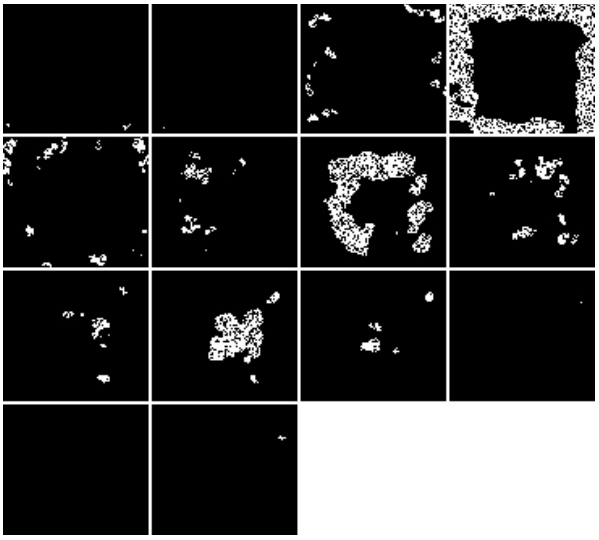

Figure 26. Activity after 7 iterations (left image mapped on top).

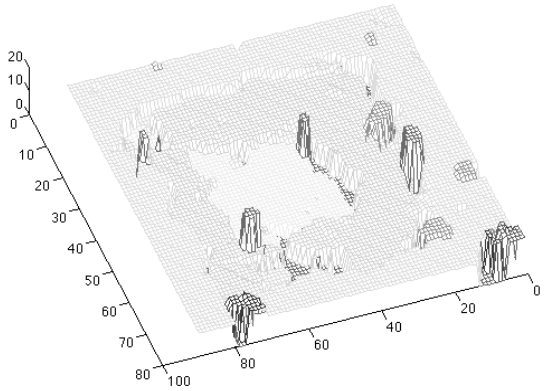

Figure 27. Active nodes after 7 iterations. Although the three planes are somewhat distorted, due to remaining false matches, they are still clearly distinguishable.

7.4 Trial 4

The input in this and the following trial consists of stereograms of natural images, and are simply intended to demonstrate how the model performs with "natural" input. In this example, the stereogram in figure 28 will be fused (it shows a picture of the author, with some bookshelves and a window in the background). Apart from the earlier presentations, here the results from some of the intermediate iterations will be displayed as well. This is to show how the activity within the CDS gradually changes and eventually becomes stabilised.

Figure 28. The author.

Figure 29 shows the activity directly after the initial matching procedure. As can be seen there is a large amount of activity at almost every level of the CDS. Clearly most of these nodes have been incorrectly activated.

After the first iteration (fig. 30) a large amount of falsely activated nodes have been extinguished and the surfaces of the face and background have become (relatively) stronger activated. As the process continues, for each successive iteration (fig. 31) there are less false matches present and the correctly matched surfaces grows more strongly activated. After the 7th iteration (fig. 32 and 33), only a few false matches remain and most of the active regions have been correctly matched. For readers capable of fusing the stereogram above (figure 28) this is easily verified.

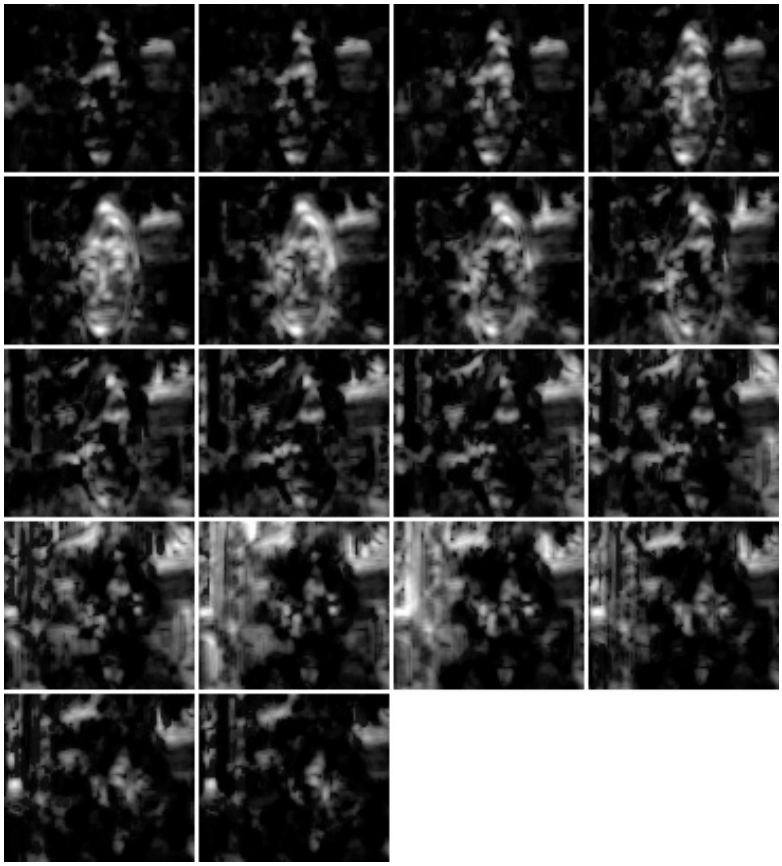

Figure 29. Initial matching result.

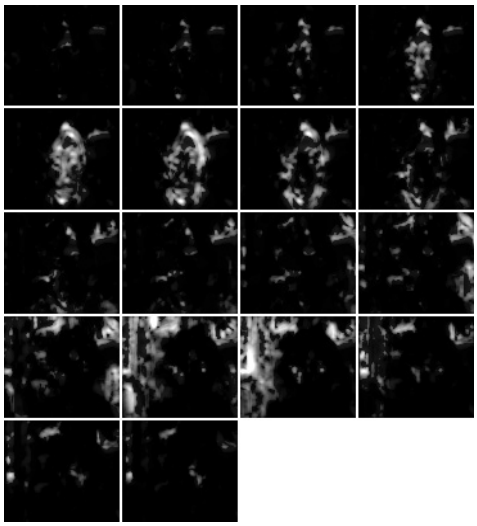

Figure 30. Activity after the first iteration.

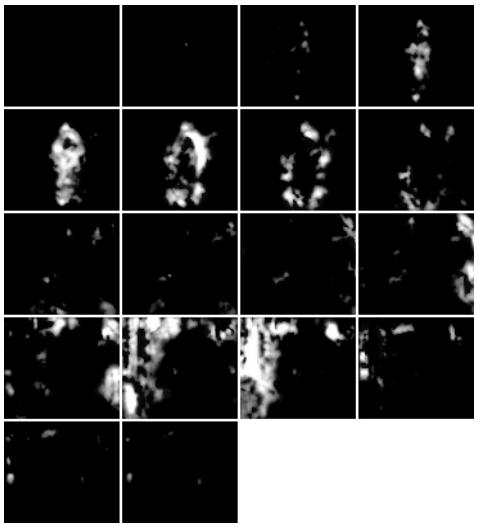

Figure 31. Third iteration.

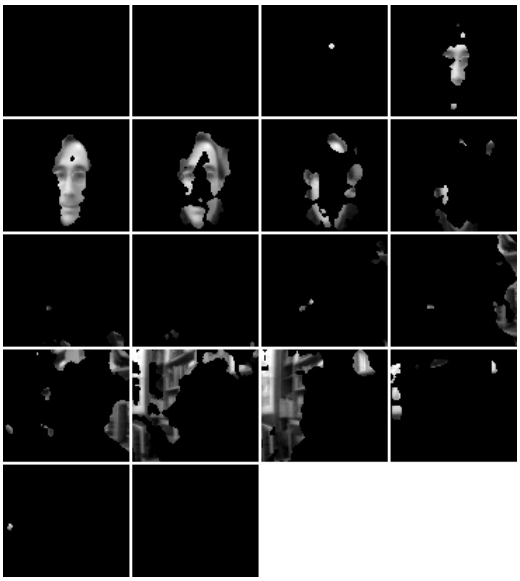

Figure 32. Activity after 7 iterations (with the original left image mapped on top of the most active regions).

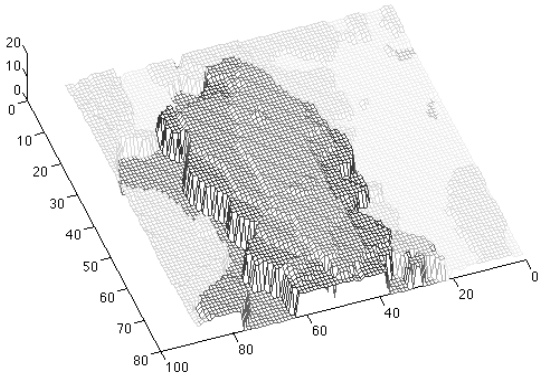

Figure 33. Disparity-map after 7 iterations.

7.5 Trial 5

The last stereogram (fig. 34) shows my tutor holding a white sheet of paper away from the camera. In the background there is a student, a round table and a supporting pillar (with increasing depth in that order).

Figure 34. Input (my supervisor).

Figure 35. Activity after the initial matching.

This last stereogram is the technically most complex one and therefore the most difficult for the model to fuse correctly. At a first glance it might not seem very different from the one in the previous trial, but at a closer look there are a

few things about the motif that causes problems for the model. First of all, considering the higher resolution channels, there are several relatively large regions where no contrast information can be detected (e.g. the inner part of the paper, the ceiling, my tutors shirt etc.). Another problem is that there, in several regions within the image, is relatively little horizontal disparity information. A majority of the edges in the scene are in fact horizontal, which can be seen from the results of the initial matching (fig. 35). As the results show the horizontal edges causes activity in almost every level of the CDS, and are therefore difficult for the model to extinguish.

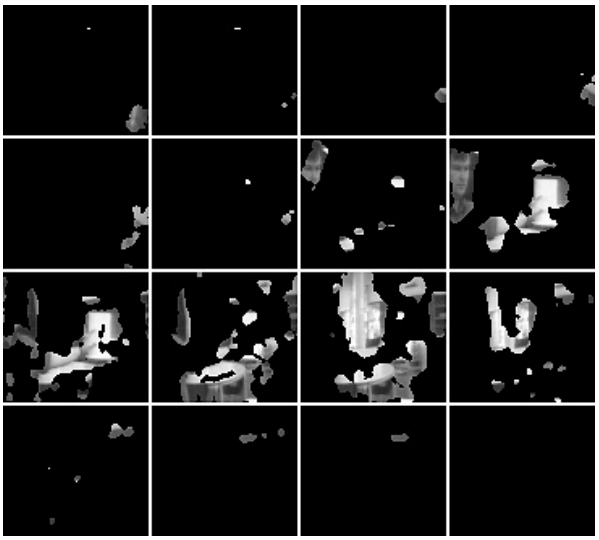

Figure 36. Active nodes after 7 iterations.

Due to these difficulties, the resulting disparity-map after seven iterations (fig. 36 and 37) is not as "clean" and continuous as in the previous examples, but it is still (on a rough scale) correct.

Before turning to the discussion I would like to point out a positive aspect in the results, which I have not yet mentioned. This positive aspect is the fact that the results (the stabilised activity within the CDS) are produced rather fast, i.e. the activity in the disparity-spaces are stabilised after only a few "iterations". I believe this "speed" could be considered as a strength of the model (as a model of the human stereopsis mechanism). The reason for this is that if the human stereopsis mechanism was a slow process, i.e. needed a long time to dissolve the false matches, there would seemingly have to be a delay in the experienced sen-

sation of depth, in comparison to what is monocularly seen, and such a delay does not seem to exist.

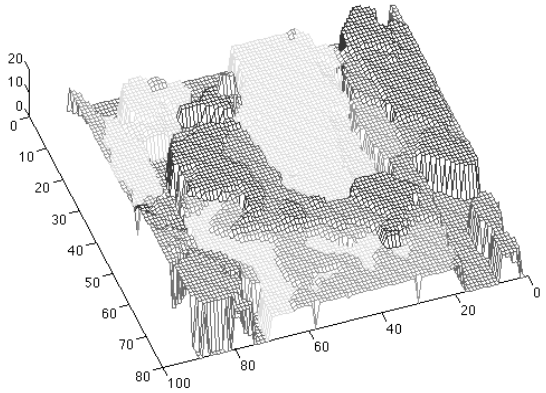

Figure 37. Disparity-map after 7 iterations.

8 Discussion

For natural reasons it is difficult to make any deeper analysis of how well the results presented above correspond to the "results", or output, of the human stereopsis mechanism. What makes this difficult is that there is yet no efficient way of simultaneously measuring the activity in a large number of cells in the human brain. Even if there were one would still have to know exactly where, in which region of the brain, the "result" was represented, and such precise knowledge of the anatomy of the brain still has to be found. Thus, the only way of analysing the results of the model is to compare them with the conscious perception of depth we experience when looking at the same pair of images as are feed into the model. What complicates this further is that our conscious perception of depth is a result of many contributing processes, which vary in their degree of cognitive complexity. Apart from the stereopsis mechanism, which can be considered as a relatively low level or early process, there are many higher cognitive functions involved in the interpretations of the various monocular cues (e.g. shading, perspective and size e.t.c.), which also affects the way we perceive depth. Even such high cognitive functions as expectations, reasoning and memory or knowledge about objects and the world affects the way we interpret the depth of a visual scene. Thus, even if the mechanism of stereopsis probably is the most im-

portant (for most types of visual scenes), the conscious perception of depth is still biased by all these other processes. For these reasons, it is difficult to draw any precise conclusions about the behaviour of the model and the following discussion will therefore be held at a quite general level.

Also important to realise, in order to make a fair judgement of the model, is that there are a number of cues available to the human stereopsis mechanism, that for practical reasons have not been possible to incorporate into the computer implementation of the model. Of particular interest are the information about the *convergence* of the eyes, the *accommodation* of the lenses and possibly even the information of *colour*.

Most likely, vergence movements (i.e. the smooth changes of the convergence of the eyes) have an important role in stereopsis. Human subjects rarely just stare at one point of visual scene, but instead we often make saccadic eye movements to bring in different parts of the image to the centre of our visual field. If these different parts of the image lie at different depths our eyes also initiate a vergence movement, so that the particular detail will fall on the centre part in both retinas. Thus, for the same visual scene several different representations of the depth can be constructed, which each and one is initiated from a different point of focus. Clearly this information could be very useful to the process of eliminating false matches. Since if these different representations are inconsistent for some part of the image an eye movement could be made to bring that particular part into focus, and thus make it possible to better establish the depth of that particular detail/region.

As explained earlier the accommodation of the lens can be a powerful cue to depth in combination with the visual input. In order to produce a sharp image on the retina, the lens has to be shaped differently, depending on the distance to the feature or surface of attention. The closer a surface is, the thicker the lens must be. Thus, by finding the optimal resolution of the image, of the surface of attention, the distance to the particular surface can indirectly be approximated from the information of the accommodation of the lens. It is quite obvious how this information could be used by the stereopsis mechanism to further restrict the domain of potential matches. Since the further a match are lying from the depth, estimated from the accommodation of the lens, the greater is the probability that it is a false match.

A final cue, or type of information, which possibly could be useful to the stereopsis mechanism is colour. Although the information of colour is not necessary, it clearly could be used to avoid at least some false matches, if the primitives to be matched were restricted to only those that showed similar colour compositions.

The main reason why the computer implementation has not been designed to take advantage of these cues is simply that the necessary “hardware” has not been available. However, provided that the necessary input could be feed into

the model, these cues (particularly the last two) could quite easily be incorporated into the model, with only minor changes to the implementation.

Finally, I would like to discuss some of the more general problems that one has to face when trying to model something as complicated as the human brain. Just as a chain is no stronger than its weakest link, the accuracy of any model is determined by the accuracy of how its smallest building blocks are modelled. In the case of modelling the brain, or part of it, the smallest building blocks are neurones. Now, the problems one has to face when trying to simulate the behaviour of neurones on a computer are mostly of practical nature but nevertheless quite complicated.

One such problem is how to simulate the continuous and parallel exchange of information between cells, on a computer that can only perform one operation at a time. The only way to model such continuous processes on computers is to split time into a number of discrete intervals and then, within each interval, compute an approximation of the behaviour of the processes over that particular time. Thus, just as when calculating the integral of a function, the accuracy of the resulting approximation will depend on the number of intervals. Desirably, the process would be divided into an infinite number of intervals. Unfortunately, this is where the problem arises since the processing time needed to compute the approximation for an interval is constant. Thus, the total time required to approximate the process grows very rapidly with the number of intervals. In practice this simply means that in order to receive the results of the process within a reasonable amount of time, one can not divide the process into too many intervals. This, in turn, means that the approximations of the processes often will be quite rough, which under poor circumstances can cause the whole model to behave strangely.

Another practical problem (closely related with the one above) with simulating neurological systems on computers is how to realistically model, with limited computer resources, the behaviour of the individual cells within the system. The problem is that such systems are often built up by a very large number of cells, and therefore, in order to save computer resources, the individual modelling of these cells often has to be quite crude. This is very unfortunate since neurones are far from being just on/off-devices. The response of a neurone is often not just determined by the current degree of incoming activation from neighbouring cells, but its response is also determined by its earlier activation history. Thus, could any particular neuron's threshold potential, firing and decay rate, vary from time to time. My point here is that without modelling the individual behaviour of the cells, in such systems, in a considerably more elaborate way than is done in most models (including the one presented in this paper), it is difficult to simulate many of the more dynamic properties of such systems. I also believe that some phenomena that usually are ascribed to processes or systems at higher levels, better could be accounted for by such lower level, “within-neur-

one" processes. As an example of such a phenomenon consider *hysterisis*. In the context of stereopsis *hysterisis* refers to the phenomenon that once the depth of a visual scene has been perceived (or stabilised), it is hard to break it up even if the images are slightly distorted or separated horizontally. Marr (1982) has commented on hysterisis as follows: "... It therefore seems unlikely that hysterisis is a consequence of the matching process, and much more likely that it is due to a cortical memory that stores the result of the matching process but is distinct from it". I believe this is a good example of a "high level" explanation of hysterisis in the sense that an entire, and separate, memory structure has to be introduced, in order to account for the phenomenon. As I see it such a high level explanation of hysterisis is not necessary. If one considers the neurones in the brain that would correspond to the nodes in the combined disparity-space (of the model presented in this paper), or possibly the neurones at the next higher level were the absolute depth is represented. It is possible to imagine how hysterisis could be accounted for at a "lower" (cellular) level by considering how these cells could be adapted to be less recipient to change and/or have a relatively sustained response profile, in order to bridge the gap between changing inputs.

I would like to emphasise that this example should not, at first hand, be seen as an attempt to explain the phenomenon of hysterisis, but merely to point out the possibility that some of the phenomena, displayed by the human stereopsis system, better could be accounted for by processes at a lower, cellular, level.

Considering the various problems described above, I believe there is no shortcut to building a "truly realistic" model of human stereopsis. I am convinced that many of the properties of human stereopsis only can be reconstructed if the behaviour of the fundamental building blocks, i.e. the neurones, are modelled so that the more dynamic aspects of their behaviour can be simulated. And to do this efficiently the problem of simulating continuous processes on computers must be solved. This might just be a matter of waiting for computers that are faster and have larger memories, but it might also mean that an entirely new form of hardware has to be used. A type of hardware better adapted to handle continuous and parallel processes.

9 Summary

I have in this paper tried to show how the correspondence problem could be solved more efficiently by a direct comparison of contrast values, within different spatial frequencies, rather than by the comparison of some set of more symbolic, or "predefined", features (e.g. bars, edges, blobs e.t.c.). I have also suggested how groups of simple and complex cells, with common receptive fields, possibly could represent the configuration of contrast within their receptive fields, and thus pointed to the possibility that such a strategy might be used by the human stereopsis mechanism. Unfortunately, the computer implementation of the

suggested model was, for practical reasons (mainly due to limitations in computer resources), not designed to support this latter assumption, but merely designed to show that the correspondence problem can be satisfactorily solved by comparing the "raw" contrast information within a stereogram.

A natural future improvement to the computer implementation, that better could support the assumption about the simple and complex cells, would therefore be to replace the current initial matching procedure with a procedure where the individual responses of the simple and complex cells, within the suggested groups, were more explicitly modelled.

Considering the later processing levels of the implementation, I also believe that the combination of the disparity-information, from the different channels, could be modelled in a more sophisticated way. A problem with just multiplying the disparity-values together is that if there is only contrast within the higher frequencies, even correctly matched regions, within the finer channels, could be suppressed by the lack of activity within the coarser channels. A possible solution to this problem could be to let the activity in the coarser channels exclusively amplify the activity in the finer channels. However, without having specified exactly what the result of this processing step should be, it is difficult to come up with a clear and general idea of what computations should be performed. Considering the human visual system it is not unlikely that our attention could shift between these channels or at least have the effect of making one, or several, of these more dominant than the rest. Clearly, this would affect the result of the cross-channels combination and also make it very hard to establish a general rule for how this combination should be performed.

Despite these shortcomings, the computer implementation performs quite satisfactory for both natural and artificially produced stereograms, and in several aspects the performance also shows signs of being consistent with the performance of the human stereopsis mechanism. For example: 1) the model seems to be quite robust, i.e. it is not very sensitive to distortions such as uncorrelated "noise" or slight vertical shifts in the relative positions of the two images, 2) it is relatively fast, only a few iterations are required to stabilise the activity in the disparity-spaces, 3) the combination of disparity-information from three different channels makes it possible to rapidly established the correct match even if several false matches are present within the finer channels.

However, even though these results are encouraging, computer implementations such as this one are still rather primitive, and can only model some of the most fundamental aspects of the human stereopsis mechanism. In order to construct a more "complete" model of this system, that better could account for some of the more dynamic properties of the human stereopsis mechanism (such as hysterisis and the establishment of depth by vergence movements), I believe it is necessary to more explicitly model the individual behaviour of the cells within such a system. Without a correct model of the dynamic behaviour, at the cellular

level, it is hard to see how such a model, realistically, could simulate the dynamic behaviour at a macro-level. Unfortunately, such an explicit model would require far more computer power than is commonly available today, but if the development of computers continue at the same rate as in the past, it will hopefully not be too long before such a model will see the light of day.

Acknowledgements

I would like to thank Christian Balkenius for having given valuable comments on earlier versions of this paper, and for the time he has spent helping me to arrange the results (into a somewhat viewable form).

References

Barlow, H. B., Blakemore, C., and Pettigrew, J. D. (1967). "The neural mechanism of binocular depth discrimination". *J. Physiol*. (Lond). 193, 327–342.

Felton, B., Richards, W. and Smith, A. Jr. (1972). "Disparity processing of spatial frequencies in man". *J. Physiol*. 225: 319–62.

Hubel, D. H. and Wiesel, T. N. (1959). "Receptive fields of single neurons in the cat's striate cortex". *J. Physiol*. (Lond.) 148: 574–591.

Hubel, D. H. (1988). *Eye, Brain and Vision.* Scientific American Library.

Julesz, B. (1960). "Binocular depth perception of computer-generated patterns." *Bell Systems Techn. J* 39, pp 1125–1162.

Julesz, B. (1971). *Foundations of cyclopean perception*. Chicago: Univ. Chicago Press.

Julesz, B. and Hill, M. (1978). "Global stereopsis: Cooperative phenomena in stereoscopic depth perception". *Handbook of sensory physiology*: v. 8, Held, R. (ed.), 7:pp236. Springer-Verlag. Berlin.

Kulikowski, J. J. (1978). "Limit of single vision in stereopsis depends on contour sharpness". *Nature*, 275: 126–27.

Levinson, E. and Blake, R. (1979). "Stereopsis by harmonic analysis". *Vision Res*. 19: 73–78.

Marr, D and Poggio, T. (1976). "Cooperative computation of stereo disparity". *Science*. 194: 283–87.

Marr, D and Poggio, T. (1979). "A computational theory of human stereo vision". *Proc. R. Soc. London Ser*. B 204:301–28.

Marr, D. and Hildreth, E. (1980). "Theory of edge detection". *Proc. R. Soc*. Lond. B 207, 187–217.

Marr, D. (1982). *Vision.* W. H. Freeman and Company.

Mayhew, J. E. W. and Frisby, J. P. (1981). "Psychophysical and computational studies towards a theory of human stereopsis". *Artificial Intelligence* 17: 349–385. North-Holland Publishing Company.

Poggio, G. F. and Poggio, T. (1984) "The analysis of stereopsis". *Ann. Rev. Neurosci*. (7): pp 392, 393–395, 400.

Pollard, S. B., Mayhew, J. E. W. and Frisby, J. P. (1985). "PMF: A stereo correspondance algorithm using a disparity gradient limit". *Perception*, vol. 14, pp 449–470. Pion Publication.

PAPER II

Occluding Contours

A Computational Model of Suppressive Mechanisms in Human Contour Perception

Abstract — A fundamental problem in vision is how to identify the occluding contours of objects and surfaces, given the ambiguity inherent in low-level visual input. A computational model is proposed for how occluding contours could be identified by making use of simple *heuristics* that reduce the ambiguity of individual features. In the striate cortex, a large majority of cells are selective for both contrast and orientation; i.e., they respond preferentially to simple features like contrast edges or lines. The heuristics we propose enhance or suppress the outputs of model striate-cortical cells, depending on the orientation and spatial distribution of stimuli present outside of the "classical" receptive field of these cells. In particular, the output of a cell is suppressed if the cell responds to a feature embedded in a texture, in which the "component features" are oriented in accordance with the orientation-selectivity of the cell. The model has been implemented and tested on natural as well as artificial grey-scale images. The model produces results that in several aspects are consistent with human contour/form perception. For example, it reproduces a number of known visual phenomena such as *illusory contours*, *contour masking*, pre-attentive *pop-out* (due to orientation-contrast), and it enhances contours that human observers often report perceiving as more salient.

1 Introduction

This paper addresses various questions related to human form perception, with a particular emphasis on how occluding contours might be processed in the visual cortex. An occluding contour can technically be defined as a contour that marks a discontinuity in depth (Marr, 1982). That is, if traced back to its source in the physical world, an occluding contour corresponds to the line on a surface where the view-line touches both the object and the background; or more formally where the view-line is tangent to the surface (fig. 1). Also considered here are contours that arise due to sharp changes in the orientation, or slant, of a surface;

such as along edges/ridges. Although not always occluding, these contours are similar in that they define, or mark, abrupt changes in depth.

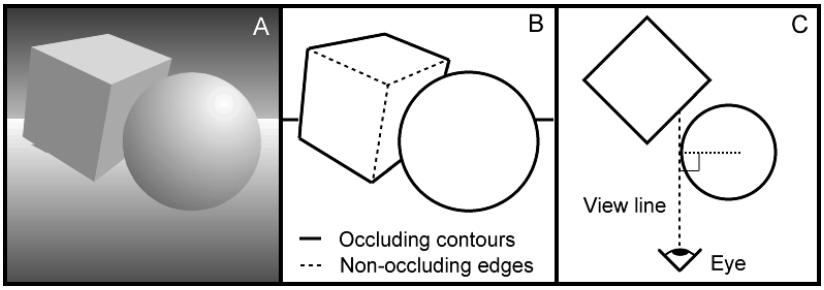

Figure 1. (A) A cube, partly occluded, by a sphere. (B) Occluding "outlines" (whole lines) and non-occluding edges (dashed). (C) Along an occluding contour, the view-line is orthogonal to the normal of the (occluding) surface.

Occluding contours are interesting entities of early vision for a very simple reason: they mediate fundamentally important information about the 3-D structure of the physical environment. If accurately identified they can provide information on the position, orientation and extension of object and surface boundaries. This information in turn is crucial for a number of our visual abilities such as determining foreground-background relationships, segmenting the visual input into meaningful entities (objects), and recognising objects from shape, etc. Given how dependent we are on these abilities for solving even the simplest task, it is evident that accurate identification of occluding contours is an important key to effective and reliable visual scene analysis. Surely, the ability to identify (and the capability to use the information on) occluding contours gave our early ancestors an advantage over species who did not have it. And surely this ability will be an important component in the perceptual system of any artificial agent that is to interact with a "real world" environment.

In general, there is some kind of visual contrast along a depth-discontinuity; e.g. a contrast in luminance or colour, a difference in motion or binocular-disparity, or a discontinuity of pattern. Hence, the human visual system could potentially use a variety of different cues to identify occluding contours. Given that binocular-disparity and (relative) motion are particularly powerful cues to depth, our visual system most likely relies heavily on such information when available. However, even in the absence of such direct cues to depth (-discontinuities), we are often remarkably good at identifying occluding boundaries in a visual scene. Given, for example, a black and white photograph, we can usually

rapidly identify the occluding contours of objects and surfaces, even if the scene or the objects in it have not been encountered before.

An interesting aspect of this ability is that we are usually not aware of the underlying process, or the computational difficulties that this process deals with. This suggests that the neural mechanism responsible for the identification of occluding contours mainly is a pre-attentive one, i.e. that it operates relatively autonomously from conscious influence. A growing body of neurophysiological, psychophysical and anatomical studies, described below, supports this view.

The main question this paper addresses is what operations these early low-level mechanisms perform on the visual input in order to produce useful representations of occluding contours; i.e. useful in the sense of assisting movement in, and manipulation of, the physical environment. In this paper, the discussion of possible mechanisms will be limited to visual input that is monocular, static and monochromatic. Of particular interest is the effect the local visual surround has in modulating how we perceive occluding contours. More precisely, how the arrangement and orientation of various low-level contrast features (e.g. edge and line segments) in the nearby surround could determine whether we perceive a visual structure as an occluding contour, or as a part of a surface texture/pattern.

2 Computational Considerations

In a static monochromatic (2-D) image, the only information that may reveal an occluding contour is the presence of some kind of luminance contrast along the contour; e.g. a contrast edge, a line or a pattern discontinuity. However, in images of natural scenes, contrast information may not only be found along the occluding boundaries of surfaces, but may also be found in the surfaces themselves, due to, for example, textures, patterns, shadow-lines and reflections (fig. 2). Hence, a major problem with identifying occluding contours is to discriminate contrast features that are caused by occluding contours from features that are produced by other physical structures and phenomena.

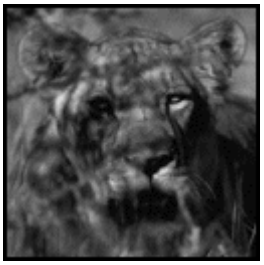

Figure 2. Surfaces often contain a variety of different contrast markings; due to for example texture (changes) and shadows.

Several factors make this discrimination difficult. First, the type of visual trace that exists along an occluding contour often changes from one point to another. That is, the type of feature, or local luminance pattern, that visually defines different parts of a contour, may change from, for example, a pattern discontinuity at one point, to a contrast edge or a line at some other point (fig. 3a). Such changes are caused by a wide variety of factors, such as variations along a contour in the texturing or reflection properties of a surface; or changes in the orientation of a surface, causing different amounts of light to be reflected into the eye. In other situations, the only physical evidence of an occluding contour may be a small disruption in a texture density, or a misalignment of the component features in a texture (fig 3b). From a computational perspective, this large variation in the type of feature/visual trace that can define a contour poses a delicate discrimination-problem to any visual system. That is, any useful discrimination strategy must not only "tolerate", or generalise over, many different features that may be present along an occluding contour, but must also be sensitive to features that are not caused by occluding structures, and discard the latter.

Further, there may be no visual trace at all along parts of an occluding contour. Such situations arise when, for example, there is no difference in the reflected luminance from (between) the occluded and the occluding surfaces, and there are no visible surface markings (see the middle left section of the sphere in fig. 3a). In such situations, the problem of identifying occluding contours is not so much a matter discrimination, but rather one of "filling-in", or reconstruction.

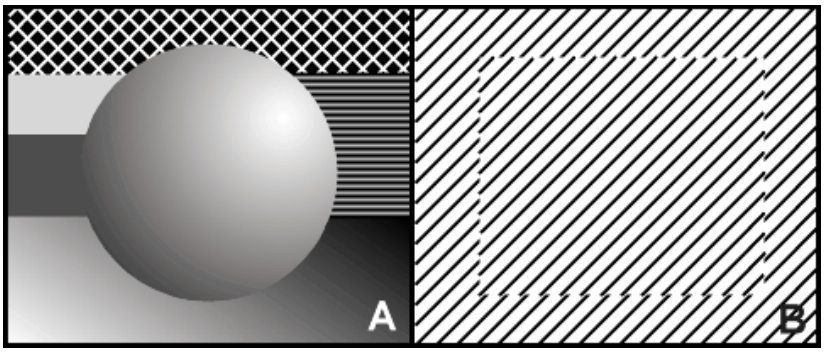

Figure 3. (A) Following the contour of the sphere around, the type (and polarity) of the contrast markings changes from point to point. (B) Rectangle defined only by a small pattern discontinuity.

Finally, problems may arise due to the loss of explicit depth information that occurs when a visual scene is projected onto a 2-D surface. Because the 3-D structure of a scene is compressed in the retinal projection, some contour (parts)

may, for example, end up closer to contours that are caused by other objects than to contours originating from the same object. For the same reason, contour parts originating from different objects may also overlap each other in the retinal image. Hence, even if the individual parts of a number of contours have been correctly identified and/or filled-in, it may still be difficult to bind, or integrate, these parts appropriately into meaningful structures (objects etc.).

Considering these computational difficulties, it is remarkable how easily and rapidly we are able to pick out the occluding contours in a 2-D image, and how biased we are to perceive these structures as coherent entities, even if the visual information along them has a complex composition and/or is partly missing. Before turning to the question of how the human visual system handles these perceptual difficulties, it is first appropriate (in order to pose the proper questions) to briefly consider how visual input is represented at the cortical level.

3 Representation of Visual Input in the Striate-Cortex

In the primary visual cortex (area V1) a large majority of cells are highly sensitive to visual stimuli that contain some oriented contrast. Depending on their response-properties to basic visual stimuli/features, these cells can be divided into three broad categories: simple, complex or hypercomplex (or end-stopped) cells (Hubel & Wiesel, 1962; Hubel, 1988). The receptive field of a typical simple cell is divided into two, three or more alternately excitatory and inhibitory sub regions, arranged in parallel bands along a common axis of orientation. Due to this receptive-field mapping, these cells respond strongly to stimuli such as contrast edges or lines of a particular orientation and polarity (i.e. contrast direction). Complex cells have slightly larger receptive fields and are not sensitive to the exact positioning of a stimulus/feature within their receptive field, but otherwise respond to similar stimuli as simple cells. The third category of cells, the hypercomplex or end-stopped cells, are also sensitive to oriented contrast patterns, but differ in one major aspect from the simple and complex cells. As the name suggests, the end-stopped cells only respond to features that terminate within their receptive fields (e.g. line- endings, corners). If the stimulus extends over their whole receptive field, the response is weakened or totally suppressed.

Another interesting but more global feature of the cortical organisation is that the topography of the retinal image, in general, is preserved in the striate-cortical representation (Kandel, 1991). That is, stimuli that are close together in the retinal image, will in general be represented by cells in area V1 that are situated near each other in the cortical tissue.

4 Inherent Ambiguities

How does our visual system identify occluding contours given 1) the computational difficulties discussed above, 2) the response properties of the cells in area V1, and 3) the fact that the retinal image is retinotopichally represented over the cortical surface? Because the receptive fields of the various simple and complex cells are relatively small compared to the whole visual field, it is evident that no individual cell can represent the presence of an occluding contour that spans a larger region of the visual field. Consequently, our visual system must, at some level of processing, integrate the responses from a potentially large number of simple/complex cells that may be firing along a contour line. However, in order for this integration mechanism to produce meaningful results, it should avoid integrating the responses from cells that fire due to causes other than occluding contours. That is, it should avoid integrating the responses from cells that fire to features caused by, for example, surface textures or shadow lines etc. But this situation creates somewhat of a paradox. Before it has identified an occluding contour, how can our visual system "know" which cells fire due to occluding contours, and which cells fire due to other causes? The problem is that, in general, the reason why any individual cell fires can not be unambiguously established by only considering the type of stimuli a cell is sensitive to. Consider, for example, a simple cell that responds optimally when a contrast edge is present within its receptive field. This cell will fire with equal strength whether the contrast edge is caused by an occluding contour, a shadow line, a reflection or some detail in a texture pattern. How then could this ambiguity be resolved?

One conceivable solution to this problem would be that some central higher-level process, which could integrate information from all over the visual field, simply tried out every possible combination of grouping the responses from the cells in V1 into contours, and then somehow determine the solution that seemed most plausible. This could involve comparing the results to stored representations of objects and scenes, consulting higher-level knowledge and experiences, and considering the context in which the stimuli was perceived.

Occasionally, such high-level processing might be needed to resolve certain perceptual ambiguities, but in general the perceptual process seems to be much faster and less accessible for conscious manipulation than such a scheme would suggest. Nor would it explain how we are able to identify occluding contours of unknown, or partly hidden, objects in unfamiliar contexts where no high-level knowledge or experience is relevant.

Moreover, leaving the disambiguation of low-level stimuli to such a late stage of processing leads to a combinatorial explosion in the number of ways there are to combine the responses from the cells in area V1 into different contour paths, even when the visual input is modestly complex. In other words, in its pure form the above scheme does not seem to account for how we identify occluding con-

tours, but instead suggests that some of the response-ambiguity must be resolved at a much earlier stage, before or at the level where spatial integration takes place.

5 Neurophysiology and Psychophysics

A growing body of empirical evidence supports the view that important aspects of contour, or form, processing is carried out at a relatively early stage in the visual pathway. For example, Peterhans and von der Heydt (1993) described "contour neurones" in area V2 that respond not only to contrast edges or lines of a particular orientation, but also to pattern discontinuities (i.e. when the discontinuity is orientated in accordance with the cells orientation selectivity) and even to broken edges and lines (i.e. illusory contours). The fact that these cells seem to respond to discontinuities, invariant to the exact composition of the luminance pattern within their receptive field, strongly suggests that these cells are important for identifying occluding contours.

However, a number of other recent studies have shown that some form-related processing which could serve to facilitate the identification of occluding contours may be done as early as in area V1. Single cell recordings (Gilbert & Wiesel, 1990; Knierim & van Essen, 1992; Kapadia et al., 1995), and real-time optical imaging (Grinvald et al., 1994), have demonstrated that the firing rate of individual cells in area V1 is not exclusively determined by the stimulus present within a cells receptive field, but can be modulated (i.e. enhanced or suppressed) by stimuli located outside the receptive field.

More specifically, Kapadia et al. (1995) have shown that the firing rate of an individual complex cell, which in isolation responds to a bar of a certain orientation, can be enhanced if one or several other similarly oriented bars are positioned along the cell's axis of orientation, but outside its receptive field. They further showed that the enhancement effect decreased as the bars were i) separated along the common axis of orientation, ii) separated from co-linearity, or iii) separated in orientation (fig 4).

A related but suppressive effect has also been reported by Knierim and van Essen (1992), who have demonstrated that the firing rate of an individual cell can be significantly reduced if a number of bars that are oriented similarly to the preferred orientation of the cell are placed outside the receptive field (fig. 5). If the surrounding bars are oriented orthogonal to the central bar, the suppressive effect is reduced but still present.

Further, Kapadia et al. (1995) demonstrated that the suppression observed in a cell's response when a large number of randomly oriented bars are placed outside its receptive field can be considerably reduced, or even eliminated, if some of the surrounding bars are positioned along the cell's axis of orientation and are oriented in the same direction as the central bar (fig. 6).

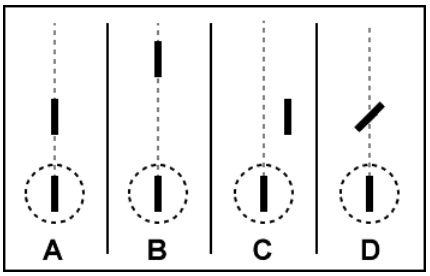

Figure 4. According to the study of Kapadia et al. (1995), the response of a cell was (A) enhanced when a bar, co-aligned with the cell's axis of orientation selectivity, was placed outside its receptive field (dashed circle). Further, the enhancement decreased if the bars were (B) separated along the common axis of orientation, (C) separated from co-linearity or (D) separated in orientation.

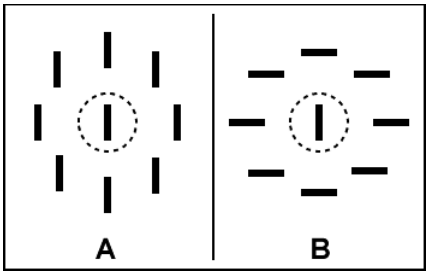

Figure 5. The response of a cell is suppressed more when (A) surrounded by similarly oriented features, than when (B) surronded by differently oriented ones (Knierim & van Essen, 1992).

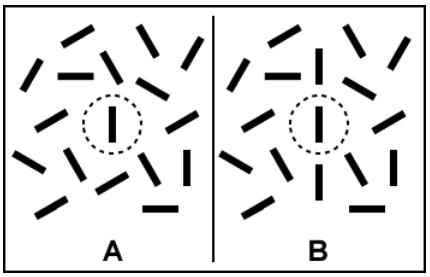

Figure 6. Kapadia et al. (1995) also observed a reduction in the suppression (caused by randomly oriented bars; A) in a cells response, if some of the bars were co-aligned with the orientation selectivity of the cell (B).

Similar findings have been reported in a number of psychophysical studies. The contrast detection threshold of a central low-contrast Gabor-patch can be increased or decreased depending on the position and orientation of surrounding Gabor-patches (Polat & Sagi, 1993, 1994) or gratings (Cannon & Fullenkamp, 1991). It has also been shown that a path of Gabor-patches, presented against a background of evenly distributed and randomly oriented patches, can be more easily detected when the relative angle between the adjacent elements in the path is less than +/-60° (Field et al., 1993), or the elements form a closed rather than open path (Kovacs & Julesz, 1993).

An interesting parallel, in this context, is how closely several of the above findings coincide with the Gestalt laws (Wertheimer, 1923; see also Rock & Palmer 1990) that were formulated to account for how we group low-level stimuli. Of particular note are the laws which postulate that we are perceptually biased to group together features that are arranged into smooth paths (*good continuation*), form closed curves (*closure*), and are close to each other (*proximity*); see fig. 7.

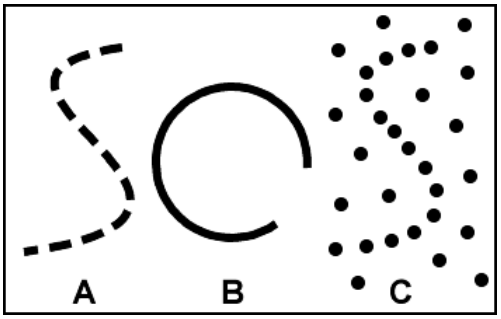

Figure 7. Illustration of the Gestalt (grouping) laws of (A) *good continuation*, (B) *closure* and (C) *proximity*.

It is not yet clear whether the effects (described above) arise within the striate cortex, or are produced by feedback connections from higher visual areas. Kapadia et al. (1995) have suggested that the long-range horizontal connections formed by pyramidal cells in the striate cortex could constitute the physiological substrate allowing spatial integration of information over several hypercolumns. These long-range connections enable the target cells to integrate information over regions well beyond the classical receptive field, but preferentially from cells having similar orientation tuning that are positioned along the target cells axis of orientation. However this may be, feedback connections from area V2 and other visual areas can not, of course, be ruled out. Nor can it be ruled out

that different mechanisms may be responsible for different modulatory effects. Knierim and van Essen (1992) reported a time delay between the onset of the general (orientation independent) [~7 ms] suppressive effect, and the orientation-dependent [~18-20 ms] suppressive effect, which may indicate different origins.

6 Possible Functional Significance

Kapadia et al. (1995) have suggested that the purpose of the selective enhancement in the firing rate of certain cells may be to make contours more salient, particularly when perceived against noisy and textured backgrounds. Given that the enhancement effect seems to be stronger for stimuli-configurations that consist of smoothly aligned features, and that occluding contours in general tend to produce such constellations in the retinal image, this interpretation seems highly plausible. The idea is appealing also because it is consistent with the computationally recognised need for mechanisms that can reduce the response-ambiguity of the simple/complex cells at an early stage of visual processing. Another interesting aspect of this interpretation is that, if correct, it might not only provide an explanation as to why we experience the Gestalt laws of *good continuation, closure and proximity* (i.e. to aid the identification of occluding contours), but it may also place the origin of these phenomena at a much earlier stage of visual processing than previously thought.

Regarding the suppressive effect, Knierim and van Essen (1992) have suggested that the observed difference in the suppression of a cell's response depending on the difference in orientation between the central and the surrounding stimuli may be important for texture segregation; and that it may be responsible for certain orientation-dependent pop-out phenomena such as our ability to quickly spot a single "V" embedded in a 2-D array of "T´s"; see for example Treisman and Gelade (1980). While basically agreeing that the suppression could be involved in both texture segregation and pop-out, a slightly different interpretation is here made on what the main functional significance of the suppression is. That is, we rather emphasise the possibility that the primary purpose of the orientation-dependent suppression -like possibly the corresponding enhancement effect- may be to aid the identification of occluding contours.

From the earlier discussion on the combinatorial explosion in the number of possible ways there are of grouping the responses from the simple/complex cells into contour paths, it is evident that our visual system somehow must constrain the grouping process. The observed enhancement in certain cells firing rates could be seen as such a constraint, as a way to guide higher-level integration processes. Letting a simple *heuristic* which prefers smoothly aligned features control the enhancement, seems a reasonable first approach to narrowing down the number of potential visual structures that may correspond to occluding contours.

However, because any heuristic by definition occasionally will be wrong, there needs to be an opposing, or complementary, mechanism that can balance or even override the effect of the enhancement. In many visual contexts, selective enhancement of co-aligned stimuli will not be a helpful strategy for identifying important occluding boundaries. Consider, for example, the fact that most surfaces in nature are heavily textured (e.g. fur, feathers, grass, leaves, rocks) and may produce regions with periodic or quasi-periodic patterns in the retinal image. Often, the components of such patterns consists of locally smoothly aligned features. Because such stimuli "fit the description" of co-alignment they would inappropriately be integrated into contours, unless some opposing system could counteract, or suppress, the integration mechanism.

Consider also a visual scene such as, a hungry lion lurking behind some high but possible-to-see-through grass; in such a context, a visual system would not serve its owner well if it enhanced every single straw of grass, but not the partially hidden outline of the lion. Clearly, not all occluding contours are equally important to us, but some deserves more attention than others. Preferably those that mark the peripheral boundaries of regions, objects and surfaces.

For these reasons, it seems that a more reliable representation of occluding boundaries, less "polluted" with nonsense contours, would be obtained if the integration of low-level stimuli into contours was suppressed within densely textured regions of the visual field; particularly if the features within such regions are periodically or quasi-periodically arranged, and are oriented in accordance with the axis along which the contour-integration is carried out.

Apart from computational considerations, ecological speculations, and the earlier reviewed physiological and psychophysical observations of suppressive effects, there is a rather compelling phenomenon referred to as *contour masking* (Kanizsa, 1979) which indicates that such a suppressive mechanism may control contour-integration in the human visual system. When the rectangle in figure 8a is viewed on its own, the vertical lines are clearly perceived as contours of the rectangle. However, when embedded into a texture such as in figure 8b, the vertical lines are no longer perceived as contours, but rather appear as if they are parts of a surface that seems to lie in front of the rectangle. What is perhaps the most interesting aspect of this phenomenon is that the experienced difference between the two viewing conditions seems to be entirely qualitative. That is, in fig 8b there seems to be no quantitative reduction in the perceived contrast of the lines, at least not large enough to cause the contours to vanish, but only a reduction in our inclination to perceive them as contours. This suggests that a higher-level representation of contours, or the process that integrates low-level stimuli into contours, is suppressed rather than the early representation of low-level stimuli per se. If this is the case, then the observed orientation-dependent suppression of V1 cells may be due to feedback connection from these higher visual areas where the contours are integrated/suppressed. This idea is consistent with

the observed time delay between the onset of the general suppression and the orientation-dependent suppression in area V1 cells reported by Knierim and van Essen (1992).

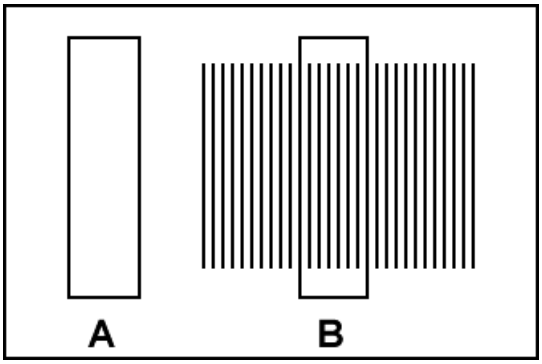

Figure 8. Example of contour "masking" (Modified after Kanizsa, 1979).

In the next section, a computational model based on these ideas is presented in which a layer of model *contour neurones* integrate oriented low-level stimuli according to a simple heuristic of co-alignment. To prevent "non-occluding" stimuli from being integrated into contours (i.e., being represented as contours), the model contour cells are suppressed depending on the magnitude and orientation of the stimuli in the near surrounds of their receptive fields. Although not intended as a quantitative description of the human visual system, the model is nevertheless consistent with several of the above described properties of both cell responses and psychophysical observations. Simulations with a computer implementation of the model do, for example, produce contour completion (e.g. illusory contours; Kanizsa, 1979), contour enhancement (i.e. increased saliency; Kapadia et al., 1995; Field et al., 1993), contour masking (Kanizsa, 1979) and orientation dependent pop-out (Treisman and Gelade, 1980), and it identifies occluding boundaries in natural images.

7 A Computational Model

The model presented below is first and foremost a hypothetical functional model of how contours might be processed in the early stages of the human visual pathway. However, although *function* has been the main constraint, most design choices in the model architecture have been influenced by known response properties of various cell types and their inter-connections.

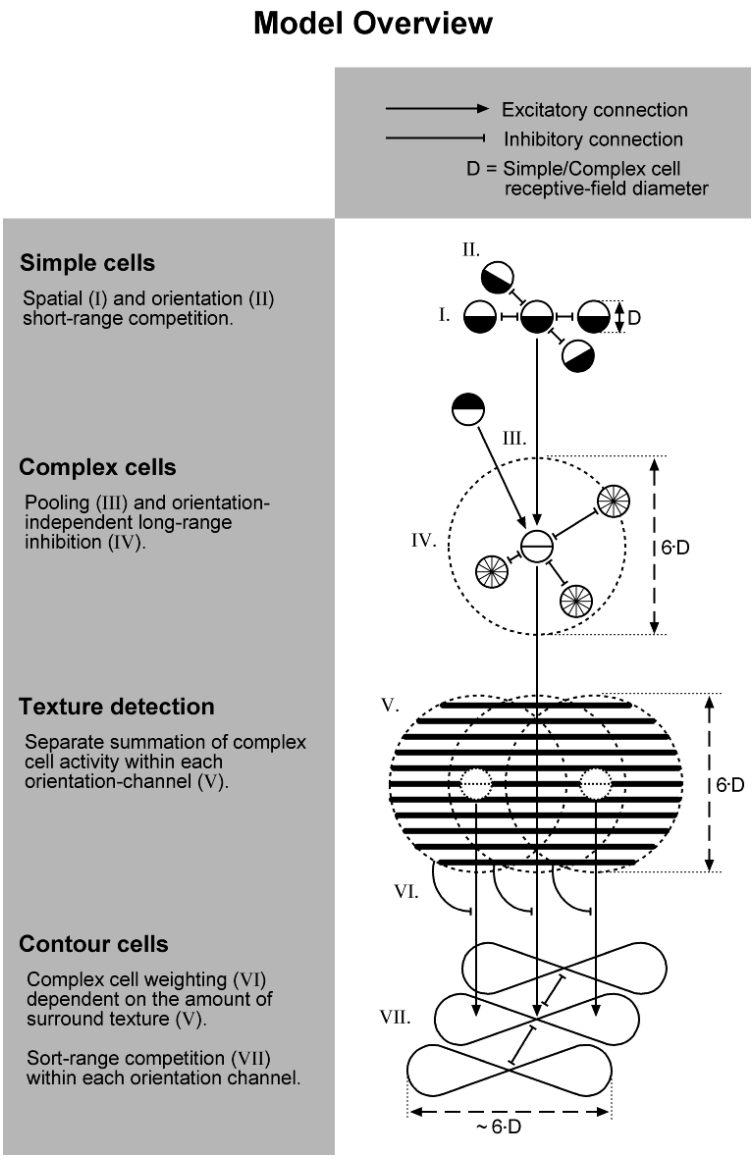

Figure 9. Model overview. (I) Spatial and (II) orientation short-range competition competition between similarly tuned simple cells. (III) Pooling of simple cell responses (by the complex cells). (IV) Mutual Long-range complex cell suppression (orientation independent). (V) "Texture" detection. (VI) Complex cell output weightied inversely proportional to the amount of texture-surround. (VII) Spatial integration of the complex cells' outputs along the common axis of orientation; and short-range (contour cell) competition.

Figure 9 gives an overview of the model. On a coarse scale the model can be divided into two major levels of processing, roughly corresponding to the processing carried out by the simple and complex cells (Hubel & Wiesel, 1962) in area V1, and by the contour neurones/cells (Peterhans and von der Heydt 1993) in area V2.

At the first level, oriented contrast features in an input image are detected by a layer of model simple cells with anti-symmetric receptive fields (fig. 10a). To increase the spatial and orientation selectivity of these cells, all nearby simple cells (i.e. near in both the spatial and orientation domain) laterally inhibit each other. The responses from the simple cells are then fed into a layer of complex cells. Any given model complex cell pools the information from two simple cells that are separated by π rad (180°) in orientation tuning, and that are positioned at the same location in the visual/image field. Like the model simple cells, the model complex cells mutually suppress one another. However, the complex cells do so over a much larger distance than the simple cells (approximately 6 compared to 1 times the radius of a cell's receptive field) and they do so independently of orientation selectivity.

At the second level, a layer of model contour cells sum the outputs from the level 1 complex cells. The contour cells are also orientation selective, and any given cell only sums the outputs from complex cells with a particular orientation tuning. The receptive fields of these cells are (approximately 6 times) larger than the simple/complex cells, and are divided into two drop-shaped sub-receptive fields (fig 11). Only when there is sufficient activity from complex cells within both of a contour cells two half-fields does it become activated. To prevent stimuli within densely textured regions from being integrated into contours, all complex cells that respond to such stimuli are given a lesser weighting than those that respond to stimuli not embedded into textures. Finally, all nearby model contour cells with the same orientation selectivity inhibit each other along an axis orthogonal to the axis of their orientation selectivity.

In the following sub-sections, a more thorough presentation of the various processing steps and their functional motivation is given. For technical and implementation details, the reader is directed to appendix A.

7.1 Level 1

7.1.1 Model Simple Cells

The receptive fields of the simple cells are modelled with anti-symmetric Gabor-functions (i.e. the product of a sine and a Gaussian function). Symmetrical and anti-symmetrical elementary Gabor-signals (fig. 10) have been shown to correspond well with the receptive field-mappings of real simple cells (Marcelja, 1980). In order to capture contrast stimuli of different polarity and at all different orientations, 12 model cells each differing $\pi/6$ in orientation from the next are

used to sample the image structure at each given position. The receptive field response is then half-rectified (i.e. negative values are ignored) and normalised for contrast by a divisive gain mechanism.

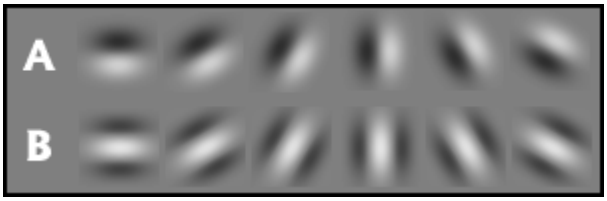

Figure 10. (A) Anti-symmetrical Gabor-filters used to detect oriented contrast features in the input image. (B) Symmetrical Gabor-filters. Not drawn to scale.

While cells with symmetrical and anti-symmetrical receptive fields respond optimally to different stimuli (i.e. a bar and edge respectively), each type of cell also responds to the optimal stimuli of the other type, although less so and at a slightly shifted position. From a computational point of view, using either kind of receptive field mapping is therefore sufficient to detect oriented contrast stimuli. In the computer implementation of the model, only one type of mapping was chosen to hold the computational cost down. The choice of anti-symmetrical ones was arbitrary, except for the observation that some kind of contour completion phenomena seems to be stronger when the inducers are solid edges rather than thin lines (Kanizsa, 1979). However, in the human and primate brain, both cell types most likely contribute to the processing of form.

7.1.2 Lateral Inhibition

Because the model simple cells have partially overlapping receptive fields and because they are quite broadly tuned to orientation, any given stimulus will evoke activity in a number of cells nearby in both the spatial and orientation domain. Hence, the representation of an image will initially be somewhat blurred. In order to obtain a higher spatial and orientation acuity in the array of simple cells all near cells (i.e. near in either the spatial or the orientation domain, or both) laterally inhibit each other. Apart from sharpening the spatial and orientation selectivity of the model cells, this operation also has the effect of creating a relative activity enhancement in cells that respond to line-ends and corners, compared to those that respond to the interior parts of such stimuli (this mechanism is similar to the "end-cut" mechanism of Grossberg & Mingolla, 1985). In general, these "end-points" are the ones of interest for a contour completion mechanism (see also von der Heydt, 1995).

7.1.3 Model Complex Cells

While complex cells, like simple cells, are sensitive to stimuli having a particular orientation, many complex cells, unlike simple cells, fire independently of the polarity of a contrast stimulus (Livingstone et al., 1987). The intuitive observation that we easily can complete and integrate fragments differing in contrast into contours (see fig. 3a) suggests that the complex cells rather than the simple cells provide the main input to the neural mechanism responsible for form analysis.

In the current model, the responses of the complex cells are obtained by simply taking the absolute value of the difference between each two simple cells that share the same position and are separated in orientation selectivity by π rad (for a more sophisticated model of complex cell responses, see for example Heeger, 1991).

7.1.4 Orientation-Independent Long-Range Suppression

Apart from what is present within its classical, or primary, receptive field, the response of a model complex cell is also determined by the degree of general activity within a larger region surrounding its receptive field. More precisely, the activity of the cell is suppressed proportionally to the squared and weighted sum over all orientations of the complex cell activity within a Gaussian envelope of approximately 6 times the radius of the complex cell.

The functional motivation for this suppression is two-fold. First, it further enhances the activity of cells that are responding to edge- and line-ends, which are important for identifying texture borders and points were contours should be completed, or filled-in. Second, it creates an initial relative difference in the strength of activity in cells that respond to stimuli positioned at the periphery of textured, or otherwise crowded, regions, compared to those cells that are positioned at the interior of a texture-field. In general, such peripheral stimuli are statistically more likely to correspond to parts of surface/object borders.

Long-range suppression of a striate complex cells, induced by stimuli positioned outside the classical receptive field, have been observed in several studies (Knierim & van Essen, 1992; Kapadia et al., 1995; Grinvald et al., 1994).

7.2 Level 2

7.2.1 Model Contour Cells

At the second level of processing, the suppressed outputs from the complex cells are integrated by a layer of model contour cells. Like the model simple and complex cells, the contour cells are selective to stimuli of a particular orientation. The contour cells, however, have considerably larger receptive fields, which allow them to integrate information from several complex cells along their axis of orientation. Another important feature of the model contour cells is that they are

heavily suppressed by stimuli within textured regions, if the stimuli making up the texture are oriented similarly to the axis of orientation to which the cell is tuned.

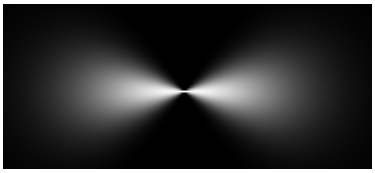

Figure 11. Receptive field of a model contour cell. Not drawn to scale.

7.2.2 Spatial Integration and Texture Suppression

The receptive field of each model contour cell is divided into two drop-shaped half-fields (fig. 11). Each half-field "hangs" down, along the axis of orientation-selectivity, from the centre of the receptive field, and reaches out to a distance of about 6 times that of the radius of the (primary) receptive field of a model simple/complex cell. Further, each sub-field separately sums the weighted outputs from all complex cells within its range that are selective to the same orientation as the contour cell. How the output from any particular complex cell is weighted is determined by two factors. First, the response is weighted by a factor that is determined by the spatial respectively angular distance from the contour cell's centre, respectively, axis of orientation. The effect of this weighting is that only relatively co-aligned stimuli will become integrated. Second, the response of a complex cell is also weighted by an iso-orientation-measure, τ^{θ}_{iso} ,(see Appendix A for details) that is inversely proportional to the degree of activity of all other complex cells tuned to the same orientation within a larger region around the complex cell (6 times the diameter of a model complex cell's primary receptive field). In other words, a model complex cell that responds to an isolated stimulus will be more heavily weighted than one that responds to a stimulus surrounded by other similarly oriented stimuli. Further, to prevent the contour-cells from becoming active at points in the image where there are no contours to fill-in or complete such as outside of corners and line-terminators, sufficient activity in both sub-receptive-fields is needed to make it respond. The contribution from each half-field is therefore integrated in a multiplicative fashion. The result of the integration is then passed through a threshold-function (an inverted Gaussian) that particularly compresses the higher response-interval, but also reduces the amount of noise in the lowest response-interval.

7.2.3 Lateral Inhibition

Because the contour-integration is performed on a relatively coarse scale (i.e. with relatively wide receptive fields), the positioning of the boundaries within the resulting representation will not be precise. Therefore, in order to better locate the spatial positions of the boundaries, all contour cells that are sensitive to the same orientation and lie near each other along an axis orthogonal to their axis of orientation-selectivity inhibit each others output.

8 Simulation Results

The model presented in the previous section has been implemented as a computer program, and simulations have been run with images of both artificial and natural scenes. For all simulations presented here, the model parameters were set as described in appendix A, and all input-images were 128×128 pixels.

Apart from producing results that are consistent with humanly observed phenomena such as illusory contours (Kanizsa, 1979) and the Gestalt laws of *good continuity*, *proximity* and *closure* (Wertheimer, 1923; Rock & Palmer 1990), the model also reproduces various contour masking (Kanizsa, 1979) and orientation-dependent pop-out (Treisman and Gelade, 1980) phenomena. Further, some capacity for texture segregation has been observed, provided that the major components of the textures differ in their orientation by more than approximately 60°, or the textures have significantly different periodicities (densities).

In each of the examples below, the input image is depicted to the left, the initial model simple cell response in the middle and the model output to the right. High intensity in the middle and rightmost images corresponds to high activity in the model simple and model contour cells respectively. The intensity value at each point in these representations was obtained by pooling the activity in all orientation-channels (see appendix A, section 4). The simple cell representation is shown only for comparison. All images presented below are also available at: *www.lucs.lu.se/people/jens.mansson/contours/index.html*

8.1 Artificial Images

8.1.1 Contour Masking and Pop-Out

Figure 12-14 shows examples of the contour masking effect, caused by the suppression of the contour-integration mechanism within regions containing densely positioned parallel lines, or other iso-oriented stimuli. Note also that in all of these three artificial images, illusory contours are formed at the ends of the lines, and the contour cell activity at these points is significantly higher (i.e. the illusory contours are more salient) than the activity along some of the actual intensity lines.

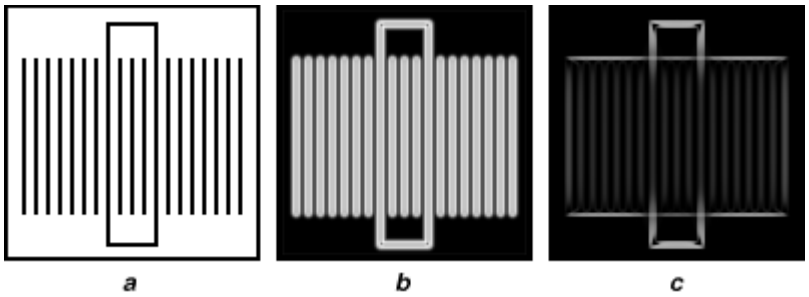

Figure 12. Example of contour "masking". Redrawn from Kanizsa (1979). a) Input image. b) Initial "simple" cell activity. c) Output, i.e. "contour" cell activity.

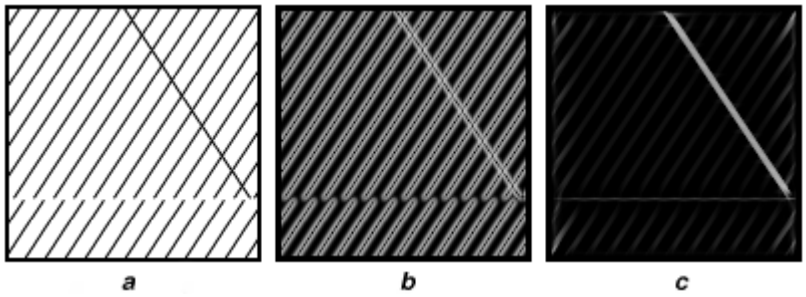

Figure 13. Partly "masked" and partly "illusory" triangle. Modified from von der Heydt (1995) who modified it from Galli and Zama (1931).

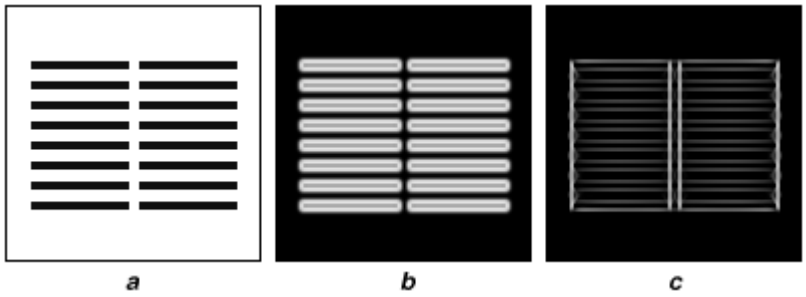

Figure 14. An illusory white bar in front of parallel horizontal lines.

The suppressive mechanism that produces the above masking effect also makes a single oriented feature relatively more enhanced than the features in a surrounding array, if these are differently oriented (see figure 15). This could explain why attention is drawn to such parts of an image, and why the search for such stimuli is considerably faster than for stimuli that differ less, or in more than one of several possible aspects (orientation, colour, motion etc.) compared to surrounding features (see for example Treisman and Gelade, 1980).

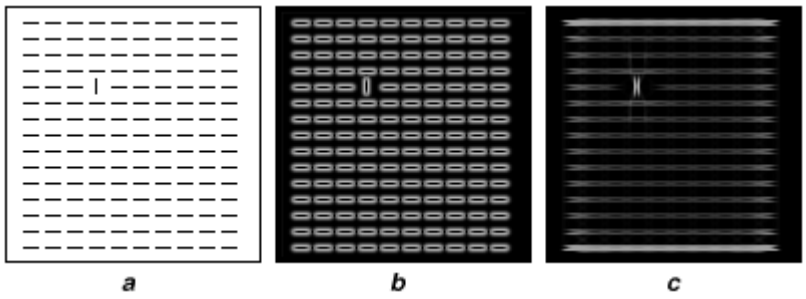

Figure 15. "Pop-out" of a single oriented line in an array of orthogonaly oriented lines. a) Attention is automatically drawn to the vertical bar. c) the output activity is stronger at the position of the vertical bar.

8.1.2 Illusory Contours

As figure 16 and 17 show the model produces both straight and smoothly curved illusory contours at positions were human observers generally report perceiving these. These results are produced because a model contour cell integrates the information from the orientation-selective complex cells over relatively large region of the image; and hence can be activated even if there is no stimulus at the centre of its receptive field.

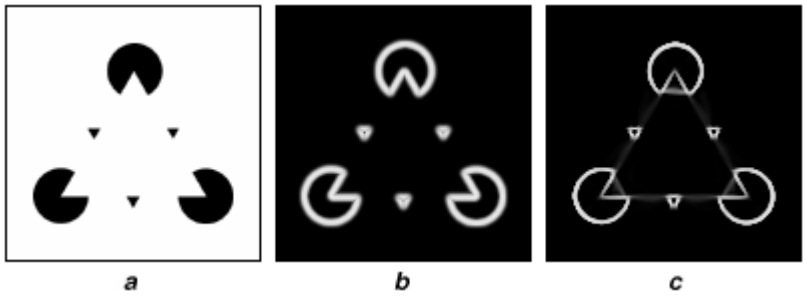

Figure 16. Kanizsa triangle. Modified from Kanizsa (1979). Note the illusory lines that have been formed between the black discs.

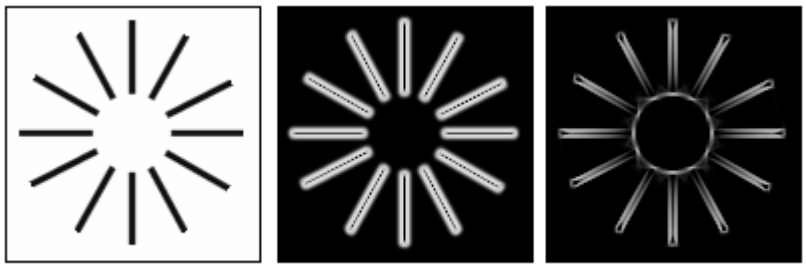

Figure 17. Illusory white disc covering the black radial lines.

8.2 Natural Images

The remaining examples demonstrate the model's performance on natural input. Figure 18-21 show how periodic textures are suppressed while leaving the majority of "real", or object, contours intact. Particularly note how the horizontal lines in figure 19c become relatively enhanced in the output representation, even though they are barely present in the initial simple cell representation (19b); and in figure 20, note how most of the contours of the fruit are left intact while the background table cloth pattern is suppressed. Also note in figure 22 and 23 how not only periodic iso-oriented textures are suppressed, but also random textures, or otherwise busy regions if there is sufficient activity in all orientation channels to drive the suppression mechanism.

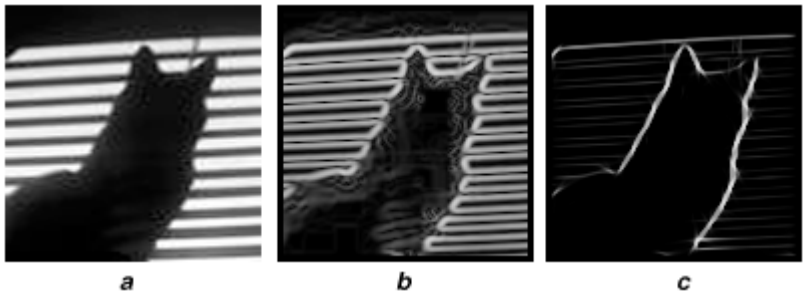

Figure 18. Shadow on a wall of a cat. The contour of the cat is both filled-in and enhanced, while the horizontal lines, on the wall in the background, are suppressed.

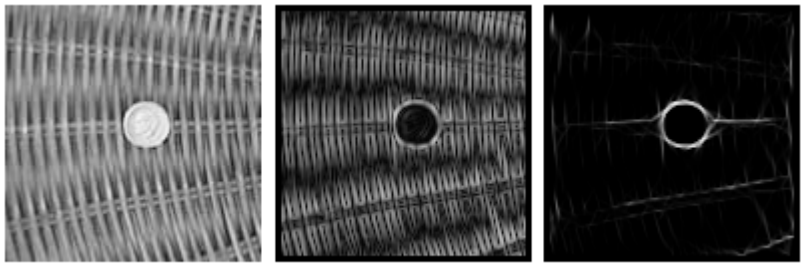

Figure 19. Coin on a table. Note how the outline of the coin, and some of the horizontal lines, which barely are present in the simple cell representation, are strongly enhanced in the output.

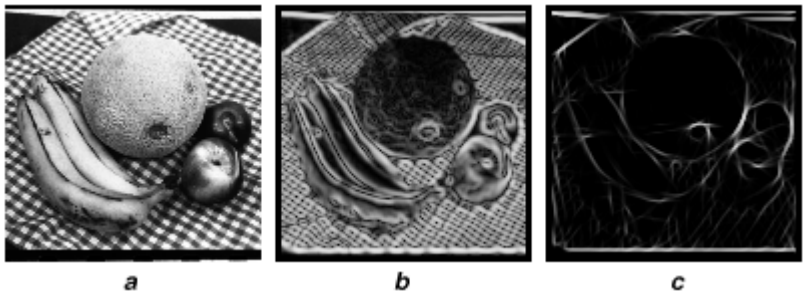

Figure 20. Compared to the simple cell representation (b), the contour representation (c) is much sparser, and almost entirely confined along the occluding contours of the fruit.

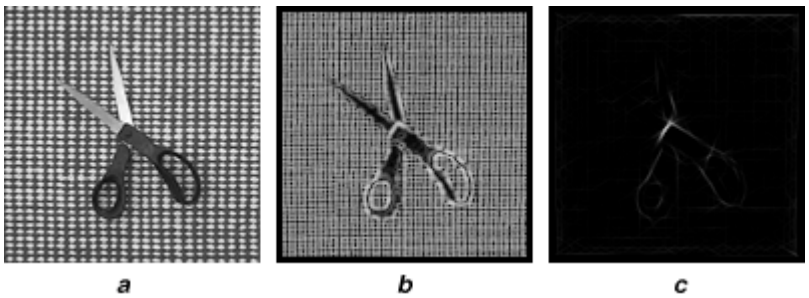

Figure 21. Pop-out of a pair of scissors on a carpet.

Figure 22. A lion resting in the shadow of a tree. In regions where there is model simple/complex cell activity in all different orientation channels (e.g. a lot of noise), most stimuli are suppressed.

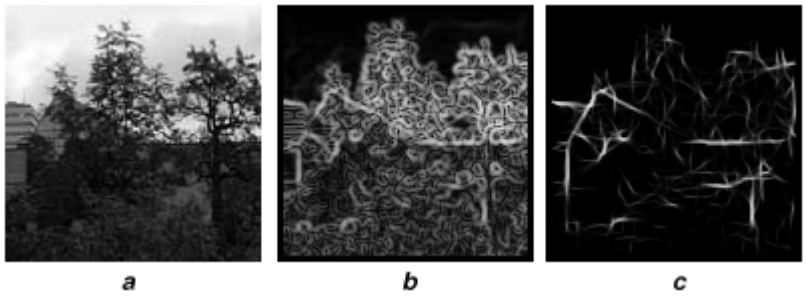

Figure 23. The skyline of a building behind some trees. Stimuli within crowded regions, such as the tree tops and bushes in front of the house, are suppressed in the output representation.

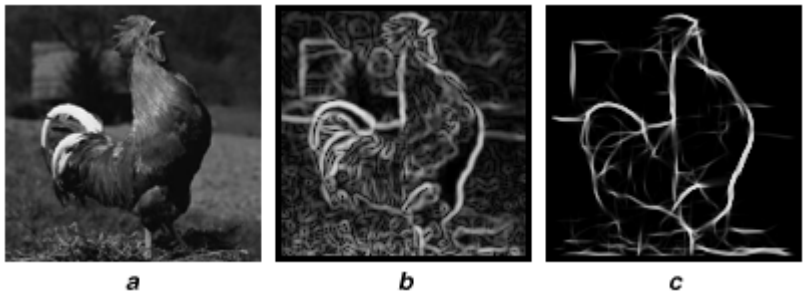

Figure 24. A rooster. Note how the majority of "false" contours in (b), caused by the feathers and grass, have been reduced in the final output representation (c).

9 Discussion

9.1 Related Work

The model presented in this paper shares several features with the models on contour perception suggested by Ullman (1976), Grossberg and Mingolla (1985), Gove et al. (1995), Heitger and von der Heydt (1993) and Yen and Finkel (1998). Although these models differ in various assumptions, for example in the proposed contour inducing elements, they all share the assumption that occluding contours, in general, produce relatively smoothly aligned features in an image. Hence, all the models locally constrain the spatial integration to features that are similarly oriented and relatively co-aligned. Due to this common feature, most of the models produce results that are more or less consistent with each other, and could account for why we perceive illusory contours and why we experience the Gestalt grouping laws of *good continuity*, *closure* and *proximity*.

However, in none of these models is the integration of low-level stimuli modulated by the contextual information available in the local surroundings, in the sense described in this paper. Therefore it seems unlikely that any of these models can account for phenomena like contour masking and orientation-dependent pop-out, considering that these phenomena seem to be highly context dependent. Further, given that textures often produce locally co-aligned visual stimuli, it is likely that these models will be relatively poor at discriminating such stimuli from actual occluding contours.

9.2 Texture Discrimination

Because the model presented here is only intended as a functional model of how contour information might be processed early on in the human visual pathway, the individual processing steps described in the model can be only loosely mapped onto particular neurological structures. A particularly loose mapping is the one between the proposed contour-suppression-mechanism and a possible neural substrate that could implement it. In the current model, "textures" are crudely sensed by simply integrating the responses from a large number of model complex cells. From a computational point of view, this is most likely not the best procedure for detecting and discriminating between textures. An interesting question that therefore arises is what neural substrates other than the complex cells could provide information about texture.

One type, or category, of cells that seem particularly fit for this job are the "grating-cells" (von Heydt et al., 1992). These cells not only respond vigorously to gratings, but often fail altogether to respond to isolated bars or edges. Further, they are narrowly tuned to both orientation and spatial frequency, and have low contrast thresholds. According to von der Heydt et al. (1992), about 4% of all cell in area V1 and 1.6% of the cells in area V2 are of this type. An interesting property of these cells is that they not only respond to gratings of a particular

frequency and orientation, but also to a number of other periodic, or quasi-periodic, patterns such as checkerboard patterns (when the diagonal rows are aligned with the preferred orientation of the cell), or patterns with "jittered" periodicity (e.g. lines separated by alternately small and large distances). von der Heydt et al. concluded that these cells do not perform a spatial-frequency analysis of the stimulus, but instead seem to be specialised for detecting periodic patterns. Considering the narrow tuning for both orientation and spatial frequency, as well as the low contrast-detection threshold, it clearly seems these cells are better fit than complex cells for performing discriminative texture detection. And hence, could provide more detailed/sophisticated information to a contour suppression mechanism. Whether this is the case will of course have to be shown in empirical studies.

9.3 Possible Role of Spatial Frequency

A final consideration, not yet either discussed nor modelled is the possible role the spatial frequency of the stimuli have on our perception of contours. Intuitively, it seems that the phenomena of *contour masking* (Kanizsa, 1975) can be reduced, or even eliminated, if the lines of the rectangle in figure 8 are made considerably thicker than the row of parallel lines (see fig. 25). This suggests that not only the periodicity and orientation of stimuli in the surround control contour-integration in the human visual system, but that the spatial frequency of the stimuli also control it. The output of the complex cells are maybe more suppressed (or less weighted by an integration mechanism) when the spatial frequency of surrounding stimuli is similar to the frequency that to which a cell is tuned.

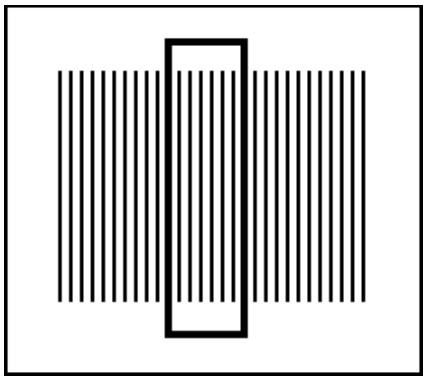

Figure 25. The "contour masking" effect is lost if the lines of the rectangle are made thicker .

10 Summary

A computational model is proposed for how information on occluding contours might be processed in the early cortical visual areas (roughly V1 and V1). A central subsystem in the model is a mechanism which suppresses the integration of oriented low-level stimuli into contours, if these stimuli are embedded into a texture composed of similarly oriented stimuli/features. This operation is motivated by the fact that features in natural scenes which are situated inside patterned regions are more likely (from a statistical point of view) to have been produced by surface textures, than they are likely to have arisen due to occluding structures. A computer implementation of the model demonstrates results consistent with the percepts that are reported by human observers. The model does, for example, fill-in missing segments of contours (i.e., produce illusory contours; Kanizsa, 1979) and enhances weak ones (i.e., increase the saliency; Field et al., 1993). Further it reproduces the phenomena of contour masking (Kanizsa, 1979) and certain orientation-dependent pop-out effects (Treisman and Gelade, 1980). It also works well on natural images where noise and, particularly, ambiguous stimuli may present problems to models that do not consider the contextual information available in the local surround.

Appendix – Technical Specification

A. Simple Cells

A.1 Receptive Fields

The receptive fields of the simple cells were modelled with 12 rotated copies each separated π/6 rad from the next [$\theta=n\pi/6$; $n=1...12$], of the following anti-symmetric Gabor-function:

$$G_{edge}(x,y) = \sin\left(2\pi f\left[x-x_c\right]\right)\cdot \mathrm{e}^{-\frac{1}{4}\left(\frac{x^2}{\sigma_{x1}^2}+\frac{y^2}{\sigma_{y1}^2}\right)} \qquad \text{(A.1.1)}$$

$$f=\frac{1}{2\,R}\ ,\quad \sigma_{x1}=\frac{R}{4}\ ,\quad \sigma_{y1}=\frac{R}{3.3}$$

where f is the frequency, R is the radius (3.5 pixels in the implementation) of the cells receptive field, σ_{x1} and σ_{y1} are space constants and x_c the centre of the receptive field.

A.2 Normalisation and Half-Rectification

The response from any given simple cell $S^{\theta}(x,y)$, at position (x,y), and tuned to orientation, θ, is obtained by convolving the raw image, I, with the correspond-

ing (Gabor) mask $G_{edge}{}^{\theta}$. The result is then normalised for contrast and half-rectified (denoted by floor brackets).

$$S^{\theta}(x,y) = \left\lfloor \frac{\sum_{s=x-R}^{x+R}\sum_{t=y-R}^{y+R} I(s,t)\cdot G^{\theta}_{edge}(s,t)}{\kappa + \sum_{s=x-R}^{x+R}\sum_{t=y-R}^{y+R} I(s,t)\cdot \left|G^{\theta}_{edge}(s,t)\right|} \right\rfloor \qquad \text{(A.2.1)}$$

$$\kappa = 0.02 R^2 I_{max}$$

Here, κ is a threshold constant, which is determined by R (same as above) and the maximum possible intensity value, I_{max}, in the image representation (e.g. 256 for an 8 bit grey-scale coding). The floor-brackets denotes half-rectification.

A.3 Lateral Inhibition

The inhibited output of a model simple cell, $S^{\theta}{}_{I}$, tuned to orientation θ and positioned at (x,y), is given by:

$$S^{\theta}_{I}(S^{\theta},x,y) = 1 - e^{-\frac{1}{2}\left[\frac{A(S^{\theta},x,y)^2}{\sigma_A^2} + \frac{B(S^{\theta},x,y)^2}{\sigma_B^2}\right]} \qquad \text{(A.3.1)}$$

The terms *A()* and *B()* provide the contribution from the orientation- and spatial-dependent inhibition respectively. σ_A and σ_B are saturation constants that control the contribution of *A()* and *B()* respectively (see below).

$$A(S^{\theta},x,y) = \sum_{n=-\Omega}^{\Omega} G_1(n\cdot\alpha,\sigma_o)\cdot\left[S^{\theta}(x,y) - S^{\theta+n\cdot\alpha}(x,y)\right] \qquad \text{(A.3.2)}$$

$$\sigma_A = 0.2 \sum_{n=-\Omega}^{\Omega} G_1(n\alpha,\sigma_0)\ ,\ n\neq 0\ \ ,\ \ \Omega = 3\ \ ,\ \ \alpha = \frac{\pi}{6}\ ,\ \ \sigma_0 = \frac{2\pi R}{3}$$

$$G_1(r,\sigma) = e^{-\frac{r^2}{\sigma^2}} \qquad \text{(A.3.3)}$$

The constant Ω determines the angular range of the inhibition, and α is the minimum angular separation between two differently tuned cells. σ_0 is a space constant that determines the shape of the Gaussian envelope provided by the

function $G_1(r,\sigma)$, which determine how much neighbouring cells contribute to the inhibition:

$$B(S^{\theta},x,y)=\sum_{i,j=-R}^{R} G_1\left(\sqrt{i^2+j^2},\sigma_s\right)\cdot\left(S^{\theta}(x,y)-S^{\theta}(x+i,y+j)\right) \tag{A.3.5}$$

$$\sigma_B=0.2\sum_{i=-R}^{R}\sum_{j=-R}^{R} G_1\left(\sqrt{i^2+j^2},\sigma_s\right)\ ,\ i\neq j\neq 0\ \ ,\ \ \sigma_s=\frac{R}{2}$$

Again, σ_s is a space constant that determines how fast the Gaussian envelope (eq. A.3.3) falls off.

B. Complex Cells

B.1 Pooling

The initial complex cell response, $C^{\theta}(x,y)$, for a cell tuned to orientation θ:

$$C^{\theta}(x,y)=\left|S_I^{\theta}(x,y)-S_I^{\theta+\pi}(x,y)\right| \tag{B.1.1}$$

B.2 Orientation-Independent Long-Range Suppression

The activity in a model complex cell after long-range suppression, $C^{\theta}_{LS}()$, is given by:

$$C^{\theta}_{LS}(C,x,y)=C^{\theta}(x,y)\cdot\left(1-\frac{1}{2}\Psi\left(\Phi(C,x,y),\ \sigma_C\right)\right)\ ,\ \ \sigma_c=16R \tag{B.2.1}$$

$$\Psi(x,\sigma)=1-e^{-\frac{x^2}{\sigma^2}} \tag{B.2.2}$$

$$\Phi(C,x,y)=\sum_{i,j=-W}^{W} G_2\left(\sqrt{i^2+j^2},\sigma_{C1},\sigma_{C2}\right)\cdot\sum_{n=0}^{5} C^{n\cdot\pi/6}(x+i,y+j)^2 \tag{B.2.3}$$

$$W=6R\ ,\ \sigma_{C1}=4R\ ,\ \sigma_{C2}=\frac{R}{1.2}$$

$$G_2(r,\sigma_1,\sigma_2)=e^{\frac{-r^2}{\sigma_1^2}}-e^{\frac{-r^2}{\sigma_2^2}} \tag{B.2.4}$$

Complex cells beyond the distance *W* do not contribute to the suppression. The rightmost summation in equation B.2.3 is summation over all orientations. The purpose of the squaring is to preferentially let high-contrast stimuli contribute to the suppression. The weight function, G_2, is the difference between two Gaussians with space constants σ_1 and σ_2. The latter creates an inner region approximately the size of the receptive field of a model complex cell, with near zero values so that a given cell does not suppress itself. σ_C is a saturation parameter for the threshold-function $\Psi()$ (eq. B.2.2).

C. Contour Cells

C.1 Sub-Receptive Fields

Depending on the distance, r, (respectively, the angular deviation, $\alpha-\theta$) from a contour cell's receptive-field centre (respectively axis of orientation), the two sub-receptive-fields ($\underset{\leftarrow}{F}^{\theta}$ and $\underset{\rightarrow}{F}^{\theta}$) of a contour cell (tuned to orientation θ) weights the outputs from all complex cells tuned to orientation θ according to (borrowed from Heitger and von der Heydt, 1993):

$$\underset{\rightarrow}{F}^{\theta}(r,\alpha,\theta,\sigma_f)=\cos^{2n}(\alpha-\theta)\cdot e^{-\frac{r^2}{2\sigma_f^2}} \qquad \text{(C.1.1)}$$

$$if\ -\pi/4 < \alpha-\theta < \pi/4\ \ else\ \ \underset{\rightarrow}{F}^{\theta}=0\ \ ;\quad n=4\ ,\quad \sigma_f=3R$$

$$\underset{\leftarrow}{F}^{\theta}(r,\alpha,\theta,\sigma)=\underset{\rightarrow}{F}^{\theta}(r,\alpha,\theta+\pi,\sigma)$$

C.2 Iso-Orientated Stimuli Density

The function $\tau_{iso}^{\theta}\left(C_{LS}^{\theta},x,y\right)$ is a measure of the amount of stimuli (with orientation θ) present within a region of radius *W,* centred at (x,y):

$$\tau_{iso}^{\theta}\left(C_{LS}^{\theta},x,y\right)=\Psi\left(\sum_{i,j=-W}^{W}G_2\left(\sqrt{i^2+j^2},\sigma_{w1},\sigma_{w2}\right)\cdot\Psi\left(C_{LS}^{\theta}(x+i,y+j),\sigma_{c2}\right),\sigma_{iso}\right),$$

$$if\ \sqrt{i^2+j^2}\le W\ \ else\ \ 0 \qquad \text{(C.2.1)}$$

$$\sigma_{iso}=0.15\cdot\sum_{i,j=-W}^{W}G_2\left(\sqrt{i^2+j^2},\sigma_{w1},\sigma_{w2}\right)\ ,\quad \sigma_{w1}=4R\ ,\quad \sigma_{w2}=\frac{R}{1.2}\ ,\quad \sigma_{c2}=0.15$$

The purpose of letting the the complex cell response, $C^{\theta}_{LS}(x,y)$, first pass through the (inner) threshold function, Ψ, (in eq. C.2.1) with the low saturation constant σ_{c2}, is to enhance weak responses and thereby emphasise the orientation of the stimuli and not the contrast-intensity. The function $G_2()$ (see e.q B2.4) with the space constants σ_{w1} and $\sigma_{2w,}$ determines how a stimuli at distance $\sqrt{i^2+j^2}$ from point (x,y) is weighted. σ_{c2} and σ_{iso} are saturation constants for the threshold-function Ψ (eq. C.2.1).

C.3 Spatial Integration and Texture Suppression

The summed activity, $\underset{\rightarrow}{K}^{\theta}$, within a contour cell's sub-receptive field, $\underset{\rightarrow}{F}^{\theta}$, is given by:

$$\underset{\rightarrow}{K}^{\theta}=\sum_{i,j=-L}^{L}\underset{\rightarrow}{F}^{\theta}\left(\sqrt{i^2+j^2},\arctan\left(\frac{j}{i}\right)-\theta,\sigma_f\right)\cdot C^{\theta}_{LS}(x+i,y+j)\cdot e^{-k\cdot\tau^{\theta}_{iso}(x+i,y+j)},$$

$$if\ -\pi/4<\arctan\left(\frac{j}{i}\right)-\theta<\pi/4\ \ else\ \underset{\rightarrow}{K}^{\theta}=0 \qquad \text{(C.3.1)}$$

$$L=5R\ \ ,\ \ k=2.2\ \ ,\ \ \sigma_f=3R$$

Note in eq. C.3.1 how both the sub-receptive field ($\underset{\rightarrow}{F}^{\theta}$) and the stimuli-density measure (τ^{θ}_{iso}) together determine how any given complex cell is weighted.

The combined responses, K^{θ}_{C}, from the half-fields $\underset{\leftarrow}{K}^{\theta}$ and $\underset{\rightarrow}{K}^{\theta}$ is:

$$K^{\theta}_{C}(x,y)=\Psi\left(\sqrt{\underset{\leftarrow}{K}^{\theta}(x,y)\cdot\underset{\rightarrow}{K}^{\theta}(x,y)},\ \sigma_K\right),\quad \sigma_K=\frac{1}{3} \qquad \text{(C.3.2)}$$

Ψ is the threshold-function (eq. B.2.2), and σ_K is a constant that determines how early the threshold-function saturates (i.e. reaches its maximum value).

C.4 Lateral Inhibition

The final contour representation K^{θ} for orientation θ is obtained by convolving the combined representation $K^{\theta}{}_{C}$ with a symmetric Gabor-function $G^{\theta}{}_{bar}(x,y)$ followed by half-rectification and filtering through the threshold-function Ψ (eq. B.2.2):

$$K^{\theta}(x,y)=\Psi\left(\left|\sum_{i=x-R}^{x+R}\sum_{j=y-R}^{y+R}K^{\theta}_{C}(i,j)\cdot G^{\theta}_{bar}(i,j)\right|,\ 0.1\cdot\sum_{i=x-R}^{x+R}\sum_{j=y-R}^{y+R}\left|G^{\theta}_{bar}(i,j)\right|\right) \qquad \text{(C.4.1)}$$

$$G_{bar}(x,y) = \cos\left(2\pi f_2\left[x - x_c\right]\right) \cdot \mathrm{e}^{-\left(\frac{x^2}{\sigma_{x2}^2} + \frac{y^2}{\sigma_{y2}^2}\right)} \qquad \text{(C.4.2)}$$

$$f_2 = \frac{3R}{4}\ , \quad \sigma_{x2} = \frac{R}{3.5}\ , \quad \sigma_{y2} = \frac{R}{3}$$

D. Simulation Output-Representation

All output images, I_O, presented in section 8 were obtained by pooling the activity in all 6 orientation channels [$\theta = n\pi/6$; n = 0 ... 5] as shown below. $\chi^{\theta}(x,y)$ is here, either, the initial simple cell activity, or the final contour cell activity, at image position (x,y).

$$I_O(\chi,x,y) = 1 - \mathrm{e}^{-\sum_{n=0}^{5} \chi^{n\pi/6}(x,y)} \qquad \text{(D.1.1)}$$

References

Cannon, M.W., Fullenkamp, S.C. 1991, Spatial interactions in apparent contrast: Inhibitory effects among grating patterns of different spatial frequencies, spatial positions and orientations. *Vision Research*, Vol. 31, No 11, pp 1985-1998

Field, D.J., Hayes, A., Hess, R. F. 1993, Contour integration by the human visual system: Evidence for a local "Association field". *Vision Research*, Vol. 33, No 2, pp 173-193

Galli, A., Zama, A., 1931, Beiträge zur Theorie der Wahrnehmung. *Z. Psychol*. 123:308-348

Gilbert, C., Wiesel, T. N., 1990, The influence of contextual stimuli on the orientation of cells in the primary visual cortex of the cat. *Vision Research*, Vol. 30, No 11, pp 1689-1701

Gove, A., Grossberg, S., Mingolla, E., 1995, Brightness perception, illusory contours, and corticogeniculate feedback. *Visual neuroscience*, 12, pp 1027-1052

Grinvald, A., Edmund, E., Frostig, R. D., Hildesheim, R. 1994, Cortical point-spread function and long-range lateral interactions revealed by real-time optical imaging of Macaque monkey primary visual cortex. *The Journal of Neuroscience*, May, 14(5): 2545-2568

Grossberg, S., Mingolla, E., 1985, Neural dynamics of form perception: Boundary completion, Illusory contours and Neon color spreading. *Psychological Review*, Vol. 92, No. 2, 173-211

Heeger, D.J., 1991, Nonlinear model of neural responses in cat visual cortex, *In Computational Models of Visual Processing*. eds. Landy M.S., Movshon J. A., MIT Press, Cambridge

Heitger, F., von der Heydt, R., 1993, A computational model of neural contour processing: Figure-ground segregation and illusory contours, In *Proceedings of the Fourth Int. Conf. On Computer Vision*. IEEE Computer Society Press

Hubel, D. H., Wiesel, T. N., 1962, Receptive fields, binocuar interaction and functional architecture in the cats visual cortex. *Journal of Physiology*, 160, pp 106-154.

Hubel, D. H. 1988, *Eye, brain and vision.* Scientific American Library.

Kandel, E. R., 1991, Perception of motion, depth, and form. In Kandel, E. R., Schwartz, J. H., Jessel, J. J. (eds.), 1991, *Principles of neural science*. Prentice-Hall International, London, UK.

Kanizsa, G. 1979, *Organization in vision: Essays on Gestalt perception*. Praeger Publishers, New York

Kapadia, M.K., Ito, M., Gilbert, C. D. 1995, Improvement in visual sensitivity by changes in local context: Parallel studies in human observers and V1 of alert monkeys. *Neuron*, Vol. 15, pp 843-856

Knierim, J.J., van Essen, D.C. 1992, Neuronal responses to static texture patterns in area V1 of the alert Macaque monkey. *Journal of Neurophysiology*, Vol. 67, No 4, April

Kovacs, I., Julesz, B. 1993, A closed curve is much more than an incomplete one. Effect of closure in figure-ground segmentation. *Proc. Natl. Acad. Sci*. USA, Vol. 90, p 7495-7497, August

Livingstone, M. S., Hubel, D. H., 1987, Psychophysical evidence for separate channels for the perception of form, color, movement, and depth. *Journal of Neuroscience*, 7 (11), pp 34163468

Marcelja, S., 1980, Mathematical description of the responses of simple cells. *J. Opt. Soc. Am*., Vol. 70, no. 11

Marr, D. 1982, *Vision*. W. H. Freeman and Company, New York

Peterhans, E., von der Heydt, R. 1993, Functional organization of area V2 in the alert Macaque. *European Journal of Neuroscience*, Vol. 5, pp. 509-524

Polat, U., Sagi, D. 1993, Lateral interactions between spatial channels: Suppression and facilitation revealed by lateral masking experiments. *Vision Research*, Vol. 33, No 7, pp 993-999

Polat, U., Sagi, D. 1994, The architecture of perceptual spatial interactions. *Vision Research*, Vol. 34, No 1, pp 73-78

Rock, I., Palmer, S., 1990, The legacy of Gestalt psychology. *Scientific American*, December

Treisman, A. M., Gelade, G., 1980, A feature-integration theory of attention. *Cognitive Psychology*, 12:97-136

Ullman, S., 1976, Filling-in the gaps: The shape of subjective contours and a model for their generation. Biological Cybernetics, 25, 1-6

von der Heydt, R. & Peterhans, E., 1989, Mechanisms of contour perception in monkey visual cortex: I. Lines of pattern discontinuity. *J. Neurosci*. 9:1731-1748

von der Heydt, R., Peterhans, E., Dursteler, M. R., 1992, Periodic-pattern-selective cells in monkey visual cortex. *The Journal of Neuroscience*, April 1992, 12(4): 1416-1434

von der Heydt, R., 1995, Form analysis in visual cortex. In Gazzaniga, M. S. (ed.) *The Cognitive neurosciences*, Cambridge, Ma:MIT Press. pp 383-400

Wertheimer, M., 1923, Untersuchungen zur Lehre von der Gestalt. II. *Psychologische Forschung* 4:301-350.

Yen, S-C., Finkel, L. H., 1998, Extraction of peceptually salient contours by striate cortical networks. *Vision Research*, Vol. 38 , No 5, pp 719-741

PAPER III

The Uniqueness Constraint Revisited

A symmetric Near-Far inhibitory mechanism producing ordered binocular matches

Abstract — Psychophysical studies have shown that image features, under certain conditions, can give rise to multiple visible binocular matches. These findings are difficult to reconcile with the traditional interpretation of the uniqueness constraint. A new interpretation, a *conditional uniqueness constraint*, is proposed that allows multiple matching of similar primitives when a one-to-one correspondence does *not* exist locally within corresponding image regions, but prohibits it when a one-to-one correspondence does exist. A cooperative network model and an implementation are also described, where this constraint is enforced at each network node by a simple inhibitory (dual) AND-gate mechanism. The model performs with high accuracy for a wide range of stimuli, including multiple transparent surfaces, and seems able to account for several aspects of human binocular matching that previous models have not been able to account for.

1 Introduction

Due to a frontally directed binocular visual system, humans and many other animals are able to perceive the 3-D structure of the environment, even when no monocular cues (e.g. motion-parallax, perspective, size etc) are available. Because our eyes are horizontally separated, and hence view the world from slightly different perspectives, the two different images that fall onto the left and right retinas will, in general, not be identical. With increasing distance from the plane of fixation, the relative binocular disparity between corresponding visual features in the left and right images will increase. If this binocular disparity, and the degree of convergence of the eyes, can be measured or somehow estimated, so can the distance to a visible feature. The problem of finding corresponding image features in the left and right images is basically a matching problem, and is referred to as the *correspondence problem*. A fundamental difficulty with the

correspondence problem is that it is ill-posed; i.e. the image arrays do not in general contain sufficient information to determine with certainty which solution (of multiple possible) is the correct. Thus, given the total solution space, or the set of all possible matches, the best any visual system can do is to choose the solution that seems most reasonable in the context encountered. What is reasonable in any particular context should preferably be determined with some kind of heuristic that is based on experience (learned or "built-in") of how the physical world behaves. With such knowledge it is possible to *constrain* the matching process, so that solutions that are in agreement with this knowledge are favored to solutions that are unrealistic, or "off the wall". However, if the constraints chosen are not flexible enough, or too strictly applied, there is a risk, when faced with unusual scenes, that a visual system will not deliver accurate descriptions of the environment because it will be too hard "locked into" interpreting all scenes in a certain way. Thus, an important factor in the design of both artificial and natural stereo systems should be to find an acceptable balance between the effectiveness and the flexibility of any particular constraint. A balance that may change from species to species depending on the accuracy needed, or that may change from situation to situation depending on the visual and physical context.

Over the years, a variety of different matching constraints have been suggested: The uniqueness constraint (Marr & Poggio, 1976) states that any given image primitive should be matched with one, and only one, other primitive in the opposite image, because (surfaces in general are opaque) any given point on a surface must have a unique position in space.

The cohesitivity (or continuity) constraint (Marr & Poggio, 1976) states that, because surfaces in general changes smoothly in depth, nearby image points should have similar binocular disparities.

The constraint of figural continuity (Mayhew & Frisby, 1981), or edge connectivity (Baker & Binford, 1981), is based on a similar motivation, but is concerned with edge information. This constraint basically says that an edge segment that is part of a longer continuous edge in one image, should be matched with an edge segment in the opposite image that is a part of a similar continuous edge, so that the figural continuity is preserved binocularly.

The ordering constraint (Baker & Binford, 1981) is motivated by the fact that under normal viewing conditions the relative ordering of image features, in the left and right images, is rarely broken; hence the ordering constraint imposes that binocular matches that preserve the relative ordering should be favoured to unordered ones.

Partly motivated by the same observation as that underlying the ordering constraint, and partly motivated by psychophysical observations (Burt & Julesz, 1980), Pollard, Mayhew and Frisby (1985) have suggested that a disparity gradient limit of 1 could be used to select correct matches. Given two binocular matches, the disparity gradient is defined as the difference in disparity, between

the matches, divided by their cyclopean separation (Burt & Julesz, 1980). Simply put, given a number of neighbouring image features, this constraint can be said to favour (groups of) matches that have relatively similar disparity values , i.e. lie on a smoothly changing (but not necessarily fronto-parallel) surface, over (isolated) matches that have deviating disparity values.

Except for the disparity gradient limit, it seems as if the motivation and justification for these constraints have mainly come from computational considerations, and not so much from psychophysical observation of how the human visual system actually behaves. Naturally, from this it does not necessarily follow that these constraints are wrong, or that they may not account for how the human visual system operates. The ordering constraint, for example, seems rarely - perhaps never - to be broken by the human visual system (see below). Nor does it necessarily follow that a disparity gradient limit is actually used in human vision. In fact, it is unclear both from a computational and a psychophysical perspective why this limit should be fixed to 1, since (as shown below) this does not always seem to hold in human vision. However, the main point here is that although computational considerations can be a highly valuable source of inspiration, and suggest simple and elegant solutions, one must not be blind to when these solutions fail to correspond with the performance of the human visual system.

Of all the constraints above the perhaps most influential and most often encountered in other models of human stereopsis, and other algorithms proposed for solving the correspondence problem, is the uniqueness constraint proposed by Marr and Poggio (1976). It seems as if their particular interpretation has been the prevailing, and only accepted, one. This is surprising considering its obvious shortcomings, when it comes to explaining transparency, and in accounting for known instances of multiple matching in human vision; e.g. Panum's limiting case, the “double-nail” illusion (Krol & van de Grind, 1980), Weinshall (1991, 1993).

This article will start out by taking a closer look at the computational and ecological justification for the uniqueness constraint, and show that the particular interpretation, and implementation, proposed by Marr and Poggio (1976) may be unnecessarily restrictive in which matches it allows, and therefore unable to account for certain aspects of human perception. An alternative interpretation, a *conditional uniqueness constraint*, is then proposed. In brief, this constraint operates in a similar manner (to the one proposed by Marr and Poggio) when there locally is an even number of similar matching primitives within corresponding binocular regions, but allows multiple matches when there is an uneven number of similar matching primitives within the same regions; when a one-to-one correspondence does not hold. A simple computational mechanism that enforces this constraint is then presented along with a cooperative network implementation, and some simulation results. Finally, the model's relationship to previous models, and its plausibility as a model of human depth perception, is discussed.

2 The uniqueness constraint revisited

The motivation for the uniqueness constraint (Marr & Poggio, 1976) is based on the observation that any given point on a surface must have a unique position in space. In their model, this observation was translated into a matching rule that requires any given image feature to be matched with only one feature in the opposite image. Although this particular interpretation has proved to be a highly effective constraint for a wide range of binocular stimuli - and seemingly undisputed - there are several reasons to question its general validity. The motivation for this comes from both computational and ecological considerations, but the main reason - which from the perspective of wanting to understand human stereopsis is more important - comes from the inability of this constraint to account for transparency and known instances of multiple matching in human vision; e.g. Panum's limiting case, the “double-nail” illusion (Krol & van de Grind), and Weinshall's random-dot stereograms (Weinshall, 1991,1993) discussed below.

First of all, it should be pointed out that although multiple matches violate the uniqueness constraint, it does not strictly speaking violate the principle on which the uniqueness constraint is based. That is, allowing multiple matches may not always be sensible, or facilitate the correspondence problem, but it is not in itself contradictory to the fact that any given surface point has a unique position in space; since even if a feature in one image were matched to two or more features in the opposite image, each individual match would still correspond to a unique position in space.

More important in this context is the fact that the uniqueness constraint is not always justified (a fact that Marr and Poggio, of course, were fully aware of). Although it is true that any given point on a surface must have a unique position in space, it is *not* always true that a given image point will correspond to a unique position space, nor that a given surface point will always be represented in both images. For example, when two overlapping transparent surfaces are seen, there is not one, but two “true” disparity values associated with each image point, one for each surface. Further, due to surface slant (and the difference in perspective between the eyes) some features, or surface regions, may appear differently sized, or shaped, on the two retinas. Finally, because the images in our eyes are decompositions of a 3-D world it is unavoidable that some surface points become occluded to one eye but not the other (half occlusion).

In order for the uniqueness constraint to be meaningful there need to exists a one-to-one correspondence between matching primitives. When surfaces are opaque and smoothly curved in depth, and all surface points are visible from both eyes, such a relationship does exists, and the uniqueness constraint can then be an important key to solving the correspondence problem. However, when surfaces do not have these “nice” properties, but for instance there are many occurrences of half-occlusion, there is no guarantee that a one-to-one relationship ex-

ists between the left and right image features. A given image feature may have one, none or many potential matches in the opposite image half. In situations like these, when there locally is an uneven number of identical, or similar, matching primitives in the two images halves, there simply is no way of uniquely matching the primitives without leaving some primitives out. Why some image primitives should be ignored altogether is not easily motivated. One could argue that, while it is true (and indisputable) that any given surface point must have a *unique* position in space, it is equally true that any given surface point must have *some* position in space. And therefore, in situations when the one-to-one correspondence condition does not hold, the price to pay for ignoring some features might sometimes be higher than the price for allowing certain multiple matches (see discussion below).

In previous models of stereopsis (Marr & Poggio, 1976; Marr & Poggio, 1979; Pollard, Mayhew & Frisby, 1985) this problem seems to have been basically ignored, or avoided by assuming that surfaces in general are opaque and smooth. However, if 3-D surface reconstruction is one of the important goals of early vision (Marr, 1982), this is a strange approach to take since any cues of transparency, and particularly half-occlusion would be highly valuable information to such a process.

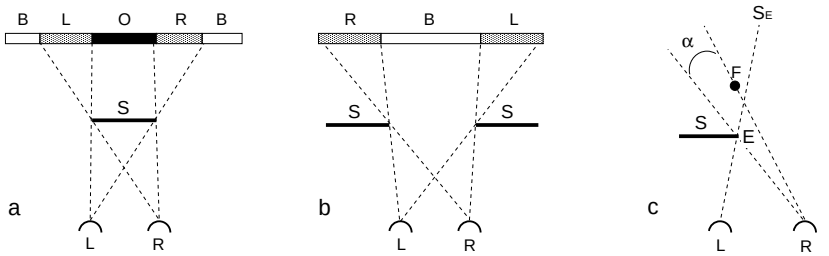

Figure 1. Schematic (top) view of the 3-D layout in cases where unpaired monocular regions may be visible. (a and b) S=occluding foreground surface, B=binocularly visible regions, O=occluded region, L=visible to left eye only, R=visible to right eye only. Left eye monocular regions can only be seen to left of an occluding surface; and right eye monocular regions can only be seen to the right of an occluding surface.(c) Given a binocular fusible surface edge, E, the (minimum) depth of the monocular (right eye) feature, F, depends on the monocular separation, α, between the edge, E, and the feature, F. (Modified after Nakayama & Shimojo, 1990).

Because half-occlusion, in one way or another, is the only fundamental reason why a one-to-one correspondence may *not* exist between image features, it is interesting to consider in some detail the possible variations of 3-D surface layouts in these cases. Regarding an unpaired monocular feature (or region), Nakayama

and Shimojo (1990) have pointed out that there are only two ecologically valid situations where such stimuli can arise (figure 1a,b). If a monocular feature is seen by the left eye only, the feature must be behind an occluding surface that is located to the right of the feature; or vice versa, if seen by the right eye only, the feature must be behind a surface that is located to the left of it. They further pointed out that the depth ambiguity a monocular feature presents, to some extent is constrained if a nearby binocularly fusible edge is present. In figure 1c for example, the region of space from where the monocular feature F could have arisen is delimited by the line RF and the line LS_E (which tangents the fusible right edge, E, of the surface, S). That is, the feature F must have arisen from some point along the line RF that is located to the left of the line LS_E, otherwise it would be visible to both eyes. Thus, the minimum depth corresponding with F is were LS_E and RF intersect. Moreover, this minimum depth depends on the angular separation (α) between the binocularly fusible edge (E) and the monocular feature (F). Hence, the greater the angular separation (α), the greater the minimum depth. Interestingly, Nakayama and Shimojo showed in one of their experiments that subjects did seem to use this information in their depth judgements. When subjects were asked to estimate the depth of an unpaired monocular stimuli their estimates varied quantitatively with the angle of monocular separation from a nearby fusible edge. The larger the angle of separation, the further away the monocular target was perceived.

The case when there is an uneven number of identical, or similar, features in the two images halves (and multiple matching is an option) is highly similar to the case with a single monocular feature described by Nakayama and Shimojo, and it is interesting to consider the problems posed to a binocular matching system when faced with such stimuli.

Consider for example the simple and well known Panum's limiting case (figure 2a) where the left eye sees only one vertical bar, but the right eye sees two. Basically, there are three possible 3-D layouts (figure 2b,c,d) that could give rise to such an image pair: In the simplest case, 2b, the two right eye features (R_1 and R_2) are exactly aligned along the left eye's line-of-sight from L_1. In 2c the single visible left eye feature (L_1) corresponds to the leftmost right eye feature (R_1), and is located on (or in front of) an opaque surface (S) that terminates somewhere between the two features R_1 and R_2. The rightmost feature (R_2) is here occluded to the left eye by the surface (S). Note that this surface is not directly visible in itself, but only indirectly due to the half-occlusion. Finally, in 2d the single left eye feature (L_1) corresponds to the rightmost, right eye, feature (R_2), and is located further away than an opaque (invisible) surface that terminates somewhere to the left of L_1 and R_1. The leftmost right eye feature (R_1) is also further away than the occluding surface, but is only visible to the right eye.

Now, from a computational perspective, given no direct visual evidence of the occluding surface (S) in case 2c and d, all three of these 3-D layouts are equally

likely to have produced the visible stimuli. According to the uniqueness constraint, however, only one of the matches $M_{1,1}$ and $M_{1,2}$ can be allowed. In case 2b, whichever match was chosen, this would never result in a true representation of the situation. And in case 2c and 2d there would be a 50% chance of getting one of the matches right. Worse yet, which is true in all three cases, is that the gist of the scene, the important fact that there exists a disparity jump, would be lost. If on the other hand, both matches were allowed to co-exist in this type of situation, this important fact would not be lost. And not only would it correctly represent the situation in case 2b, but it would also correspond to a fairly good approximation of the true layout in case 2c and 2d. Moreover, as Nakayama and Shimojo (1990) pointed out in their analysis of the strictly monocular case, this approximation would not only be qualitative, but to some extent also quantitative since the correct match (whichever it is) will delimit the possible locations of the other match.

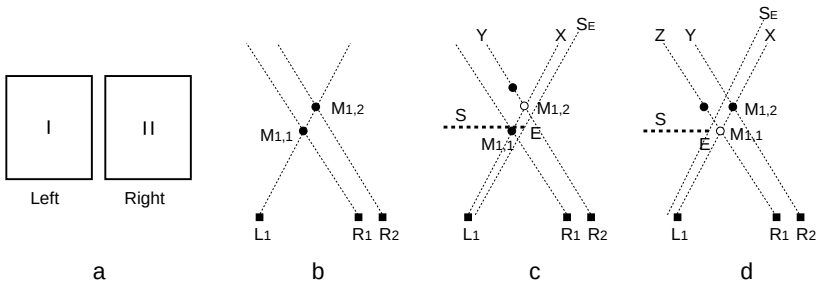

Figure 2. Panum's limiting case (a), and the only three plausible (ecologically valid) 3-D layouts that could have produced it (b, c and d). The filled circles mark the actual positions of the visible features L_1, R_1 and R_2 (filled squares). The thick dashed line is an occluding surface, S, that is not in itself visible. (b) Both match $M_{1,1}$, and $M_{1,2}$, correspond to the actual position of R_1, and R_2, respectively. (c) The true position of R_1 is on, or in front of, the surface S, and R_2 is behind S. Match $M_{1,1}$ is correct, but $M_{1,2}$ is false. (d) Both R_1 and R_2 correspond to points/regions behind the "invisible" occluding surface, S. Match $M_{1,1}$ is false, but $M_{1,2}$ is correct.

In case 2b it is obvious that matching L_1 with both R_1 and R_2 would correspond to the actual positions ($M_{1,1}$ and $M_{1,2}$) of the two features. In case 2c, match $M_{1,1}$ would be a correct match and although $M_{1,2}$ may not correspond to the actual position of R_2 we know (from the analysis of Nakayama and Shimojo) that it must be somewhere along the line-of-sight from R_2, but also to the left of the line L_1X. Therefore match $M_{1,2}$ at least conveys an approximation of the actual layout, to the degree that it corresponds closely to the minimum depth of R_2. Strictly speaking the minimum depth would correspond to where line ES_E inter-

sect with R_2Y, but since neither the surface (S), nor the edge (E), are visible this information is not available. However, if the edge where too far to the right of L_1 it would occlude feature R_2 as well, and therefore the difference should not be very large. Case 2d is perhaps the most interesting because here it is obvious that $M_{1,1}$ is neither the correct position of L_1 and R_1, nor a very good approximation of it. However, using the reverse argument from Nakayama and Shimojo (regarding the location of an unpaired monocular feature), one can say that if $M_{1,2}$ is a correct match (true in this case) and there exists an occluding surface that is not directly visible (as is the case in 2b); the position of this surface edge will be delimited by the angular separation between match $M_{1,2}$ and the monocular feature R_1. That is, this surface (edge) must lie to the left of both line L_1X (otherwise none of the features would be visible to the left eye), and line R_1Z (otherwise R_1 would be occluded to the right eye); but we also know that it can not lie too far to the left of L_1 and R_1, since otherwise R_1 would be visible to the left eye as well. In case 2d, one can therefore say that although $M_{1,1}$ is an incorrect match of the visible *features*, it does represent the 3-D layout in the sense that it approximates the position of the *edge* in the same manner as $M_{1,2}$, in case 2c approximates the position of feature R_2.

To rephrase this in different words, because the only three ecologically valid surface layouts that could generate the stimuli in Panums's limiting case, all involve an occluding surface that terminates in the near vicinity to the left of the occluded feature (and also immediately to the right in the border case 2b); it does not really matter which match is actually the correct when it comes to capturing the essence of the scene; i.e. that there exists a disparity discontinuity near L. Thus the best any visual system could do in such cases seems to be to accept both matches and, given the available (lack of) evidence, or cues, treat them as the best possible interpretation of the 3-D layout. Contrasting this approach with the traditional version of the uniqueness constraint will at any rate result in a less grave error in the interpretation of the scene, and a smaller loss of potentially valuable information.

Given the strong similarity between Panum's limiting case, discussed above, and the strictly monocular case discussed by Nakayama and Shimojo (1990), two things are worth pointing out about their study. First, in the experiment where subjects experienced the depth of the unpaired stimulus to depend on the monocular separation from the matchable surface edge, it is quite possible that the binocularly visible edge was matched twice. The stimuli used for both the binocularly fusible surface, and the monocular target, were white rectangular regions. If the surface edge were matched twice, this would explain why the unpaired bar appeared further away with increasing separation from the binocularly fusible edge. Second, in another of their experiments where unpaired vertical bars had been inserted (in an ecologically valid manner) to simulate an occluding

(but in itself invisible) surface, subjects did perceive illusory, or subjective, contours at the appropriate side of the unpaired (half-occluded) features.

3 Psychophysical evidence for multiple matching

There are several examples where the human visual system, when faced with an uneven number of similar features, seems to make multiple matches. In the Panum's limiting case (figure 3a) it has been known since long that, given a small separation between the bars, the fused percept is not that of a single bar, but instead one sees two bars (one in front of the other). In the trivial case where both eyes sees two bars (figure 3b), it is interesting to note that the fused result is not that of two bars superimposed on two bars that lie further away (schematically depicted to the right in figure 3b) even though this is a possible interpretation, but instead that of two stable bars in the same depth plane. In this trivial case it seems as if the two redundant (unordered) matches indeed are suppressed. This shows that the visual system clearly separates between situations, and treats stimuli different, depending on if there is an even or uneven number of matching primitives; and it seems to indicate that the visual system chooses the least complicated interpretation, given the available information.

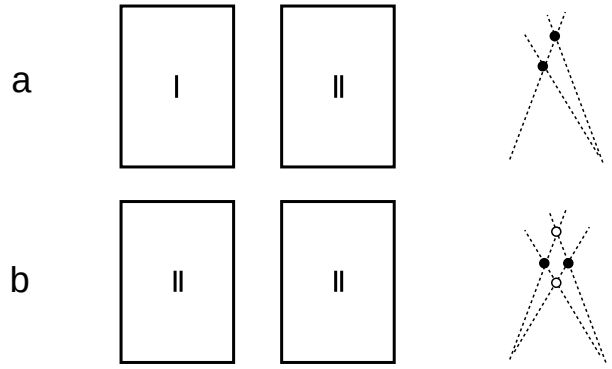

Figure 3. a) Panum's limiting case, and b) the “trivial” case where only the two ordered matches (filled circles) are seen.

Another example where multiple stable matches have been demonstrated is in the “double-nail” illusion (Krol & van de Grind, 1980). In the basic version of this illusion (figure 4a) two nails of the same length are placed, with their heads vertically aligned, on a lath, which is aligned with the midsaggital plane. When the setup is viewed with both eyes, one does not see the nails at their actual positions, but instead at the positions corresponding to the false, or ghost, matches

(figure 4b). In a variation of this experiment (which also can be seen as a variation of Panum's limiting case), Krol and van de Grind added a third nail that was placed behind the other two and aligned with the left eye's line-of-sight through the middle nail so that it was only visible to the right eye (figure 4c). With this setup their subjects reported that there were two stable vergence conditions (4d). In both cases the two matches $M_{1,1}$ and $M_{2,3}$ were seen, but depending on if fixation was on either $M_{1,1}$, $M_{2,2}$ was also seen, or if on $M_{2,3}$, $M_{1,2}$ was seen.

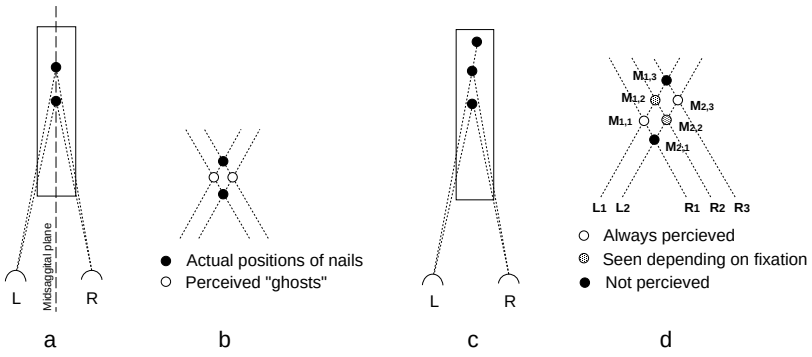

Figure 4. a) Basic setup of the double-nail illusion. b) Schematic view of the possible matches. c) The variation of the double-nail illusion, where a third (most distant) nail is placed so it is occluded to the left eye only. d) Possible matches in the latter setup. (Modifed after Krol & van de Grind; 1980).

A more massive form of multiple matching has been demonstrated by Weinshall (1991, 1993). In a series of experiments, using ambiguous random-dot stereograms (RDS), which consisted of multiple copies of the (basic) double-nail illusion stimuli, Weinshall reported that subjects saw up to four different transparent depth planes. These ambiguous RDS were constructed by making a single random-dot image which was copied twice into each half of the stereogram with an inter-dot separation that was G_L in the left image, and G_R in the right image. When the dot separation was the same ($G_L = G_R$), subjects saw only one opaque surface, the one corresponding to the ordered matches - as would be expected from the single pair case. In contrast to the single pair case, however, when the inter-dot separation was different ($G_L \neq G_R$) their subjects often reported of seeing, not only the two planes corresponding to the ordered matches, but also one or both of the ('ghost') planes corresponding to the unordered matches. Weinshall also reported of an unpublished study by Braddick (see Weinshall, 1993) in which similar RDS stimuli where used, with the difference that the RDS pattern was only copied once in one of the stereo pair images. Like in the single pair

case (Panum's), the dots in the single-copy image were matched twice and two depth planes was perceived.

As a final example, MaKee, Bravo, Smallman and Legge (1995) have found, measuring the contrast-increment threshold for high-frequency targets, that a single target in the left eye can mask both of two identical targets in the right eye, when arranged in a configuration like the Panum's limiting case; and that this binocular masking was nearly the same as when only one target was visible to the right eye. To confirm that the left target actually had been matched twice, they further conducted a depth-judgement test where subjects (over 200 trials) had to determine if a test target in the right eye where in front of, or behind, a reference target. The stimuli presentation time (200ms) was too short to allow for voluntary vergence movements. The results of these depth judgements where essentially the same as when only a single target was present in the right eye, and hence supported the idea that the left target was matched with both targets in the right eye. MaKee et al. (1995) concluded that "uniqueness is not an absolute constraint on human matching".

Taken together the above examples strongly suggest that when there locally exists an uneven number of similar features in the left and right retinal images, and there consequently does not exist a one-to-one correspondence between matching primitives, the human visual system does allow multiple matches to co-exist. If so, this clearly speaks against a strict enforcement of the uniqueness constraint; i.e. one that ignores or suppresses other potential matches under such conditions, and consequently it also speaks against models of human stereopsis that blindly rely on this constraint, since they are not flexible enough to account for the above results. What instead seems to be called for is a more sensitive usage, or enforcement, of the uniqueness constraint that, given a feature for which a match is sought, not only takes into consideration the similarity to features in the opposite eye, but also to the presence (and/or absence) of similar features in the eye-of-origin.

Returning briefly to the double-nail illusion, this is interesting not only because it demonstrates multiple matching, but also because it shows how strongly the human visual system seems to prefer matches that preserve the relative ordering of image features. In the basic version of this illusion, subjects consistently perceived the targets at their ghost positions instead of at their actual position. Perhaps counter intuitively this percept survived despite remarkably large changes in the nails dimension/appearance such as when the distal nail was longer than the other and both heads could be seen at different heights; one nail was twice the diameter of the other; or the proximal nail was rotated 5 deg around an horizontal axis through its centre. Despite the many, both monocular and binocular, cues that could have been used by the visual system to choose the correct matches the ghosts were consistently preferred. To explain this phenomena, Krol and van de Grind suggested a (somewhat vague) rule which stated that

the perceived pair was always the pair of matches closest to the fixation point. A simpler (and in their case equivalent) explanation is that our visual system always chooses matches that preserve the relative (left-right) ordering of features in the two retinal images. It should be noted that Panum's limiting case does not, strictly speaking, break this principle but can in fact be considered as a border case.

4 A revised uniqueness constraint

The alternative interpretation of the uniqueness constraint, proposed below, arose from an attempt to reconcile and explain the three following properties of, or rather assumptions about, human binocular matching: *i*) When a one-to-one correspondence *does* exists between matching primitives, any given primitive is matched only once. *ii*) When a one-to-one correspondence does *not* exists locally between *similar* matching primitives, any given primitive is matched at least once. *iii*) Binocular matches that preserve the relative ordering of image features are always preferred to unordered matches.

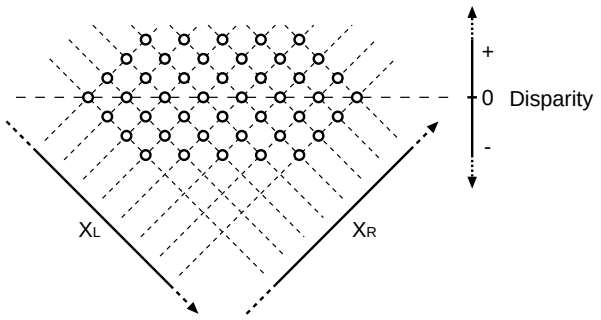

Figure 5. Schematic view of a horizontal layer in the network, showing the mapping of the left and right x-axis. Positive disparities correspond to further depths, and negative disparities correspond to positions closer to the plane of fixation.

At a first glance, these three conditions on the matching mechanism may seem unrelated and therefore not easily integrated, but in fact they can be implemented by a quite simple (dual) mechanism. The basic architecture of the model consist of a network of interconnected nodes that each represent a particular point in disparity space. Figure 5 shows a horizontal layer of this network. Initially the activity at each node $M(x_L,x_R,y)$ in the network is proportional to the similarity, between the stimulus at position (x_L,y) in the left image and position (x_R,y) in the right image. In this default mode, or rather without any additional modifications, no particular constraint would be enforced (except that matches are sought only

along corresponding epipolar lines), and given, for example, the stimuli in Panum's limiting case, both of the nodes that represent the two possible matches (figure 3a) would remain active. Although such a simple model could account for the percept in Panums's limiting case, it obviously can not account for why we only see the two ordered matches in the trivial case (figure 3b); or why we, for example, only see the five ordered matches with stimuli like that depicted in figure 6, although there are 25 possible combinations between the left (L_1...L_5) and right (R_1...R_5) features. For the model to handle such stimuli, some kind of inhibition must be introduced. Hence, to prevent multiple matching from occurring in the model when a one-to-one correspondence *does* exists, between the left and right image features, a *conditional uniqueness constraint* is introduced. Like in the model proposed by Marr & Poggio (1976), the uniqueness constraint is enforced by mutual suppression between matches that lie along the same lines-of-sight, but unlike their model this suppression is dependent upon that two conditions must hold for the binocular stimuli. The first of these two conditions is required assure property (*i*) above, and the second condition is needed to assure property (*iii*). However, although the two conditions have different motivations, they are not independent of each other but on the contrary tightly coupled. In fact, without the second condition the first would be meaningless.

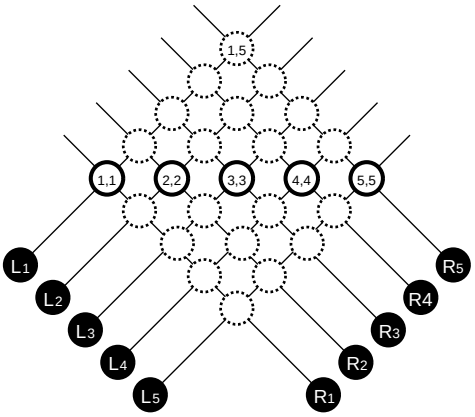

Figure 6. Example of the basic difficulty with the correspondence problem. Five identical stimuli (L_1...L_5 and R_1...R_5) are visible to both the left and right eye, and 25 different matches ($M_{1,1}$...$M_{5,5}$) are possible. In general however, only the five ordered matches are perceived (filled circles). The match $M_{1,1}$ (of L1 with R1) has no competing matches to the left of it. Neither its Near, nor Far, AND-gate (see fig. 7) will therefore be activated, and the node $M_{1,1}$ will not be suppressed. Match $M_{1,5}$, on the other hand, will be strongly suppressed, because its Near AND-gate have input from both sides (multiple competing matches along both the left, and right, line-of-sight).

Given a target match $M_T(x_L,x_R,y)$ the first condition is that other matches, lying along one of the two lines-of-sight (e.g. $M([0....n],x_R,y)$), may suppress $M_T(x_L,x_R,y)$ only if there also exists potential matches along the opposite line-of-sight (e.g. $M(x_L,[0....n],y)$). Or differently put, if a feature in, for example, the left eye can be matched with more than one feature in the right eye, the corresponding matches compete for dominance only if there also exists other (similar) features in the left eye that also can be matched with the same set of features in the right eye.

The second condition on the suppression, given the target match $M_T(x_L,x_R,y)$, is that only combinations of matches that have the same sign of the disparity relative to $M_T(x_L,x_R,y)$ contribute to the suppression (e.g. $M([(x_L+1)...n],x_R,y)$ in combination with $M(x_L,[(x_R+1)...n],y)$, and/or $M([0...(x_L-1)],x_R,y)$ in combination with $M(x_L,[0...(x_R-1)],y)$). Or in other words, matches along one line-of-sight that lie further away than the target M_T, are allowed to suppress it only if there are also matches along the opposite line-of-sight (through M_T) that lie further away; And, vice versa, matches along the line-of-sight that are closer than M_T, are allowed to suppress M_T only if there are also matches along the opposite line-of-sight that are closer.

One simple mechanism that enforces both of these conditions is depicted in figure 7, and consists of a pair of inhibitory AND-gates: a Near AND-gate (or Near-gate) and a Far AND-gate (Far-gate). Consider that each node, in the basic network described above, has such a dual mechanism connected to it (as shown for only one node in figure 7). Now, the Near-gate separately sums the activity, along the left and right line-of-sight, of all nodes that lie in front of M_T. Only if both the (left and right) sums are larger than zero does the Near-gate produce an output that suppresses node M_T. The Far-gate is identical in operation to the Near-gate, but receives its input from matches that correspond to depths further away than the target node M_T.

If one considers the dynamic interactions between the nodes in the network and their associated inhibitory mechanisms, it should become clear why this model will produce a unique and ordered set of matches for the stimuli in figure 6, i.e. the matches in the middle horizontal row (this solution set would be predicted by any known stereo constraint).

Consider for example the two features L_1 and R_1 (figure 6), from either view these two are the leftmost feature, and should therefore be matched with each other. Any other solution would be incongruent with the ordering constraint. Because both the Near-gate and the Far-gate associated with node $M_{1,1}$ only receives input from matches along one of the line-of-sights, the net output of both of these AND-gates will be zero, and consequently node $M_{1,1}$ will not be suppressed. For any other node along the line-of-sights from L_1 or R_1 this is not the case. At all these other positions there is some input (either in front of, or behind, the node) from both the left and right line-of-sight, and they will consequently

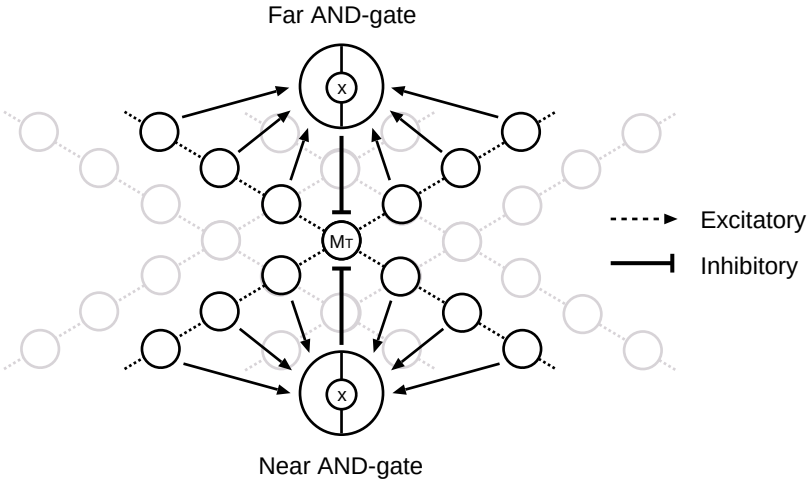

Figure 7. Depiction of the proposed dual mechanism consisting of a Near, and a Far, inhibitory AND-gate. Each AND-gate separately sums the activity along the left, and right, lines-of-sight (through the target node M_T), and multiplies the two sums. If both the left, and right, side input to an AND-gate is greater than zero, the target node, M_T, will be suppressed.

all be inhibited to some degree. Consider particularly the node $M_{1,5}$, corresponding to the match of feature L_1 with R_5. From the left view, L_1 is the leftmost feature, but from the right view R_5 is the rightmost. Clearly, matching these two features would severely violate the ordering principle, and it is therefore an unacceptable solution. In this case the Near-gate connected with $M_{1,5}$ will receive a strong input from both the left and right line-of-sight, and the node will therefore be strongly suppressed. This latter situation quite nicely illustrates the motivation for the second condition above. That is, except for cases when the binocular stimuli consists of highly repetitive patterns (such as in the Wallpaper-illusion), a high number of competing matches on both the left and right side of the target match (M_T) indicates that the target is not an ordered match. The output of the (multiplicative) AND-gate can therefore been seen as a measure (although not the only) of how well the match is ordered in relation to its neighbours. However, even if the stimuli is (finitely) repetitive as in figure 6, the proposed mechanism will produce an ordered set of matches. Consider, for example, the "correct" match $M_{3,3}$ in figure 6, which has competing matches with crossed and uncrossed disparities (relative to it) along both the left and right line-of-sight. Initially, this node will be suppressed. However, because node $M_{1,1}$ is not sup-

pressed but all other nodes along its lines-of-sight are, both of node $M_{2,2}$'s AND-gates will eventually loose their input from one side and consequently node $M_{2,2}$ will be "disinhibited". Once $M_{2,2}$ has been disinhibited, it in turn will suppress the nodes that input to the AND-gates connected to node $M_{3,3}$, leading to its revival. As this disinhibitory process propagates further in to the network eventually only nodes that preserve the ordering will remain.

5 An Implementation

A computer implementation of the proposed model have been realised and tested on a variety of different stimuli. The model consists of two major levels: a preliminary matching stage (*I*), and a secondary relaxation stage (*II*) where the conditional uniqueness constraints is enforced. At the preliminary stage, suitable matching primitives are identified independently in each image array, and subsequently binocularly matched for similarity. The result (i.e. the set of all potential matches for the stereogram in question) is then fed forward to the secondary stage. At the secondary stage, all potential matches are allowed to inhibit (disinhibit) each other, over a number of iterations (according to the principles described above), until a stable state is reached. To balance the inhibition in the network and avoid that a "dead" state is reached, there is a small continuous excitatory feed from the preliminary matching stage into the secondary stage.

Because the purpose of the current implementation is mainly to show that the principles enforced in the secondary stage are sound (i.e. that the correspondence problem can be effectively solved for a wide range of stimuli), the preliminary stage has been kept as simple as possible. In the current implementation, the preliminary matching stage simply uses the intensity values at each image point as "matching primitives", and a binary function for the similarity evaluation. That is, if the image intensity at a given point in the left image array (x_L,y), and at some point (x_R,y) in the right image array, are both greater than zero, then the node $M^I(x_L,x_R,y)$ in the preliminary (and initially also the secondary) network will represent a potential match and be assigned a value of 1. If there is no image intensity at one, or both, of these points, the node $M^I(x_L,x_R,y)$ will not represent a match and will hold a value of 0.

Because the ordering constraint is a key principle in the model, it should be obvious that the current (preliminary) matching strategy is not well suited for natural images that contain large surfaces regions with homogeneous intensity values; since surface points within such regions can not be meaningfully ordered. However, for sake of demonstrating the proposed mechanism the current preliminary stage is sufficient, since all examples below are either random-dot, or simple (vertical) line, stereograms with binary intensity values (black/white). The advantages with using artificial, over natural, stereograms as performance probes are that the stimuli can be precisely controlled and arranged

as desired, and more importantly, an accurate disparity map is available for validation, which is usually difficult to obtain for natural image stereograms.

Assuming the preliminary stage has matched each image point, with all points in the opposite image that lie within a horizontal disparity range of +/-D pixels (in the examples below D=12); and the result have been loaded into the secondary network, the value of any node in the (secondary) network is given by the semi-saturation function:

$$f_S(x,\sigma)=K\cdot\frac{x^2}{x^2+\sigma^2} \tag{5.1}$$

which reaches half K when $x=\sigma$. Here, x and σ are compound terms representing the total excitatory respectively inhibitory input to the node. More specifically (with $K=1$), given a particular node $M^{II}(\vec{p})$ in the secondary network, where $\vec{p}$ = (x_L,x_R,y) is the triplet, or vector, that defines the position in the network, its value at iteration $t+1$ is given by:

$$M^{II}_{t+1}(\vec{p}) = f_S\left(M^{II}_t(\vec{p})+M^{I}(\vec{p})\cdot A\cdot e^{-B\cdot S_t(\vec{p})},\ \sigma_s+C\cdot S_t(\vec{p})\right) \tag{5.1}$$

where $M^{I}(\vec{p})$ is the node in the preliminary network holding the result of the initial matching. A, B, C are constant weight factors, and σ_s is the semi-saturation constant. Finally, the term $S_t(\vec{p})$ is the sum of, the inhibitory input, from the Near AND-gate, $N_t(\vec{p})$, and the Far AND-gate, $F_t(\vec{p})$, connected to node $M^{II}(\vec{p})$:

$$S_t(\vec{p})=N_t(\vec{p})+F_t(\vec{p}) \tag{5.3}$$

$$N_t(x_L,x_R,y)=\sqrt{\sum_{d=1}^{D+(x_R-x_L)} M^{II}_t(x_L+d,x_R,y)\cdot\sum_{d=1}^{D+(x_R-x_L)} M^{II}_t(x_L,x_R-d,y)} \tag{5.4}$$

$$F_t(x_L,x_R,y)=\sqrt{\sum_{d=1}^{D-(x_R-x_L)} M^{II}_t(x_L,x_R+d,y)\cdot\sum_{d=1}^{D-(x_R-x_L)} M^{II}_t(x_L-d,x_R,y)} \tag{5.5}$$

6 Simulation results

For all the examples presented below the same parameter set was used ($A=\sigma_s=0.5$, $B=8$ *and* $C=4$). As of yet, no proper formal analysis of the parameter space has been carried out, but the above values were empirically found to be

a good trade-off between speed-of-convergence and the number of correct matches. In each stereo triplet, the left and middle images can be free-fused with uncrossed-crossed eyes, and the middle and right images with crossed eyes.

6.1 Example 1: Vertical lines

The first example (figure 8) contains several of the simple cases described in previous sections, and illustrates the basic implications of the conditional uniqueness constraint both in cases when there is an even number of similar features in the two halves of the stereogram, and in cases where there is an uneven number of similar features. From top to bottom, the first and second row are examples of the Panum's limiting case, and the corresponding "trivial" case, respectively, from figure 3. The third row contains the variation of the "double-nail" illusion illustrated in figure 4c. In each of the last three rows, there is an

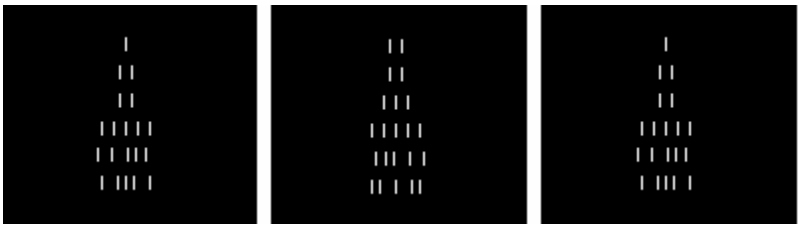

Figure 8. The left and middle images can be free fused with un-crossed eyes, and the middle and right images with crossed eyes. Row: 1. Panum's limiting case; 2. Its "trivial" case; 3. Variation of the double-nail illusion; 4. Plane at zero disparity; 5. Peak/Pyramid; 6. Peak and through.

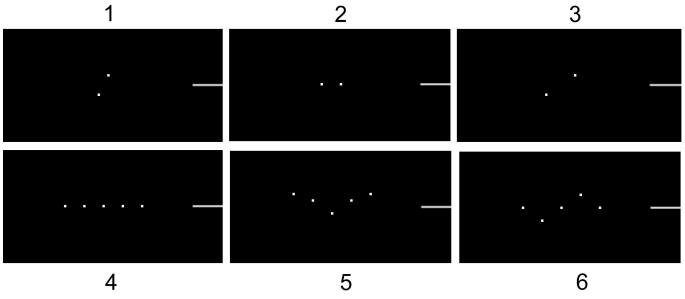

Figure 9. Horizontal cross-sections of the network, after it has converged (with the input in figure 8) into a stable state. The right horizontal line marks zero disparity.

even number of features in the two image-halves. In row four, all lines lie in a plane at zero disparity (the example from figure 6). In row five, they form a peak with the middle bar on top, and in row six, they form a peak and a trough (from left to right). The output is displayed in figure 9 as horizontal cross-sections where the rightmost horizontal line mark zero-disparity. As can be seen, the model allows both of the two matches (fig 9-1) in the Panum's limiting case, but only the two ordered matches (fig. 9-2) in its corresponding "trivial" case. Case three corresponds to the variation of the double-nail illusion (figure 4c). As explained above (in the context of figure 4d), when subjects were presented with this stimuli they reported of seeing both of the matches $M_{1,1}$ and $M_{2,3}$ (the two visible ones in figure 9-3), but also either $M_{2,2}$ if fixation was on $M_{1,1}$, or $M_{1,2}$ if the fixation was on $M_{2,2}$. This seem to suggests that human stereopsis to some degree is biased towards matches that lie in the plane of fixation. However, in the current implementation no such bias has been introduced, and therefore the two matches (in fig 9-3), that correspond to the matches $M_{1,2}$ and $M_{2,2}$ in figure 4d, are equally suppressed (not visible in fig 9-3). The final three cases are quite straight forward, and as the output shows (figure 9-4, 9-5 and 9-6) there are no instances of multiple matching since there exists a one-to-one correspondence between the features.

6.2 Example 2: RDS - Occluding square

The random-dot stereogram in figure 10 contains two opaque surfaces: a background plane at zero disparity, and an occluding square in the foreground. The dot-density in this particular example is 10%, and the disparity difference between the two planes is 4 pixels. Figure 11 shows a 3-D plot of the model output. The amount of correct, and false, matches were 93.3%, and 10.5%, respectively. 6.7% of the dots were unmatched (see table 1 for results with different dot-densities).

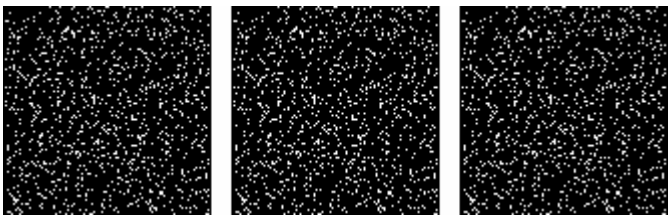

Figure 10. Random-dot stereogram displaying a smaller occluding square, in front of a background surface at zero disparity.

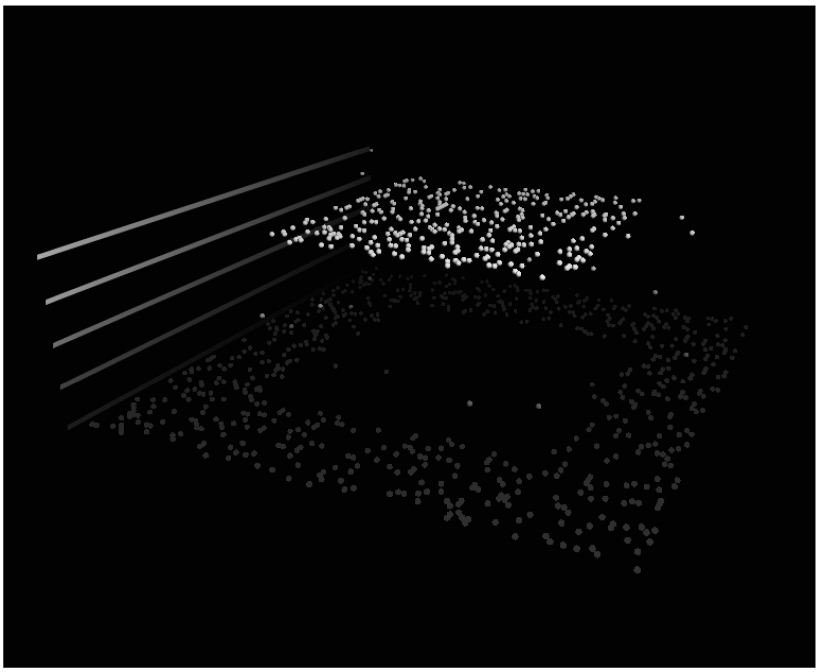

Figure 11. 3-D plot of the model output for the stereogram in figure 10. Each dot represents an active node in the network. The horizontal lines (to the left) mark the separation between different disparity layers.

6.3 Example 3: RDS - Gaussian "needle"

In the third example (figure 12), the disparity (dX_m) in the middle image, relative to the left (right), image was generated by the following Gaussian-like distribution:

$$dX_m = -10\mathrm{e}^{-\frac{(I-x)^2+(I-y)^2}{\sigma^2}} \tag{6.3.1}$$

where I is half the image width (I = 64 pixels) and σ = 12. When fused, a sharp peak is perceived pointing out from a flat background. At the steepest part (between the background and the peak), the disparity-gradient is approximately 0.7. Figure 13 shows a 3-D plot of the model output (see also table 1). This example is an interesting test because it should pose a problem for models that assume smooth (particularly fronto-parallel) surfaces, and rely on some kind of

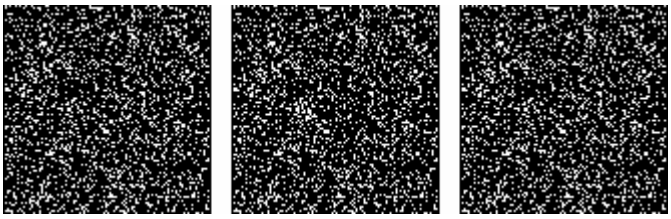

Figure 12. RDS of a flat opaque surface with a sharp narrow peak coming out in the middle. Dot-density 20%.

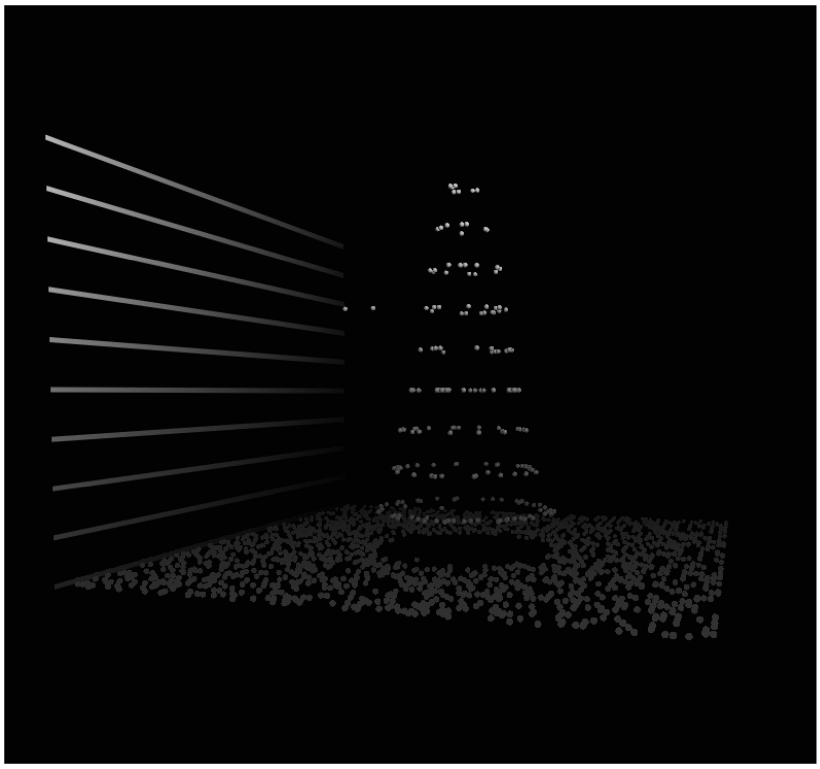

Figure 13. 3-D plot of the model output for the stereogram in figure 12.

spatial summation (i.e. mutual support between neighbouring matches with equal or similar disparities) to solve the correspondence problem. What should pose a problem for such models is that, at the top of the peak, there are only a

very few dots. The support that the correct matches give each other could easily be "drowned" by the more massive support that the background could give to potential false matches with disparities closer to the background (see also discussion).

Stimuli / density (%)	**Matches**			**Iterations**
	Correct (%)	**False** (%)	**Unmatched** (%)	
Square				
5%	98.0	4.6	2.0	24
10%	93.3	10.5	6.7	51
15%	91.6	12.3	8.4	72
20%	88.6	12.6	11.4	162
"Needle"				
5%	100.0	0.0	0.0	31
10%	99.3	0.8	0.7	65
15%	98.9	1.1	1.1	85
20%	98.0	1.3	2.0	126
Transp.				
5%	93.4	5.8	6.6	23
10%	82.3	16.0	17.7	42
15%	72.9	25.4	27.1	60
20%	65.5	33.0	34.5	121
Needle+Transp.				
5%	96.6	3.0	3.4	21
10%	89.7	9.5	10.3	44
15%	80.4	18.3	19.6	71
20%	72.8	24.6	27.2	116
RDRDS				
5%	95.3	3.8	4.7	16
10%	84.8	14.5	15.2	34
15%	80.0	19.6	20.0	72
20%	68.3	30.9	31.7	107

Table 1. Performance results of the implementation when tested on example 2-6 above, with different dot-densities. The rightmost column shows the number of iterations required for the network to settle down to a stable state; i.e. when there was no, or very small (<0.001%), change in the activity between successive iterations.

6.4 Example 4: RDS - Transparent layers

Because uniqueness is not an absolute constraint in the model, multiple transparent surfaces do not pose as serious a problem as it does to other models that allow only a single disparity value at any image point. However, as the number of planes, or the dot-density, rises, naturally, the chance increases that the relative ordering of adjacent features within the two image halves will be jumbled. The fact that the model is cooperative does to some extent counteract such mismatching, if there are strong unambiguous matches in the near vicinity. In example 4 (fig. 14) there are two transparent planes with a disparity separation of 4 pixels. Note in table 1 how the performance deteriorates with increasing dot-density. Such a deterioration, with increasing dot-density, has also been described in human vision (see Akerstrom & Todd, 1988).

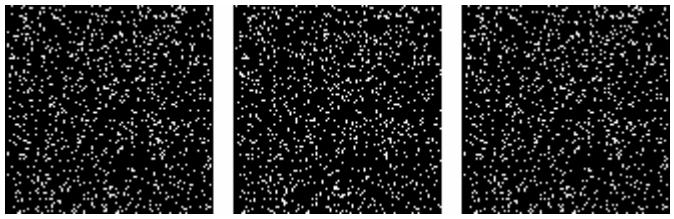

Figure 14. RDS containing two transparent planes, separated by a 4 pixel disparity. Dot density 10%.

6.5 Example 5: RDS - Gaussian “needle” with one transparent plane

In example 5 (fig. 15) a transparent layer have been added to the needle in example 3. The needle constitutes an opaque background, while the transparent layer cuts through the peak half way up.

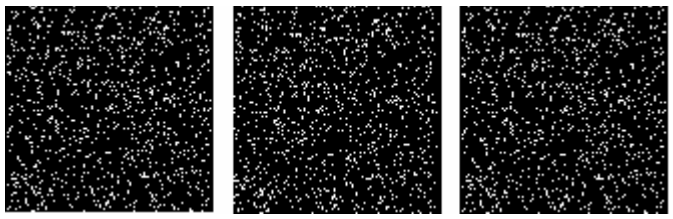

Figure 15. RDS of an opaque background surface that has a peak in the middle, which sticks through an additional flat transparent surface. Dot-density 10%.

6.6 Example 6: Random-Disparity Random-Dot Stereogram (RDRDS)

In the random-dot stereogram in figure 16, the disparity between each pair of corresponding dots was randomly set, and lies in the interval from -3 to 3 pixels around zero disparity. This type of stereogram is clearly a challenge to models that rely on assumptions about surface smoothness, since there is no correlation whatsoever between the disparity values of any two neighbouring dots. Surprisingly, this type of stimuli seem quite easily fused as long as the dot-density is relatively low, and the disparity interval is not too large (according to a highly informal study where two members at our department participated as subjects).

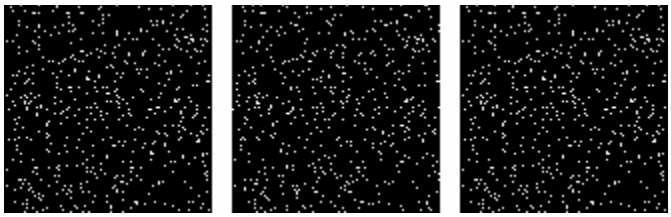

Figure 16. Random-dot stereogram with random disparities between corresponding dot-pairs. Dot-density 5%.

6.7 Example 7: RDS - Weinshall stimuli

Figure 17 shows an example of the basic stimuli used in Weinshall's studies (1991, 1993). What is particularly interesting with this stimuli is that it shows that the human visual system treats the same stimuli different when it is seen in isolation from when it is seen *en masse*. Recalling that in the (single) case of the double-nail illusion, subjects saw only the two ordered matches. In the "Weinshall-stimuli", on the other hand, where the same stimuli has been copied multiple times into the same stereogram, subjects saw up to four planes (corresponding to the disparities of both the ordered and the ghost matches). A number of variations of this stimuli were tested on the model (see Table 2), where the dot-density was the same as in the Weinshall study (9% x 2 = 18% total), and the inter-dot separation (G_L and G_R) was varied. Each column/example in table 2 shows the averaged result of three different stereograms with the same values of G_L and G_R. Interestingly, but unanticipated, the model produced results that had the same basic characteristics as that of human vision in this respect. When the inter-dot separation in one image was zero, but greater than zero in the other image (i.e. multiple copies of Panum's limiting case), the great majority of matches was concentrated at zero disparity, and the disparity corresponding with the non-zero inter-dot separation. The remaining matches were fairly evenly distributed over all other disparities. When the inter-dot separation was the same (G_L=G_R) in

the two image halves, the output was concentrated to one single plane at zero disparity. When G_L and G_R was different from each other (and both greater than zero), again the majority of activity/matches was in the two planes corresponding to the ordered matches, but there was also some activity at several other disparities with clear peaks at the two planes corresponding to the two ghost matches.

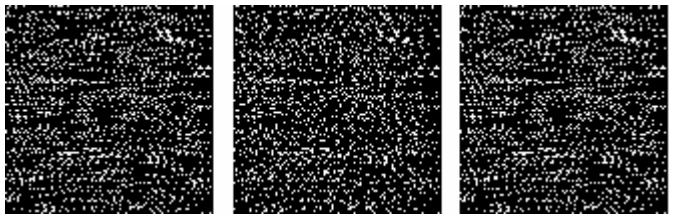

Figure 17. Example of the stimuli used by Weinshall (1991, 1993), with G_L=4, and G_R=8. Dot-density 18%.

In Weinshall's study (1991) subjects saw between two and four planes with stimuli like this. Seeing to the absolute number of matches only (in the model output), there were relatively few matches in the two planes corresponding to the two ghost matches, and even fewer than in some of the other layers which should not produce a perception of a plane. However, first of all, it only takes a dot-density as low as 0.001 (Weinshall, 1991) to produce a sensation of a plane in these stimuli, which in the examples above corresponds to about 10 dots. So clearly, the number of dots in the ghost planes are above that value in both of the examples, in table 2, of this type (G_L=4, G_R=7 and G_L=6, G_R=3). Further, it is not clear whether it is the absolute number of matches, or some relative measure that creates the sensation of a plane in human perception. In the example below with G_L=4 and G_R=7, there are for example only 16 matches in the “ghost” layer (at disparity 7), but more than that in several of the other layers (4, 2, 1, -1) which should not produce the perception of a plane. On the other hand, of all the planes that contain more than 10 matches in them, only in the layers -4, 0, 3 and 7 does the number of matches reach local peak values. Whether or not this may explain the psychophysical observations for this type of stimuli remains to be seen.

In the final two examples (last two columns in table 2), all accidental matches were removed from the stereograms. In Weinshall's study (1991), accidental matches were avoided by spacing the dots so that the disparities of possible accidental matches were larger in absolute values than the possible disparities for each corresponding (left-right) dot-pair. With the same procedure for removing accidental matches the model output was concentrated exclusively to the two planes corresponding to the ordered matches of each dot-pair, which is in

	$G_L=0$ $G_R=3$ ***(0,3)*** *It: 105*	$G_L=2$ $G_R=2$ *(-2* ***0****,2)* *It: 93*	$G_L=4$ $G_R=7$ ***(-4,0,3,7)*** *It: 135*	$G_L=6$ $G_R=3$ ***(-6,-3,0,3)*** *It: 139*	$G_L=6$ $G_R=3$ *(-6,* ***-3,0****,3)* *No accid.* *It: 63*	$G_L=2$ $G_R=4$ *(-2,* ***0,2****,4)* *No accid.* *It: 67*
Disparity (pixels)	**Number of active matches**					
12	17	0	2	0	0	0
11	18	0	1	1	0	0
10	25	0	2	1	0	0
9	25	0	2	3	0	0
8	20	0	2	1	0	0
7	25	0	**16**	1	0	0
6	28	0	9	1	0	0
5	32	0	10	3	0	0
4	31	0	28	3	0	0
3	**468**	0	**593**	**29**	0	0
2	52	0	107	15	0	**729**
1	48	0	107	18	0	0
0	**512**	**1567**	**625**	**617**	**694**	**747**
-1	32	0	32	108	0	0
-2	36	0	11	113	0	0
-3	33	0	10	**564**	**647**	0
-4	32	0	**26**	16	0	0
-5	25	0	2	15	0	0
-6	23	0	1	**31**	0	0
-7	18	0	2	5	0	0
-8	21	0	1	3	0	0
-9	19	0	1	3	0	0
-10	12	0	0	4	0	0
-11	14	0	1	2	0	0
-12	10	0	1	1	0	0

Table 2. Model output with different values for the left (G_L), and right (G_R), image inter-dot separation, in stereograms of the same type as used by Weinshall (1991, 1993). The number of active matches, and the number of iterations (It.) required, shown in each column are the average results for three different stereograms (with the same values of G_L and G_R in each trial). The numbers within parenthesis (second top row) mark at what disparities matches are possible, if only corresponding (left-right) dot-pairs are considered; i.e. all possible matches between the dots in a left image dot-pair, and the dots in the corresponding right image dot-pair (the copy). The numbers in bold correspond to disparities where Weinshall's subjects reported of seeing depth planes. No accid. stands for no accidental matches.

accordance with the results obtained by Weinshall. However, not surprisingly - considering that the ordering principle is the key constraint in the model - the model produced the same output as long as the separation between any two (following) dot-pairs was at least one pixel larger that the value of $max(G_R,G_L)$. This spacing preserves the relative ordering of the dots, but it does not guarantee that the possible disparities (in absolute values) between the dots within each (left-right) dot-pair is smaller than that between possible accidental matches.

7 Discussion

7.1 Relation to other models

The one major feature of the proposed model that separates it from, possibly all, previous models of stereopsis (and other algorithms devoted to the correspondence problem) is its more relaxed version, or interpretation, of the uniqueness constraint. Because multiple matches are not penalised when a one-to-one correspondence does not hold between the left and right image halves, the model naturally explains why humans perceive double (e.g. Panum's limiting case) or multiple (e.g. “Weinshall-stimuli”) instances of the same stimuli in the examples reviewed above. This feature of the model and the fact that no direct assumption is made about surface smoothness (only indirectly via the ordering constraint) also means that the model handles transparency fairly well (possibly in line with human performance), as long as the relative ordering of the stimuli is not broken up too much.

Due to the basic difference in interpretation, and implementation, of the uniqueness constraint compared to the traditional interpretation (Marr & Poggio, 1976), which has been the prevailing one, it is somewhat difficult to make a direct comparison to any previously described model. Apart from the uniqueness constraint, two key features of the model is that it 1) relies on the ordering constraint, and 2) is cooperative.

The ordering constraint has been explicitly used in previous models by for example Baker and Binford (1981) and Ohta and Kanade (1985). Although these models differ in detail, both models find a solution to the correspondence problem by matching the edges (Baker & Binford, 1981) or the intervals between edges (Ohta & Kanade, 1985), in the input stereogram, according to the ordering principle. And moreover, they also share the central idea of utilising edges that run vertically across each image half to guide this process. Neither of these models however address the phenomena of multiple matching, since they were not designed to explain human perception, but designed on rather strict computational grounds. Moreover, Ohta and Kanade use interpolation to determine the disparity at positions where there are no edges available. Consequently, it is un-

likely that their model can handle stereograms that contain transparency. It is unclear how the "Baker -Binford" model would handle transparency.

Three other models that share at least some traits with the one proposed here are the ones suggested by: Marr and Poggio (1976), Pollard et al. (1981) and Prazdny (1985). What all of these models share is that they i) are cooperative, ii) are mainly concerned with the correspondence problem as such, and not the nature (or composition) of the matching primitives, and iii) were all designed - at least in part - to explain human stereopsis, or aspects thereof. Each of these will be briefly discussed below.

The model of Marr and Poggio (1976), while not being the first cooperative one, is probably one of the most well known. The basic idea, or assumption, in their model is that surfaces in general are opaque and cohesive, i.e. smoothly changing in depth. This assumption led to the formulation of the uniqueness constraint, and the cohesitivity (or continuity) constraint; which in their implementation, in turn, were translated into rules stating that matches representing different disparities lying along the same lines-of-sight should inhibit each other (uniqueness), while matches that were close to each other (within some radius) and had the same disparity should support each other (cohesitivity). Their implementation could successfully solve certain types of random-dot stereograms (preferably consisting of smooth opaque surfaces), but it had significant shortcomings. First, their particular interpretation, and implementation, of the cohesitivity constraint (i.e. support between neighbouring matches with the same disparity), makes the model biased towards fronto-parallel surfaces. Consequently, it is not well suited for resolving stereograms of, for example, tilted or jagged surfaces. Another aspect of this shortcoming (discussed earlier in the context of example 3; the Gaussian "needle"), which the authors also pointed out, is that "the width of the minimal resolvable area increases with disparity". This simply means that, for example, a small surface patch in front of a background surface, becomes increasingly difficult to resolve (from the background) as the disparity difference between the surfaces increase.

Finally, as already discussed, their strict formulation of the uniqueness constraint does not allow the model to account for, neither, the phenomena of multiple matching in human perception, nor transparency, since the model can not resolve multiple overlapping surfaces or even represent them.

The PMF model proposed by Pollard et al.(1985) is highly similar to the model of Marr and Poggio (1976) in the interpretation of the uniqueness constraint, and in that neighbouring matches (representing similar disparity values) support each other. However, a major difference is that while the support region, in the latter model, is basically a 2-D disc centred around each match, it is a 3-D space in the former. More specifically, in the PMF model, the width (in disparity) of the support region grows with the distance from a match, and is bounded by the surface were the disparity gradient is equal to 1. The disparity gradient is

defined, given two binocularly visible points in space, as the difference in disparity (between the points) divided by their cyclopean separation (Burt & Julesz, 1980). The choice of using a disparity-gradient limit of 1 was made partly due to the computational argument that binocularly corresponding features in most natural scenes seldom exceed a disparity-gradient of 1, and partly due to the argument that the human visual system seems to have such a limit (Pollard et al., 1985). A major advantage, over using a "flat" support region, is that the PMF model is not biased towards fronto-parallel surfaces, but handles e.g. slanted and jagged surfaces (within the disparity limit of 1) well. However, regarding multiple matching and transparency, the PMF algorithm too falls short; since at any given image point, only the strongest match is kept and all others discarded. Thus, multiple matches and transparent surfaces can not be represented simultaneously. It is unclear if, or how, the PMF model could be modified to handle multiple matching and transparency as well, and at the same time keep its disambiguating power since, for example, in the Panum's limiting case the disparity gradient between the two bars is exactly 2, and hence well beyond the imposed limit. See also Weinshall (1993) for why the PMF model fails to account for her results.

Prazdny's (1985) model does handle transparency and can (if modified), to some extent, also account for the results in Weinshall's studies (see Weinshall, 1993). The disambiguating power of Prazdny's model has the same basic motivation as the previously two described models, i.e. that the disparity difference between neighbouring points on a surface, in general, will be small. In his model, matches support each other according to a Gaussian measure that essentially decreases with increasing distance and/or disparity-difference between the matches. The major difference from the previous two models, however, lies in that Prazdny does not assume opaque surfaces, and therefore there is no (explicit) penalty, or inhibition, in his model between matches that represent different disparities. This allows the model to represent, and resolve, stereograms that contain transparent surfaces, since each surface can be said to support itself without any interference from possible others. On the other hand, the model does not allow multiple matches at any single image point. If there are still ambiguous matches left at any image point, after the matches with the strongest support have been selected, these are simply merged into one and the disparity is determined by interpolation. Thus, in its original form, the model can not account for multiple matching.

Also worth mention is that although there are no explicit inhibitory connections in Prazdny's model, there are indeed such at the effective level (or level of implementation); both in the process of choosing the supporting matches (i.e. given any match, i, only the best supporting match, j, contributes to the activity increment of i), and in the final selection of the matches that have the highest activity. From a computational perspective it changes nothing whether you

choose to call an operation “selection” or “inhibition” as long as it performs the same function, but the point here is that, from a biological perspective, it does matter how such a selection processes could be implemented in neural circuitry. What perhaps is particularly questionable (from this perspective) is why *only* the best supporting match should contribute to any given match's activity level, and how this could be neurally implemented. Considering any possible match in a dense random-dot stereogram, it is difficult (but of course not impossible) to see how a neural mechanism could be so finely balanced as to, from a set of many possible matches with similar support values, let through only the support from the strongest, and at the same time block the support from all others. However, this critique may not be serious to Prazdy's model, and it certainly does not diminish the elegance of it, but it certainly would be interesting to see if a more biologically oriented implementation could perform equivalently.

From the perspective of human stereopsis, a more general critique can be directed, not only to the three models discussed above, but to all models that rely on surface smoothness for disambiguating power. While it is true that surfaces in general are smoothly changing in depth, and that this clearly can be a powerful constraint; it is not necessarily the case that the human visual system make use of this constraint, at least not at the level where stereopsis is achieved. Surface perception (reconstruction) does not necessarily have to be as tightly coupled, or integrated, with stereopsis as these models suggest, but it is quite possible that these are fairly separate processes. The type of stimuli used in example 6 (figure 14), the random-disparity random-dot stereogram (RDRDS), should be a good probe for testing whether human stereopsis is actually biased towards producing matches that appear as cohesive surfaces, even when no obvious surface cues are available. If we are not so biased then there is no reason to believe that it is an important constraint in human stereopsis.

7.2 Possible neural mechanisms

In the current model and implementation, a number of important simplifications have been made, particularly regarding the preliminary matching stage, which makes it difficult to say to what degree it can be considered a model of human stereopsis. Needless to say, however one chooses to look at it, the model lacks too many of the known features of human stereopsis to be called a complete model of human stereopsis. On the other hand, despite the many simplifications made, the close agreement between the model output and the reviewed psychophysical data does seem to suggest that - while the model as a whole may not map onto human stereopsis - the proposed mechanism (i.e. the dual inhibitory near-far AND-gates) may very well have some correlate in human depth perception.

Exactly what this correlate might consist of, or correspond to in human vision, is of course more difficult to specify. It is however tempting- perhaps danger-

ously so - to make the association from the Near and Far AND-gates proposed above, to the Near and Far cells described by Poggio and Fisher (1977). Could these two mechanisms have anything in common other than parts of their names?

By measuring the impulse activity of cells in foveal striate (A17) and prestriate (A18) cortex of the Rhesus monkey, Poggio and Fisher (1977) classified neurones, depending on their response to binocular stimuli, into four different categories: *Tuned Excitatory*, *Tuned Inhibitory*, *Near* or *Far*. The Tuned Excitatory cells were ocularly balanced and gave an excitatory response over a narrow range near the fixation distance. The Tuned Inhibitory cells were ocularly unbalanced and responded to stimuli from the dominant eye over a wide range of disparities except near the fixation distance. The Near cells were also (predominantly) ocularly unbalanced and responded to stimuli, over a broad range, in front of the fixation distance, but not at, or beyond it. Finally, the Far cells mirrored the Near cells, in that they were ocularly unbalanced and responded to stimuli over a broad range, but differed in that they only responded to stimuli beyond the plane of fixation.

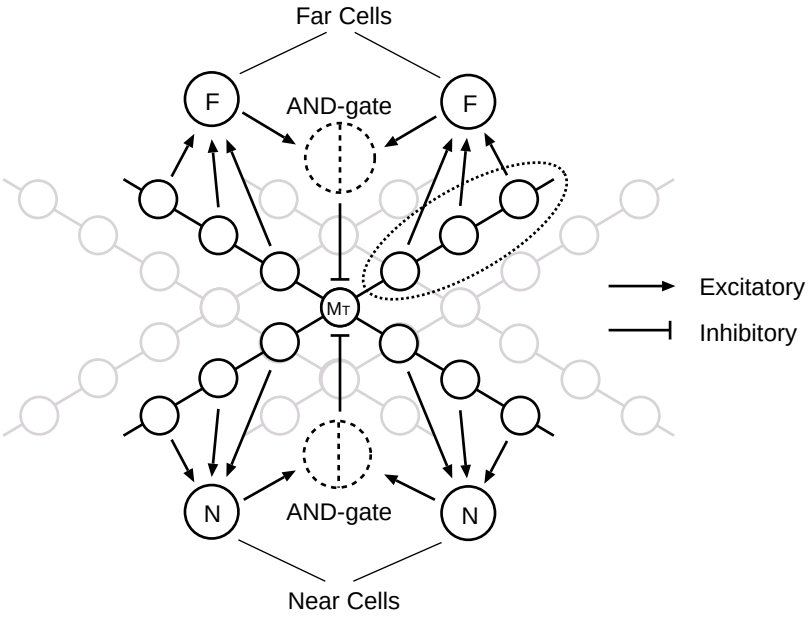

Figure 18. Schematic view of how two Near, and two Far cells, could be arranged to subserve inhibitory cells with possible AND-gate properties.

The Near and Far cells have two properties that, in combination, makes them seem particularly fit as components in a neural implementation of the proposed Near/Far AND-gate mechanism. First, they only respond to stimuli that lies either in front of (Near cells), or behind (Far cells) the plane of fixation. Second, they respond to stimuli over a broad range of disparities, which may suggests that they receive their response properties by summation of a number of subunits sensitive to different disparities. Given these two properties it is not difficult to imagine how a group of two Near, and two Far, cells could subserve one AND-gate each, as depicted in figure 18. What is missing in the data from Poggio and Fisher's study is any direct evidence of cells with response properties like the proposed AND-gates themselves. However, this is not surprising since they did not use ambiguous stimuli in their study. Using only a single bar, or some similar stimuli, there will never be any ambiguity and hence never any simultaneous activation of either two near, or two far, neurones that could drive the postulated AND-gates. If any direct evidence of these AND-gates is to be found, some ambiguous, preferably repetitive, stimuli will have to be used.

Finally, regarding the fact that the great majority of the Near and Far cells were ocularly unbalanced; it seems that such a distinguishing feature should reflect some major functionality of the cells. Unfortunately, the study of Poggio and Fisher did not reveal sufficient details about the nature of this property to allow for any (here) meaningful speculation of how it could fit with the proposed model, but it is nevertheless worth mentioning that several different reconcilable interpretations are possible.

References

Marr, D & Poggio, T. (1976). Cooperative computation of stereo disparity. *Science, 194*, 283-287

Mayhew, J.E.W. & Frisby, J. P. (1981). Psychophysical and computational studies towards a theory of human stereopsis. *Artificial Intelligence, 17*, 349-387

Baker, H.H. & Binford, T.O. (1981). Depth from edge and intensity based stereo. *In Proceedings of the 7th IJCAI* (Los Altos, CA: William Kaufmann), 631-636

Burt, P. & Julesz, B. (1980). Modifications of the classical notion of Panum's fusional area. *Perception, 9*, 671-682

Pollard, S. B., Mayhew, J. E. W. & Frisby, J. P. (1985). PMF: A stereo correspondence algorithm using a disparity gradient limit. *Perception, 14*, 449-470

Krol, J. D. & van de Grind, W. (1980). The double-nail illusion: experiments on binocular vision with nails, needles, and pins. *Perception, 9*, 651-669

Weinshall, D. (1991). Seeing “ghost” planes in stereo vision. *Vision Research, 31*, 1731-1748

Weinshall, D. (1993). The computation of multiple matching doubly ambiguous stereograms with transparent planes. *Spatial Vision, 7*, 183-198

Marr, D. & Poggio, T. (1979). A computational theory of human stereo vision. *Proc. R. Soc. Lond. B. 204*, 301-328

Marr, D. (1982). *Vision*. New York: W. H. Freeman and Company.

Nakayama, K. & Shimojo, S. (1990). Da Vinci stereopsis: Depth and subjective occluding contours from unpaired image points. *Vision Research, 30*, 1811-1825

MaKee, S. P., Bravo, M. J., Smallman, H. S. & Legge, G. E. (1995). The "uniqueness constraint" and binocular masking. *Perception, 24*, 49-65

Akerstrom, R. & Todd, J. T. (1988). The perception of stereoscopic transparency. *Perception & Psychophysics, 44*, 421-432

Ohta, Y. & Kanade, T. (1985). Stereo by intra- and inter-scanline search using dynamic programming. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-7*, 139-154

Prazdny, K. (1985). Detection of binocular disparities. *Biological Cybernetics, 52*, 93-99

Poggio, G. F. & Fisher, B. (1977). Binocular interaction and depth sensitivity in Striate and Prestriate Cortex of behaving Rhesus monkey. *Journal of Neurophysiology, 40*, 1392-1405

Paper IV

The Perception of Binocular Depth in Ambiguous Image Regions

Toward a Computational Theory of Surface Perception

Jens Månsson Christian Balkenius

Lund University Cognitive Science
Kungshuset, Lundagård
S-222 22 LUND, Sweden

Abstract — The perception of binocular depth in image regions that lack explicit disparity information was investigated using sparse random-dot stereograms. The basic design consisted of two different depth planes; a foreground that covered the entire scene, and a background that covered only half the scene; from the centre to either the left, or the right, end of the display. Despite the fact that the foreground covered the entire scene, subjects typically reported that the ambiguous image regions, in between the foreground dots, belonged to the background. When, however, a few unpaired dots were added along the centre, to suggest an occluding opaque surface, subjects tended to perceive the same ambiguous region as belonging to the foreground, which suggest interaction between the binocularly paired, and unpaired, stimuli. A number of variations, of this basic theme was investigated, including: changing the density of the dots in the two depth planes; changing the number, and positioning, of the unpaired dots along the centre; using other cues, e.g. a 2-D contour, or a pair of "Kanizsa"

inducers, to suggest occlusion. Our results can not be accounted for by any simple disparity interpolation scheme, but seem to require additional processing within the disparity domain, as well as interaction with processes devoted to the identification of occluding boundaries.

1 Introduction

Neurones in the primary visual cortex are typically tuned to stimuli of a particular frequency, orientation and disparity (Hubel & Wiesel, 1962), and respond strongly to edges and other local luminance features, but only weakly, or not at all, to light that is evenly distributed within their receptive fields. Consequently, this early cortical region seem to produce something like a sketch drawing of the objects and surfaces that are projected onto the retinas. Remarkably, despite this dramatic reduction of information, we do not usually perceive objects as wireframes, or as clouds of image primitives hanging freely in space; but usually such features are integrated, by the visual system, into cohesive surfaces, where depth appear to vary smoothly over the "empty" image regions in between edges and other isolated image features.

Several previous studies have addressed the question of how depth is filled-in, by the visual system, in ambiguous image regions. Collett (1985), for example, used a random-dot stereograms where one half-image contained a blank region. He found that the perceived depth in this region was, not only, determined by the disparity of neighbouring flanking image regions, which contained explicit disparity information, but was also affected by the surface orientation at these surrounding regions. He also found evidence that depth can be extrapolated into a blank image region, from a single flanking binocularly defined sloping surface.

Evidence for disparity interpolation have also been provided by Würger and Landy (1989), who used stereo displays with a random-dot background, on which a rectangle with uniform luminance was drawn, which vertical edges had slightly different disparities. They found that the perceived depth, at a probe location within the uniform rectangle, varied relatively smoothly with the distance from the vertical edges, and was essentially consistent with linear interpolation of the depth at the edges.

Using a slightly different approach, Blake and Yang (1995) explored how accurately the peaks of two narrow (band-like) surfaces, which depth profiles were defined by a Gabor-function, could be aligned. In one profile, depth varied smoothly, while in the other it was only periodically sampled so that the peak was not explicitly defined by disparity. They found that up to a sampling period of approximately 0.3 degrees, the two peaks could be accurately located relative each other, but that performance degraded abruptly for larger sampling periods. Blake and Yang interpreted the accuracy in performance, below this limit, as evidence that the sparsely sampled surface had been reconstructed; i.e. that depth was interpolated across the gaps. It is noteworthy, however, that the gaps in the periodically sampled profile were not blank, or empty, but had a random-dot texture with zero-disparity, which made them form a background plane that could be seen through the gaps.

These and other (Buckley et al., 1989; Wilcox, 1999; Likova & Tyler, 2003) studies strongly suggests that the processing of disparity information, both directly and indirectly through inter-/extrapolation, plays a central role in the reconstruction of surfaces. That there is more to surface perception than disparity processing, however, seems obvious considering that (phantom) surfaces can be induced by interocularly unpaired retinal stimuli alone (Lawson & Mount, 1967; Nakayama & Shimojo, 1990; Liu et al., 1994). That binocularly unpaired stimuli typically induce illusory contours, and are strongly associated with surface discontinuities, is not surprising considering that they predominantly arise due to half-occlusion (see Nakayama & Shimojo, 1990). That a few interocularly unpaired image components, inserted to suggest occlusion, dramatically can alter the perceived (binocular) depth of a stereoscopically viewed scene have been pointed in a number of previous studies (Ramachandran & Cavanagh, 1985; Anderson, 1994; Nakayama, 1996; Anderson, 1998).

An example of this effect is illustrated in figure 1. In both stereograms the binocular disparity content is identical, and contain a background plane defined by the outlined dots, and a foreground plane defined by the black dots. In figure 1b, however, small sections of the outlined dots, in the background, have been erased to suggest an occluding disc-shaped

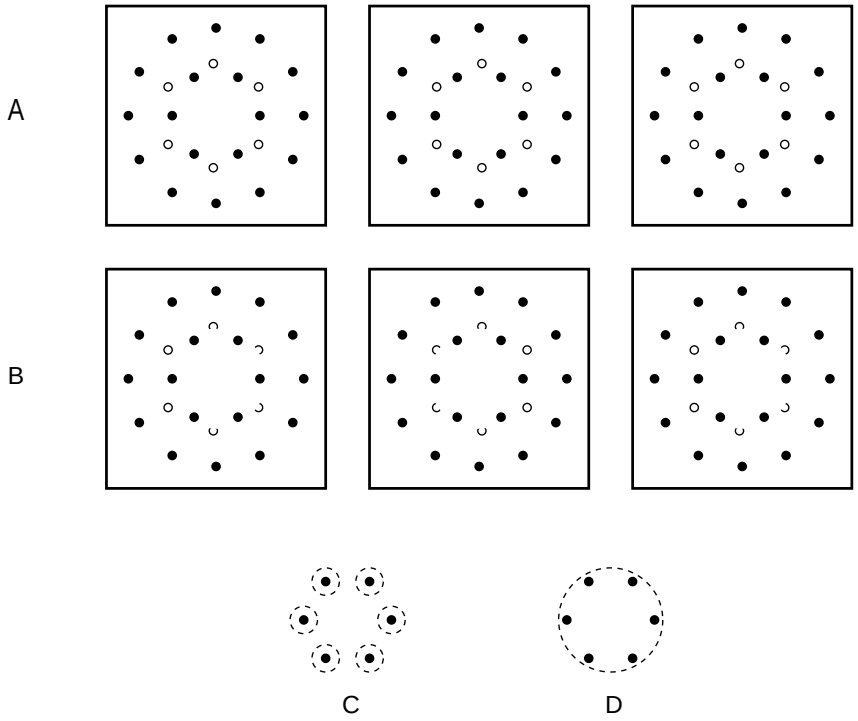

Figure 1: A & B. Example of how the depth in an ambiguous image region depends on the interaction between disparity and occlusion information. The stimulus can be viewed either by crossing the eyes and fusing the right and middle image, or by focusing behind the image and fusing the left and middle images. See text for explanation. C & D. Schematic illustration in the different surface completions in A and B.

surface in the foreground. When correctly fused, the difference in perceived depth of the ambiguous central white region, is striking. In figure 1a, the central white region is perceived as belonging to the background, and the black dots in the foreground are perceived as isolated islands, as schematically depicted in figure 1c. In figure 1b, however, the central ambiguous white region is perceived as belonging to the foreground, and forms a distinctly opaque disc together with the neighbouring black dots, as depicted in 1d.

In previous examples of stereo capture (Ramachandran & Cavanagh, 1985; Ramachandran, 1986; Häkkinen & Nyman, 2001) it has been demonstrated that a few isolated unpaired image components, marking the boundaries of an illusory surface, can produce discontinuities in the binocular matching of a periodic patter; so that the components of the pattern are matched at one depth outside the illusory border, and at another (closer) depth within the enclosing illusory border.

The example in figure 1 differ slightly, but significantly, from such instances of stereo capture, in that there is no ambiguity, whatsoever, with respect to the binocular correspondence of the visible image components. Consequently, the difference in perceived depth of the central ambiguous region in figure 1a and b can not be attributed to any difference in how the individual dots are binocularly matched in the two conditions. Considering that the only difference between figure 1a and b is whether small sections of the outlined dots are erased, or not, it appears obvious that the difference in perceived depth must be attributed to these section being interpreted as evidence for an occluding surface. What is, perhaps, most interesting about the central region, in figure 1b, however, is that the perceived depth is not confined to the illusory boundary that is created by the missing sections, but appears to cover, or spread into, the whole region. One conceivable explanation for this would be that the depth induced by the unpaired sections, is interpolated across the region. This explanation feels akward, however, since interocularly unpaired stimuli predominantly provide qualitative (however see Liu et al., 1994; Gillam & Nakayama, 1999) depth information, which clearly is unsuitable input to an interpolation mechanism. A more reasonable explanation, considering how smoothly the illusory edges, the black dots, and the "empty" central region are integrated into an opaque surfaces, which has the same depth as the black dots, is that the perceived depth of this region is a result of interactions between the processing of the unpaired stimuli (illusory boundaries), and the processing of the binocular disparity content.

In order to explore, in a more systematic manner, how binocularly paired, and unpaired, stimuli interact to form the surfaces we ultimately perceive; we designed a random-dot stereogram, in principle similar to

the one in figure 1; were the perceived depth of an ambiguous image region could be manipulated depending on the amount of unpaired dots along a (virtual) boundary, and the density of the dots in two different depth planes.

The basic design of the stimulus (see figure 2) consisted of two different layers of randomly distributed dots. The dots in the foreground layer had zero disparity, and were randomly positioned anywhere within an enclosing quadratic frame. The dots in the background were also randomly distributed, but within a smaller rectangular region that spanned only half of each image; i.e. from left to the centre, or vice versa, from right to the centre. In this display, the empty image regions, on the side that contain no background dots, have ambiguous depth, and can be perceived to be part of either the background, or the foreground.

In this basic version of the stimulus (bottom stereogram, figure 2), however, most people report that this ambiguous region belong to the

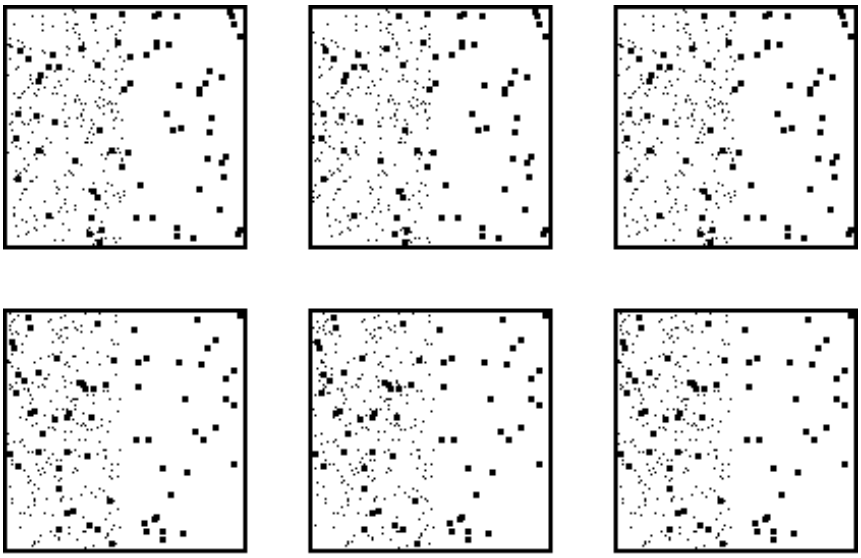

Figure 2: Example of the stimulus used in the experiments. In the actual experiment, the background and forground dots had the same size, but were differently colored (black/white) against a gray fond.

background, which suggests that the depth of the background dots is extrapolated across this entire region, causing the foreground to be perceived as a (binocularly) transparent plane.

This is in itself remarkable since it suggests that extra-/interpolation in a background plane can override the contrasting information that the foreground dots provide, despite the fact that these are much closer than the inducing background dots.

The perceived depth of this ambiguous region is, however, easily manipulated by, for example, adding a few binocularly unpaired dots along the (virtual) boundary where the background dots end. In this case (top stereogram, figure 2), most people report that the ambiguous region belong to the foreground, and that it creates a distinctly opaque surface together with the foreground dots. That is, the foreground appears divided into one transparent half (on the side where dots can be seen in the background), and one opaque half.

Since the resulting percepts in the two conditions are qualitatively very different from each other, and since the ambiguous image regions, essentially, is forced to take on the depth of either the background, or the foreground, no explicit depth probe is needed to determine how depth in the ambiguous region is interpolated, but the sensation of opacity/transparency of the foreground can in itself be used as an indicator of this.

One of the things we wanted to investigating with this display was to what extent the density of the dots in the foreground determine whether the ambiguous region is perceived as opaque, or transparent. The answer to this question is interesting, not least, from a computational perspective. In virtually all computational models of stereopsis, it is assumed that surfaces, in general, are opaque (however see Prazdny, 1985) and relatively smooth. Typically this assumption is translated into binocular matching constraints (Marr & Poggio, 1976; Pollard et al., 1981) that enhance the strength of neighbouring matches that have similar disparities. If such constraints are in fact utilized by the visual system, it would suggest that a densely textured region would be more likely to be perceived as a coherent opaque surface, than an image region with a low density, where

the individual texture components would receive less support from each other, and hence would be more weakly connected.

In the first experiment, we manipulated the density of the foreground and background dots too see how this would influence the perception of opacity of the ambiguous image region. In addition, we changed the number of unpaired dots to investigate how this would interact with the dot density. In the following two experiments, we investigated the role of the placement of the unpaired dots. This was followed by an experiment where we looked at the how the perceived depth of the ambiguous region varied depending on whether the background was defined by dots on one, or both sides of it. Finally, we investigated more explicitly the role of contours induced by different types of stimuli.

2 General Methods

2.1 Participants

The participants were recruited through posters in the university area and were mainly students. They were screened for stereo vision before they were allowed to participate in the experiment.

Before performing the actual tests, each participant was shown a series of examples of the basic stimuli (figure 2), where the type of the stimuli (opaque or transparent) alternated between each trial. In the opaque condition, there were between 8-12 unpaired dots present in the central zone. Participants were instructed to see the stimuli as a window they were looking out from, with the foreground dots lying on the glass of the window. The participants were then asked whether the left and right half foreground of the window appeared any different in the two conditions. A majority of participants spontaneously described one half of the window as being opaque, often using terms like “frosted”, “moist”, or simply “less clear”, when the stimuli contained unpaired dots along the centre. Some participants initially expressed an uncertainty of where the difference lay in the two conditions; but after having being explicitly instructed to disregard the colours of the dots, and whether or not there were any visible dots in the background, most participants did see a difference between the opaque and transparent condition. Participants who did not

perceive any difference between the two conditions, after 20 examples, did not participate in the any of the main experiments.

2.2 Apparatus

The stimuli were displayed on a Hitachi CM815ET Plus monitor on which the gamma value had been calibrated to make a fine raster, of alternate black and white pixels, appear the same shade of grey as a homogeneous surrounding intermediate grey.

The stimuli was viewed through a double mirror stereoscope, mounted on the monitor, from a viewing distance of approximately 35 cm. The size of 1 dot on the screen was 2.95 arc min, and the full width, (Fig. 3), of one half image of the stereogram including the frame was 5.9 deg. The screen was set to a colour temperature of 9600K.

2.3 Stimuli

The basic stimuli used in all experiments were constructed, according to a common 3-D layout (Fig. 3), consisting of a foreground plane with a dot density of 1.5%, surrounded by a 2 pixel wide frame, and a background plane, extending over only half the image, also with dot density of 1.5% and a disparity of 4 pixels. The same dot size (=1 pixel) was used for both the foreground and background in all experiments. To minimize possible interference from incorrect matching of dots in different planes, the foreground dots and the frame were black, and the background dots including possible monocular dots were white on half of the trials. On the other

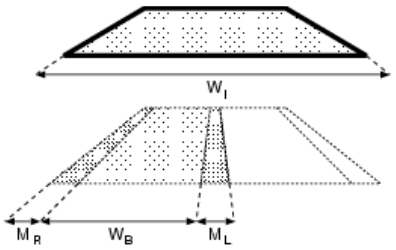

Figure 3: The general structure of the stimuli used in the different experiments. See text for explanation.

half of the trails, the colours were reversed. The colour between the dots and outside the frame was a constant intermediate 50% grey.

Another purpose of alternating foreground and background colours was to reduce potential learning effects. To further reduce learning effects, the side where the background dots were presented was also changed throughout the experiments. In the following description of the stimuli, unless explicitly stated, the background dots are assumed to span the left side.

In both the transparent and opaque condition, there were always right image unpaired dots present within a 4 pixel wide monocular zone, M_R, at the left end of the frame (Fig. 3), but only in the opaque condition were there also left image unpaired dots present in the central 4 pixels wide monocular zone, M_L.

The disparity of the background dots was evenly divided between the left and right images, so that the texture boundary created by the dots and the blank right region appeared binocularly centred at half the total image width. To make the texture boundary appear centred even when monocular dots were present in the central zone, the width of the background region, containing binocular dots, was slightly reduced so that the average width of the left and right image regions, containing dots, equalled half the total image width.

The positions of the left image unpaired dots, within the central zone, were always randomly selected. However, in order to keep control over the exact number of visible unpaired dots, the set of possible positions never included those that were already occupied by an unpaired dot, or a dot in the foreground or the background. In all experiment, the image size of the stimuli, including the frame was 120x120 pixels.

2.4 Procedure

In the main experiments, a two-alternative forced-choice procedure was utilized, in which participants were instructed to push the *down-arrow* button, on a keyboard, if the whole "window" (the foreground) appeared transparent, and the *up-arrow* button if only half the "window" appeared transparent.

Each trial started with a stereoscopic fixation cross consisting of 2 overlapping diagonal lines (41 arc min long) for 500 ms and was followed by the stimulus. After 1000 ms, the word 'half' appeared above the stimulus and the word 'whole' appeared below the stimulus. At this point, the participants would press one of the keys to indicate whether they perceived the whole front surface as transparent or only half. The stimulus remained on the screen until the answer was given, which would blank the screen in preparation for the next trial. An inter-trial interval of 1000 ms was used. Key presses before the initial 1000 ms stimulus presentation were discarded to encourage participants to view each stimulus for at least that time.

3 Experiment 1

In the first experiment, we investigated how the number of unpaired dots, and the density of dots in the foreground, and background, would influence the perception of depth in the ambiguous region. We wanted to test the hypothesis that the probability of perceiving this regions as opaque would be higher with an increasing number of unpaired dots at the centre. Further, we expected the probability of perceiving the ambiguous region as opaque to increase with increased foreground density, and to decrease with increased background density.

3.1 Materials & Methods

There were 10 participants. In total, 384 stimuli were presented to each participant and they had to judge whether the ambiguous part of the image appeared transparent or opaque.

Three different groups of stimuli (of the types shown in figure 3) were used. In stimulus group A, the number of unpaired dots was varied to investigate how it would influence the perception of a surface as opaque or transparent. Each participant were shown 48 different stimuli without unpaired dots and 12 x 8 different stimuli with 1, 2, 3, 4, 5, 6 ,7, 10 or 15 unpaired dots respectively. The dot-density in the the foreground, DF, and the background, DB, were both set to 1.5%. Stimulus group A contained 144 stimuli in total.

In stimulus group B, there were 60 different stimuli without unpaired dots and 60 stimuli with 7 unpaired dots (note that 7 dots, within the central monocular region, corresponds to a density of 1.5%). The density of dots in the foreground layer was set to either 0.375%, 0.75%, 1.5%, 3.0%, and 6.0%. The background density was held constant at 1.5%. 12 different stimuli at each foreground density were presented to the participants. Stimulus group B contained 120 stimuli in total.

Stimulus group C was identical to group B except that the the foreground density was held constant at 1.5%, and the number of unpaired dots M was either zero or 7, while the background density was varied over the set {0.375%, 0.75%, 1.5%, 3.0%, 6.0%}.

The stimuli from groups A, B, and C were mixed and presented in random order. The reaction time for each stimulus was also recorded together with the order in which the stimuli were presented.

3.2 Results

Group A

The average responses of the participants to stimuli in group A were interpreted as the probability that the the ambiguous area would be perceived as opaque given a certain amount of unpaired dots (m). The probability for opacity in each condition is shown in Fig. 4. Logistic regression, using the function

$$p(opacity) = e^{\alpha+\beta m} / \left(1+e^{\alpha+\beta m}\right),$$

gave a best fit with the parameters α=0.211 and β=-0.282. Both α and βm have significant influences on the perception of opacity ($p<0.001$ in both cases). Here, $a>0$ indicates that without unpaired dots, the participants had a tendency to perceive the ambiguous area as transparent rather than as opaque, while $b<0$ shows that the perception of opacity increased with the number of monocular dots.

Group B

The data from stimulus group B where the foreground density was varied is shown in Fig. 5. Without any unpaired dots along the centre, there is little change in the perceived opacity of the ambiguous area when the

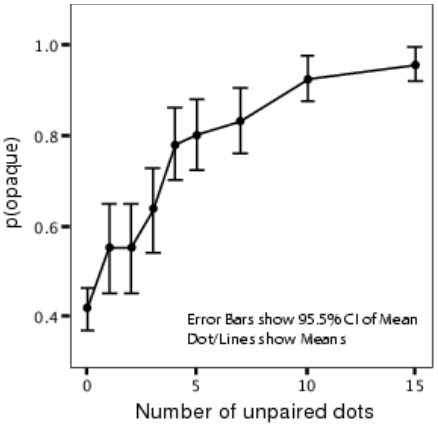

Figure 4: Opacity as a function of the number of unpaired dots in stimulus group A. The error bars indicate the 95% confidence interval around the mean.

foreground density changes. The small increase in opacity seen in Fig. 5 is not significant (ANOVA, p=0.068). A closer look at the data reveals that only the first point with a foreground density at 0.00375 differs significantly from the mean probability at 0.245 (t-test, p<0.01). Excluding this point, we find that the probability of identifying the ambiguous area as opaque is 0.313 in this stimulus group and that the remaining change in opacity as a function of background density is not significant (ANOVA, p=0.913).

With unpaired dots present, the probability of perceiving the ambiguous area as opaque *decreases* with increased foreground density as shown in Fig. 5 (p<0.001). Linear regression gives the approximation p(opacity)=0.914-5.959d, where d is the foreground density. There was a significant interaction between foreground density and the number of unpaired dots (ANOVA, p<0.001).

Group C

The perceived opacity of stimulus group C, where the background density was varied, is shown in Fig. 6. The amount of unpaired dots, and the background density both have significant positive effects on the per-

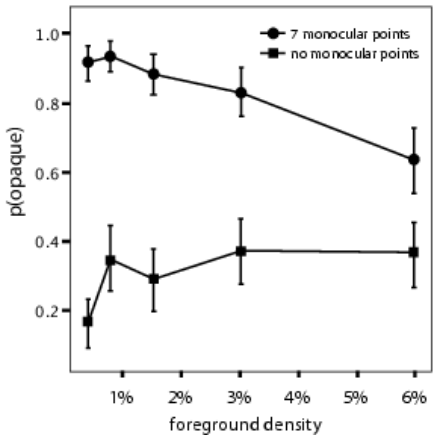

Figure 5: Opacity as a function of forground density, with and without unpaired dots. The error bars indicate the 95% confidence interval around the mean.

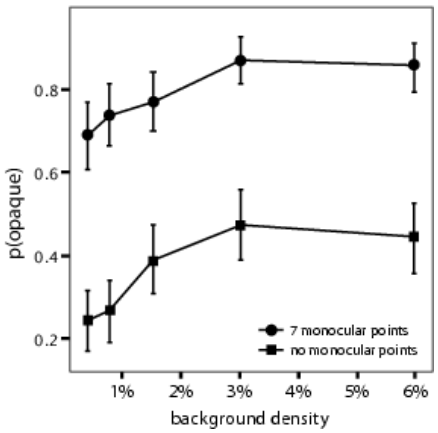

Figure 6: Opacity as a function of background density, with and without unpaired dots. The error bars indicate the 95% confidence interval around the mean.

ceived opacity (ANOVA, $p<0.001$ for both). There was no significant interaction between the number of unpaired dots and the background density (ANOVA, $p=0.241$).

Additional observations

The data was also analysed to test for influences on perceived opacity from other factors in the experiment. These included foreground colour and whether the unpaired dots were presented to the left or right eye. The foreground colour had a significant effect on perceived opacity (ANOVA, $p<0.01$). When the foreground dots were white, the ambiguous area was more likely to be perceived as opaque although the effect was very small ($\Delta p(\text{opaque})<0.05$). Surprisingly, which eye the unpaired dots were presented to also had a significant influence on opacity (ANOVA, $p<0.001$). When the unpaired dots where presented to the right eye, the probability of perceiving the ambiguous area as opaque increased on the average by 0.127. There were no significant interactions between the number of unpaired dots, foreground colour and which eye viewed the unpaired dots.

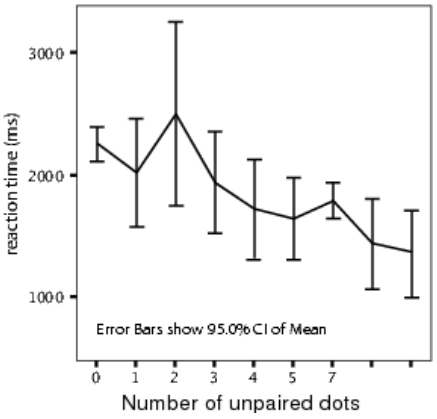

Figure 7: Reaction time as a function of the number of unpaired dots.

As could be expected, the reaction time for each participant decreased during the experiment, but the reaction time was also influenced by a number of other factors. The reaction time decreases with increasing number of monocular dots (ANOVA, $p<0.001$) as shown in Fig. 7. It was also influenced by the foreground colour (ANOVA, $p<0.01$) such that the

reaction time decreased with on the average 376 ms when the foreground was white. A difference was also found for the cases where the participants perceived the ambiguous area as opaque compared to when it was seen as transparent (ANOVA, $p<0.001$); The participants reacted on the average 451 ms faster when the region was perceived as opaque.

3.3 Discussion

The results show that the perception of opacity in the ambiguous region depends on the number of unpaired dots along the centre, as well as the density of the dots in the foreground and the background. The role of these different variables are different however.

With an increasing number of binocularly unpaired dots, present in the central monocular zone, the probability of perceiving the ambiguous image region as opaque increases (Fig. 4). This is exactly what was expected since each unpaired dot is a cue that indicates that an opaque surface is present. Evidently, several such unpaired dots are necessary to suggest an opaque surface. However, above a level of approximately 10-15 unpaired dots, the probability of seeing the ambiguous regions as opaque saturates.

When the density of the dots in the foreground increases, there is no effect on the perception of opacity without unpaired dots (Fig. 5). With unpaired dots present, however, the ambiguous area looks, relatively, more transparent with increasing foreground density. This result clearly contradicts the hypothesis that the components in a high density texture would be more strongly connected, and therefore more likely to be integrated into a coherent opaque surface. One possible explanation for this result, however, may be that the unpaired dots simply becomes relatively less salient due to the increased number of dots in the foreground, and the signal (of occlusion) they provide therefore becomes weaker. Another possibility, if transparency is considered as a surface property, is that the higher density do create a more strongly connected plane/surface, which facilitates the propagation of (the property) transparency, from the unambiguous side to the ambiguous side.

The effect of the background density is different from that of the foreground density in several ways. First, when the background density in-

creases, the ambiguous foreground region becomes more opaque. This again disconfirms the prediction that an increased background density would make the unambiguous area more transparent which in turn would influence the ambiguous area. Instead, the background density works the other way, and does so in a way that does not interact with the number of unpaired dots. Instead, the effect of background density is additive. A possible explanation for this result is that an increased background density produces a more distinct texture boundary at the centre, which in itself may be interpreted as evidence of an occluding surface.

It is interesting to note that the reaction time depends on the number of unpaired dots in the stimulus. This can possibly be explained by the fact that the ambiguity decreases with an increasing number of unpaired dots. This is in line with the observation that the binocular fusion of random dot stereograms, containing an occluding surface, is faster when there are unpaired dots present (Gillam & Borsting, 1988) that are positioned in a manner that is ecologically consistent with the occlusion geometry.

An unexpected result was that when the foreground dots were white, the ambiguous area was more likely to be perceived as opaque than when the foreground dots were black. This suggests that the lightness is somehow a part of this process. This is most likely a monocular effect that interacts with binocular transparency.

An even more surprising result was that the perception of transparency was influenced by which eye viewed the unpaired dots. The ambiguous surface was more likely to be perceived as opaque when the unpaired dots were viewed by the right eye. There can obviously not be any ecological explanation for this effect, and the result therefore seems due to some asymmetry in the experiment, or the observer. Despite an exhaustive examination of the entire experimental setup, we did not find any such asymmetry. Another possibility is that ocular dominance is a factor here, but unfortunately we did not test the participants for this, and and were thus not able to test this hypothesis.

4 Experiment 2 - *Shifted unpaired dots*

The results of experiment 1 clearly showed that the probability of perceiving the ambiguous region as opaque increased with an increasing number of unpaired dots present in the central monocular region. The only reasonable explanation for that the unpaired dots had this effect on the ambiguous region is that these are interpret as evidence for an occluding surface. However, since the reaction time decreased for each participant, one can not rule out that some kind of learning effect interfered with their responses. That is, it is possible that their judgements was based a simpler form of categorization, rather than an actual evaluation of the depth of the ambiguous region. Their judgements could, for example, be based simply on the presence/non-presence of unpaired dots at the centre. If this was the case in experiment 1, the exact positioning of the unpaired dots would not be important, but any placement close to the centre would have been equally effective. To rule out the possibility that the depth judgements were based on such a simple categorization of the stimuli, we essentially replicated experiment 1A with the exception that on some of the trials that contained unpaired dots, the unpaired dots were slightly shifted away from the centre into the binocular region, which contained the background dots. The effect of this was that there were now binocularly matched (background) dots on both sides of the unpaired ones. Such a display is not consistent with an occluding surface and should therefore not cause the ambiguous region to appear opaque, unless judgement is based simply on the presence/non-presence of unpaired dots.

4.1 Materials & Methods

There were eight participants in experiment 2. Four different stimulus types (I-IV) were used in the experiment. In all four types, the foreground and background densities were held constant at 1.5% (the same intermediate level as in experiment 1, group A). In the type-I stimuli there were no unpaired dots. In the type-II stimuli there were 7 unpaired dots within the central monocular region M_L. In the type-III stimuli there were also 7 unpaired dots, but the whole central monocular region was shifted 10 pixels to the left (Fig. 8), so that binocularly fusible (background) dots

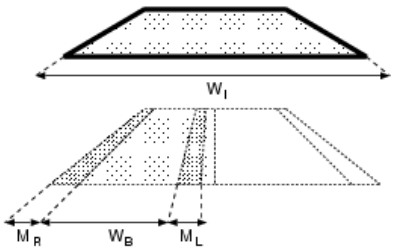

Figure 8: Overview of the stimuli used in experiment 2. The central monocular region ML was left-shifted by 10 pixels.

were present to the right of the monocular region. The type-IV stimuli were identical to the type-III stimuli except that the number of unpaired dots was doubled to 14, compared to 7 for type III. There were 12 trials of each type (48 in total).

4.2 Results

The results of the experiment are presented in Figure 9. The probability of perceiving the ambiguous areas as opaque when seven monocular dots are presented without offset (type II), is significantly higher than for stimulus types I, III and IV (t-test, $p<0.001$ in all cases). There is no significant difference between stimulus types III and IV (t-test, $p=0.25$). Comparing stimulus type I with III and IV shows that there is a significant difference in both cases (t-test, $p<0.05$ and $p<0.001$ respectively).

4.3 Discussion

As expected, the unpaired dots that were placed in the ecologically valid region had a much larger effect on the perceived opacity of the ambiguous region, even compared to the condition when there were twice as many unpaired dots but these were shifted into the binocular region. We interpreted this as evidence that the depth judgements in experiment 1 mainly were due to a global interpretation of the scene, rather than due to some simpler categorization approach, like a simple "feature-detection" of the unpaired dots.

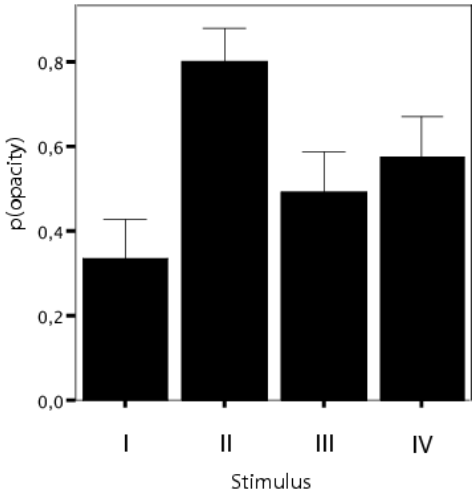

Figure 9: The probability of perceiving the ambiguous region as opaque for each of the four stimulus types: I, no unpaired dots; II, 7 unpaired dots at centre; III, 7 unpaired dots, shifted; IV, 14 unpaired dots, shifted.

5 Experiment 3 - *Randomly Distributed Unpaired Dots*

Contrary to our prediction, in experiment 1, an increased foreground density did not raise the probability of perceiving the ambiguous region as opaque, but rather, in the condition when unpaired dots were present, it made the ambiguous region appear, relatively, more transparent compared to lower foreground densities. With an increasing foreground density, the number of background dots that are "naturally" occluded by dots in foreground will increase. Because such "naturally" (in lack of a better word) occurring dots arise anywhere within the background region, these will most likely be flanked by binocularly matchable (background) dots on both the left and right side. In such isolated instances, the unpaired stimuli is a cue to (local) occlusion, but it is also a cue that the foreground is transparent. If such "naturally" occurring unpaired dots accounted for the result in experiment 1B, they should have the same effect if added at different locations in stereograms with lower foreground densities. This was explicitly tested in the third experiment, where we added different amounts of unpaired dots to the transparent region at random locations.

5.1 Materials & Methods

There were seven participants in experiment 3. Each participants were presented with a total of 200 stimuli. In half of the cases, seven unpaired dots where added along the border between the two regions, and in half of the cases, no unpaired dots where present (Fig. 10). To each stimulus was further added either 0, 2, 4, 8, or 16 unpaired dots, at randomly selected positions anywhere within the half-image that also contained binocularly matchable dots in the background (as indicated by region A in fig. 10).

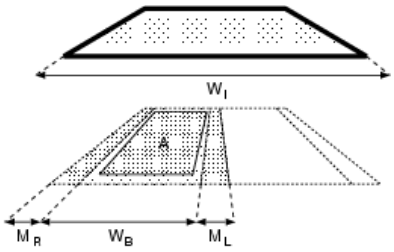

Figure 10: Layout of the stimuli used in experiment 3. The region A contained randomly placed unpaired dots.

5.2 Results

The results of experiment 3 failed to reveal any effect of the added unpaired dots in the transparent region (Fig. 11, ANOVA, p=0.7).

5.3 Discussion

Adding unpaired dots to the half of the image that is always perceived as transparent had no effect on the perceived depth of the ambiguous region. This shows that it is not the increased number of naturally occurring unpaired dots, per se, but some other factor, that causes the foreground to appear more transparent when the density increases, as shown by experiment 1B.

6 Experiment 4 - *Extrapolation vs Interpolation*

The results of experiment 1 showed that with no unpaired dots present, the ambiguous region was more often interpreted as part, or a continuation, of the background plane. This suggests that the depth, defined by

the disparity of the background dots, is extrapolated into the ambiguous region. By increasing the density of the background plane (experiment 1C), our initial prediction was that this would produce a more strongly connected surface that would be more strongly extrapolated into the ambiguous region, which in turn would cause the foreground to appear more transparent. Contrary to our prediction, an increased background density made the foreground appear more opaque. As discussed earlier, however, this result may have been caused by the fact that the texture boundary became more evident with increasing background density, and might have been interpreted as evidence of occlusion. Another, perhaps more obvious, way to induce a stronger background is to let the ambiguous region be flanked by background dots on both the left and right side. If some form of depth propagation occurs in the background plane, from the binocular region into the ambiguous region; then interpolation between two binocular background regions should have a stronger influence on the ambiguous region than extrapolation from a single flanking binocular background region. This was explicitly investigated in experiment 4, were the ambiguous region was flanked on either one, or both, sides by dots in the background.

6.1 Materials & Methods

There were eight participants in experiment 4. Each participant was shown 180 stimuli in total divided equally into three types The stimuli of type I were identical to those in experiment 1A. For type II stimuli, the binocularly matchable dots, in the background, spanned only ¼ of the whole image width (W_I), leaving ¾ of the background empty. Finally, for

Figure 11: Layout of stimulus type I, and type III, in experiment 4. Note that the different widths are not drawn to proportion.

type III stimuli, there were binocularly matchable (background) dots on both the left and right side of the central ambiguous region, both ¼ x W_I wide, and the empty region was centred in the middle (Fig. 11) and was ½ x W_I. To the 60 stimuli of type I, II and III was added either 0, 3, or 6 unpaired dots at each border. This means that for type III stimuli, there were twice as many unpaired dots, in total, compared to stimulus type I and II, since unpaired dots were added to both eyes; left unpaired dots to the left of the empty region, and vice versa right unpaired dots to the right of the empty region.

6.2 Results

The results of the experiment are summarized in Fig. 12. Without any unpaired dots, there was a significant difference in the opacity of stimulus type I and either of II (t-test, $p<0.013$) and III (t-test, $p<0.05$). There was no significant difference between stimulus type II and III. With three unpaired dots, the result was similar. There was a significant difference between stimulus type I and either of II (t-test, $p<0.001$) and III (t-test, $p<0.001$). Again, there was no significant difference between stimulus types II and III. Finally, when six unpaired dots were present, the only

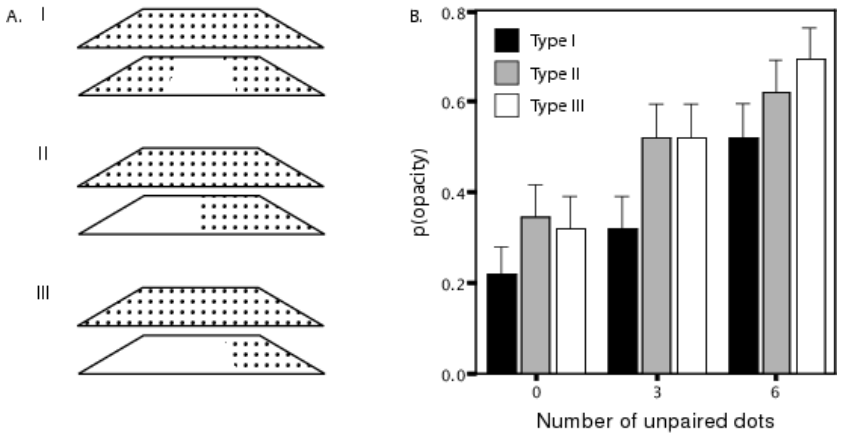

Figure 12: The probability of perceiving the ambiguous region as opaque depending on whether it is flanked by a binocular background region on one, or both sides. Type (I) 0: Both, (II) 1: Half, (III) 2: Quarter.

significant difference was between stimulus group I and III (t-test, $p<0.001$). We also compared stimulus type II and III without unpaired dots with stimulus type I with 3 unpaired dots and found no significant differences. In addition, we compared stimulus type I with three unpaired dots with stimulus type II and III with six unpaired dots. The perception of opacity was significantly different in both cases (t-test, $p<0.001$ and $p<0.001$).

6.3 Discussion

The results of experiment 4 clearly shows that when the ambiguous region is flanked on both sides, by background dots, it is more likely to be perceived as transparent than when it is flanked on only one side. This effect is seen regardless of whether unpaired dots are present or not. It is interesting to note that the ambiguous (foreground) region, in the double-flank (type 1) condition with three unpaired dots on each side, is judged to be equally transparent as the stimuli in the two different single-flank (type II and III) conditions in which there were no unpaired dots. The effect of double flanks is even more pronounced when comparing stimulus type I, with three unpaired dots along *each* border, and stimulus types II, with six unpaired dots along the single border. Although the same number of unpaired dots are present in these two condition, the perceived opacity differs dramatically. This shows that it is not the absolute number of unpaired dots, alone, that determine the perceived opacity of the foreground, but the strength of the extra-/interpolated background is also an important factor. It is also interesting to note that foreground, in the double-flank condition without unpaired dots, is judged to be more transparent than in any other stimuli in this study. The probability of perceiving this stimulus as opaque is approximately 0.2, which is well below that of any other stimuli we used. The results suggest that the increased perception of the transparency, when the ambiguous region is flanked on both sides, is the result of some form of interpolating process connecting the two flanking regions.

7 Experiment 5 - *Contours*

The results of the previous experiments suggest that the placement of the unpaired dots is a highly important factor in determining the perceived depth of the ambiguous region. When the unpaired dots are arranged in a manner that is consistent with an occlusion interpretation, and an "illusory" edge seen at the centre, the foreground is more likely perceived as opaque. When the arrangement of the unpaired dots (experiment 2 and 3) does not support an occlusion interpretation, no "illusory" edge is seen, and the foreground is more likely perceived as transparent. In experiment 6 we explicitly investigated the role of different types of contour inducers.

7.1 Materials & Methods

There were 8 participants in experiment 5. There were six types of stimuli in the experiment with different types of contours at the border between the ambiguous and unambiguous image region (Fig. 13). In stimuli of type I, two black (Kanizsa-like) quarter sections of a disc were placed in the background, at the top and bottom as depicted in figure 13. The radi-

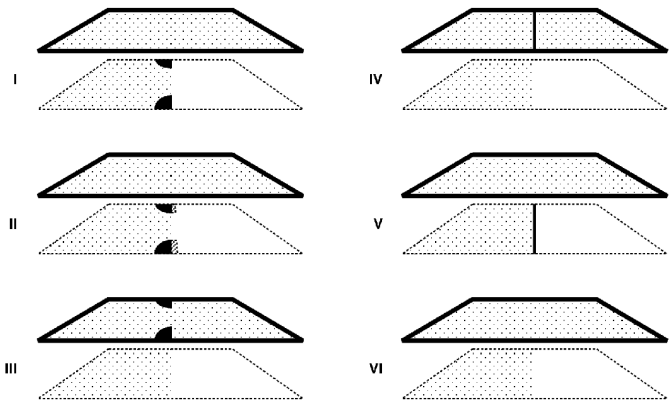

Figure 13: Stimuli layout for the different types used in experiment 6. I: Illusory (2-D) contour in the background. II: Illusory contour in the background, with unpaired sections. III: Illusory contour in the forground. IV: Line in the forground. V: Line in the background. VI: Control.

us of the black disc-section was 10 pixels. The type II stimuli was identical to type I except that part of the discs now extended into the ambiguous region in one eye, but not the other. Type III stimuli used the same Kanizsa-inducers as in the type I stimuli, but they were now positioned in the foreground rather than in the background. Type IV had a simple straight vertical line along the border in the foreground, while type V stimuli had the same line in the background layer. Finally, stimuli of type VI were the control and did not contain any line or contour. Type VI was identical to the stimuli used in experiment 1A, with no unpaired dots present.

7.2 Results

The perceived opacity for each of the stimulus types are shown in Fig. 14. There was no significant difference between stimulus type I and V (t-test, $p>0.05$), but both differs significantly from the control stimulus (t-test, $p<0.001$ in both cases). Stimuli II and IV also differs significantly from the control stimulus (t-test, $p<0.001$ for both). There was no significant difference between stimulus types II and IV. Stimulus type II also differs significantly from the control (t-test, $p<0.01$).

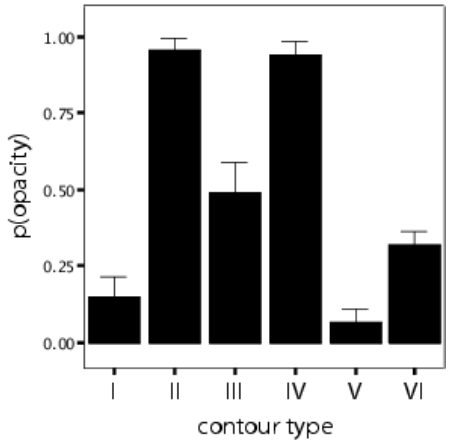

Figure 14: The probability of perceiving the ambiguous region as opaque. I: Illusory contour in background, II: Illusory contour in background, with unpaired sections III: Illusory contour in foreground, IV: Line in foreground, V: Line in background, VI: control stimulus without any boundary, nor unpaired elements.

7.3 Discussion

As expected, the results shows that when the black disc-section are fully visible to both eyes (type I) they do not contribute to making the foreground more opaque. Although the abrupt cut-off of the black disc sectors can be interpreted as a cue to occlusion (in a 2-D scene), it is obvious from the results that when unambiguous disparity information is available, it is used instead and that it overrides the signal that the cut-off discs provide. If the cut-off sections of the discs in fact were caused by an occluding foreground surface, it should have been accompanied by a binocular mismatch of the ends of the sectors as is the case in the stimuli type II. In the stimuli of type II, the disparity information and the signal provided by the cut-off discs cooperate, since they are both consistent with an occlusion interpretation. Consequently, in this condition, the foreground was almost always interpreted as opaque.

Less expected, however, was that the stimuli of type III would make the foreground appear more opaque. At first glance, the stimuli arrangement in this condition does not seem to be ecologically consistent with an occlusion interpretation. In fact, however, it is a perfectly reasonable result, if one assumes that, for example, a white paper with the foreground dots painted on it, were laid out across the discs. In other words, the explanation here is the same as that for the classic (2-D) Kanizsa-illusion, with the only difference that here it is displayed in a 3-D setting.

The perhaps most interesting result of experiment 5, however, was how large effect the placement of a single, binocularly fusible, vertical line had on the perceived depth of the ambiguous region. When placed in the foreground, the foreground was almost always perceived as opaque; and when placed in the background, the foreground was almost always perceived as transparent. Note that there were no unpaired dots present in any of the stimuli in experiment 5. Since the background dots terminate at the centre, in all stimuli we used, there is distinct difference in the 2-D appearance of the left and right half (of each half-image). One side of the background is textured, and one side is not. When the ambiguous, empty, region is perceived as part of the foreground, the lack of background dots in this region is attributed to the occluding foreground.

When, however, the ambiguous region is perceived as part of the background, the lack of dots in this region can only be attributed to that the texture ends, but not necessarily the background surface itself. Possibly, when the vertical line is inserted in the foreground, it works to reinforce this attribution since it suggest that the change, in texture, is caused by a change in the foreground. The line divides the foreground in to two different surfaces, which are more easily attributed with different properties (opaque/transparent). Likewise, when the line is placed in the background, it reinforces the probability that the texture difference is attributed to, or caused by, a change in the background, or rather the property (texture/no texture) of the background surface.

8 General Discussion

The results of experiment 1 showed that the probability of perceiving the ambiguous region as an opaque foreground surface increased with the number of unpaired dots that were added along the centre. Above approximately 10-15 dots, however, the effect saturated, and the ambiguous region was almost always perceived as opaque. Contrary to our prediction, we also found that an increase in the foreground density did not make the ambiguous region appear more opaque. When no unpaired dots were present, in the central monocular zone, changes of foreground density had basically no effect on the perception of the ambiguous region. With unpaired dots present, however, higher foreground densities made the foreground appear, relatively, more *transparent* than lower densities. This result seems counter-intuitive, but may possibly be accounted for simply by the fact that the saliency of the unpaired dots decrease with increasing foreground density. It should be noted that, due to their small size, the individual unpaired dots were very difficult to explicitly locate. Further we found that an increased density of the background dots made the ambiguous region more likely to be perceived as part of the *foreground*, whether unpaired dots were present in the monocular zone or not. Again this result stand in contrast to our initial prediction that a higher background density would more strongly define the background surface, which in turn would facilitate the extrapolation of depth into the

ambiguous region. When the background density increases, the texture boundary that is created between empty (ambiguous) region, and the region containing the background dots, become more distinct. Possibly, this texture boundary is in itself interpreted as a cue to occlusion.

The results of experiment 2 showed how important the exact placement of the unpaired dots is in order for them to make the foreground appear opaque. When the unpaired dots were shifted into the region containing the background dots, their influence on making the foreground opaque was dramatically reduced, even if there were twice as many unpaired dots as in the condition when they were placed at the centre. This clearly shows that it is not the presence of individual unpaired dots, per se, that induce the the perception of opacity of the foreground, but that rather that they have this effect only when they are ecologically consistent with an occlusion interpretation.

In experiment 3, the hypothesis was tested that the increase in perceived transparency of the foreground, with increasing foreground density (experiment 1B), could have been caused by "naturally" occurring unpaired dots. Such randomly occurring unpaired dots are (at least theoretically) a cue to transparency. The results of experiment 3 did not, however, support this hypothesis.

Experiment 4 showed that in a display where the ambiguous region is flanked, by background dots, on both the left and right side, the foreground is more likely perceived as transparent compared to a display where the ambiguous region is flanked on only one side. The results also showed that it took a greater number of unpaired dots to make the foreground appear opaque in the double-flank condition, compared to the single-flank condition.

Finally, in experiment 5, it was demonstrated that a straight vertical line (stimuli IV) that explicitly divided the foreground into two separate regions, was equally effective, in making the foreground appear opaque, as a stimuli (type II) that contained more explicit cues to occlusion, such as end-cuts and unpaired segments. Remarkably, when the straight line was instead placed in the background plane, the foreground was judged to be transparent more often than the control stimuli, which contained no

vertical line, nor any unpaired dots. In all the stimuli we have used in this study, the lack of dots in the empty region can, essentially, only be attributed to two factors; the lack of dots on one side can either be caused by an occluding surface in the foreground, or it can simply be due to the fact that the texture, but not necessarily the background surface itself, terminate at the centre. Our interpretation of the result (of experiment 5, stimuli V) that an explicit (binocularly) visible border, presented at the texture boundary, increase the likelihood of perceiving the foreground as transparent, is that such an explicit border better supports the latter alternative, and hence make the occlusion interpretation less likely.

Due to major differences in the stimuli, method and procedure used in the experiments presented here, it is difficult to compare, other than we have already done, our result to those of previous studies that have addressed binocular depth interpolation. The perhaps most similar, and therefore most relevant, study was performed by Gillam and Nakayama (2002).

In ordinary 2-D displays, a set of abutting collinear line terminators typically induce an illusory contour (see for example Kanizsa, 1974) that follows the line terminations, and that is, locally, perpendicular to the orientation of the lines. When two such set of line terminators meet, the resulting percept of an illusory contour is generally stronger. In natural scenes, such abrupt termination of a pattern, or texture, is usually a strong indication of occlusion. In a 2-D image, however, depth is ambiguous and either set could be occluding the other.

By using stereo displays where the depth of two such sets of lines were defined by binocular disparity, Gillam and Nakayama (2002) investigated the conditions under which illusory contours arise. Two different sets of lines were used. Both sets of lines were made up of randomly oriented straight lines. In one set (they referred to as the *forest*) the lines were also randomly tilted in depth. In the other set (the *plane*) the disparity was constant, and hence made all lines lie in a single fronto-parallel plane.

When the two sets were displayed, one above the other, so that the different sets of line terminators met at the same height, a distinct illusory contour was seen when the forest was further away than the plane, but

not when the plane was further away than the forest. This result is not surprising, considering that a plane can easily occlude a set of isolated objects, but a set of isolated objects (lines) can not occlude a plane; but it effectively demonstrates that illusory contours are not well predicted, in 3-D scenes, by the presence and placement of line-terminators, or other local (2-D) image features, alone. In particular, the illusory contour was not defined, in their displays, by the line-terminators of the occluding set, but by the line-terminators of the occluded set. This became evident when the two sets of lines were made to either overlap slightly, or were slightly separated from each other. In both cases, when the *plane* were closer than the *forest*, a distinct illusory contour was seen were the lines of the *forest* terminated, not were the foreground lines (*plane*) terminated.

Gillam and Nakayama argue that these findings support the view that the perception of illusory contours are intricately linked with the interpretation of occlusion and surface layout, and therefore require an analysis at the surface level of description. In their view, line-terminations and other, local, low-level cues to occlusion do not, automatically, induce illusory boundaries, but do so only in conjunction with surface layouts that support an occlusion interpretation.

We fundamentally agree with this view, and our results seem to suggest that binocularly unpaired elements (experiment 2,3), or other boundary inducers (experiment 5 I & II) are treated no differently than line-terminators in this respect. That is, the presence of such elements do not, per automaticity, induce an illusory contour, or make the ambiguous region appear opaque, but they have this effect only when their arrangement, and the surface layout, is ecologically consistent with an occlusion interpretation.

Turning to consider possible mechanisms that can account for the perceived depth in the ambiguous region, in the basic version the stimuli, we see essentially two different possibilities. A prerequisite in both accounts is that the disparity of the dots in the fore- and background have been accurately computed; i.e. that the correspondence problem, with respect to the available image primitives (the individual dots), have been successfully resolved. This assumption is reasonable, considering that the likeli-

hood for mismatching of non-corresponding dots were low, due to the different colouring of the dots (black in the foreground and white in the background), and due to the relatively low density used. Moreover, it is assumed that a preliminary surface completion process have been carried out, independently in the fore- and background, that involves interpolation and extrapolation of disparity into the empty image regions.

Subsequently, considering first the condition when no unpaired dots are present, one possibility (account I) is that the transparency that is induced on the side where dots can be seen in the background, is treated (by the visual system) as a *property* of the foreground surface, which is assigned/spread to the whole foreground (as schematically depicted in figure 15 Ia). Alternatively, in account II, the perceived transparency of the foreground is not treated as a property of the foreground per se, but arise due to the extrapolation of the background surface, which in turn inhibit the interpolation of disparity in the foreground, and so to speak resolves the foreground surface (figure 15 IIa). Considering, on the other hand, the condition where unpaired dots are present and induce an illusory edge; account I imply that the illusory edge block the spread of (the property)

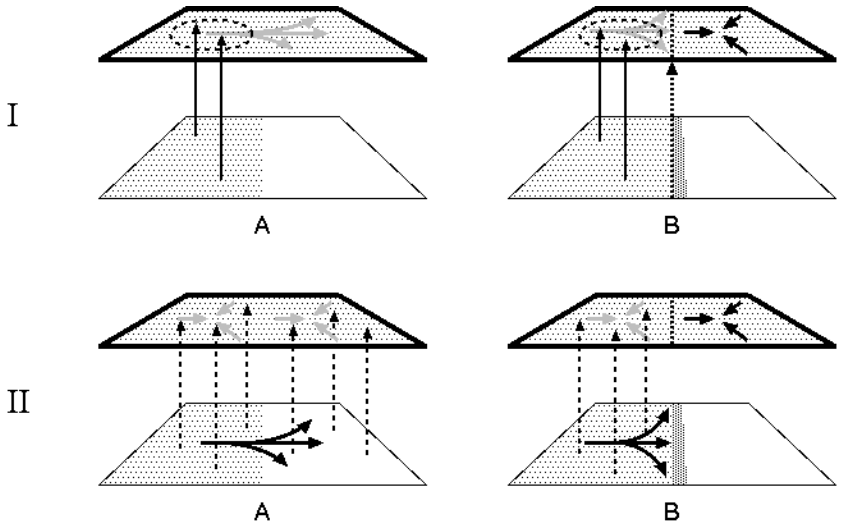

Figure 15: Outline of two possible mechanisms that can account for the percieved depth of the ambiguous region. See text for explanation.

transparency in the foreground (figure 15 Ib), while account II imply that the illusory edge have the effect of blocking the extrapolation of the *surface* in the background, which in turn prevents the background from inhibiting the disparity interpolation in the foreground (figure 15 IIb).

Although these two accounts appear very different from each other, they seem to predict remarkably similar results, and appear able to equally well account for our results. Nevertheless, we strongly favour the second of these accounts, primarily because it appears akward to treat (binocular) transparency/opacity as a property of a surface. Rather, in 3-D displays, opacity and transparency is in itself evidence for the presence/non-presence of a surface. That is, where we see opacity we see a surface, and where we see binocular transparency we see empty space. We believe that one reason why the two mechanisms appear difficult to discern from each other, is due to the arrangement of the stimuli we used. Because the dots are very clearly separated into two distinct depth planes, these are readily interpreted as surfaces, and hence account I seem equally appropriate. In more natural scenes, however, and in the more general case, it is not necessarily this obvious which collections of image primitives are part of the same surface, and which primitives that belong to different surfaces. In fact, this is ultimately the problem that needs to be explained, and it therefore seems akward to involve a *surface property* as part of the explanation for how surfaces are completed. Or, perhaps more simply put, account II appears more natural since it suggests that binocular transparency/opacity is a *result* of the surface completion process, rather than the other way around that ("free-floating") transparency/opacity is *causing* the surface completion.

Placing the mechanism, proposed in account II above, into the broader context we get the following tentative model of binocular depth processing (see figure 16).

First, individual low-level image primitives in the left and right retinal images are binocularly matched, and a preliminary solution to the correspondence problem is found, supposedly in fashion as suggested by conventional (low-level) models of stereopsis; e.g. Marr & Poggio (1976), Prazdny (1985), Månsson (2002). Given this preliminary solution, the dis-

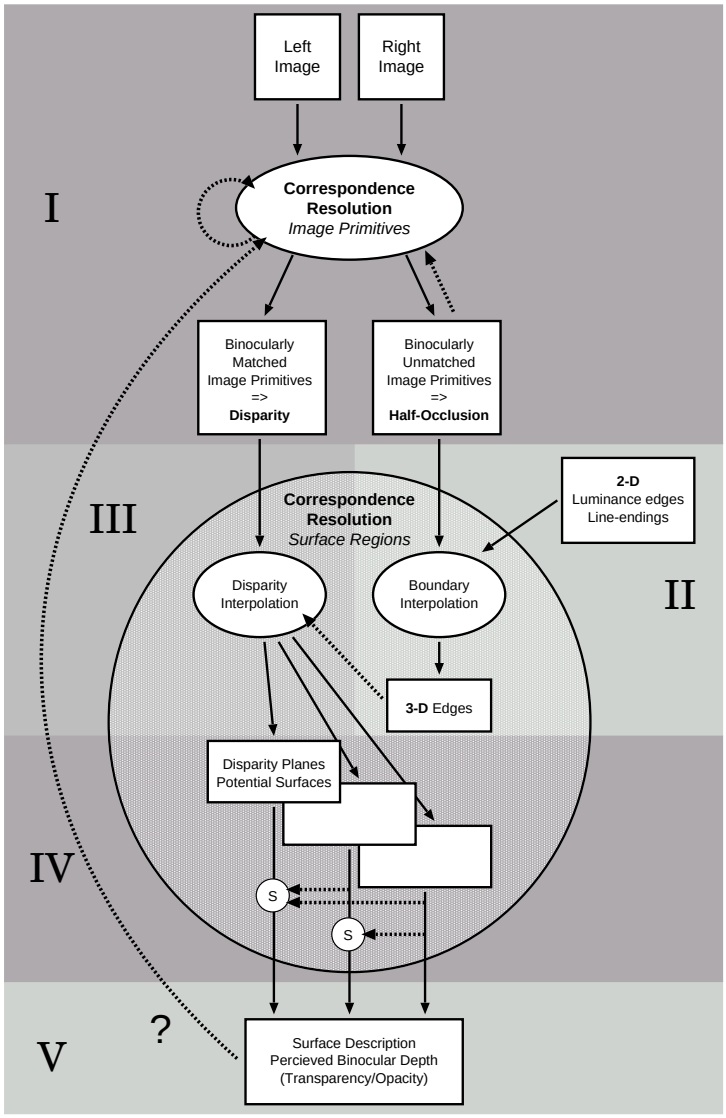

Figure 16: Outline of a model for binocular depth procesing. I: Discrete image primitives are binocularly matched, and unpaired elements are identified. II: Identification of occluding boundaries. III: Preliminary surface completion; inter-/extrapolation of explicit disparity information. IV: Background surfaces inhibit interpolation in foreground layers. V: Rendered (visible) surfaces.

parity of corresponding image elements is computed, and binocularly unpaired image elements are identified.

Second, binocularly unpaired image elements, as well as (2-D luminance) edges, line-terminators, and other low-level cues to occlusion are used to identify potential occluding edges the scene.

Third, the disparity of the binocularly matched image primitives (the output of stage I) is input to a preliminary surface completion process, which within the boundaries defined by occluding edges estimate depth within empty regions, by means of inter-/extrapolation of the explicit disparity information. Likely, this integration process is controlled by some kind of smoothness constraint, so that regions where image primitives have similar, or smoothly changing, disparities are more strongly connected, than regions where image primitives are more scattered in depth. At this stage, multiple possibly overlapping surfaces can co-exist, but do not necessarily correspond to any visible surface.

In the final processing stage, depth that has been filled-in, in empty foreground regions, is cancelled, or suppressed, at any given location where there is also a surface in the background. Explicitly matched image primitives, in the foreground, are not affected. At this stage, any remaining regions with explicit, or implicitly assigned, depth is rendered as visible surfaces; other regions are perceived as void space.

Preliminary work (unpublished) on a computer implementation of the above model, which uses the (discrete primitive) stereo model proposed by Månsson(2002) as a front-end (stage I), suggest that the model handles stimuli (of the type used in Weinshall, 1991) containing binocular transparency better than conventional stereo algorithms.

9 References

Hubel, D. H., Wiesel, T. N., 1962, Receptive fields, binocuar interaction and functional architecture in the cats visual cortex. Journal of Physiology, 160, pp 106-154.

Collett, T. S., 1985, Extrapolating and interpolating surfaces in depth, Proc. R. Soc. Lond. B 224, 43-56

Würger S. M & Landy M. S., 1989, Depth interpolation with sparse disparity cues, Perception, v 18, p 39-54

Buckley D., Frisby J. P. & Mayhew J. E. W., 1989, Integration of stereo and texture cues in the formation of discontinuities during three-dimensional surface interpolation, Perception, v 18, no 5, p 563-588

Yang, Y. & Blake, R., 1995, On the accuracy of surface reconstruction from disparity interpolation, Vision Research, v 35, no 7, p 949-960

Likova L. T. & Tyler, W., 2003, Peak localization of sparsely sampled luminance patterns is based on interpolated 3D surface representation, Vision Research, v 43, 2649-2657

Wilcox, L. M (1999), First and second-order contributions to surface interpolation, Vision Research, v 39, 2335-2347

Mount D. C & Lawson R. B., 1967, Minimum conditions for stereopsis and anomalous countour, Science, v 159, no 3802, p 804-806

Nakayama, K. & Shimojo, S., 1990, Da Vinci stereopsis: depth and subjective occluding contours from unpaired image points, Vision Res. v 30, p 1811-1825

Liu, L., Stevenson, S. B. & Schor, C. M., 1994, Quantitative stereoscopic depth without binocular correspondence. Nature, 367, 66-69

Ramachandran, V. S & Cavanagh, P., 1985, Subjective contours capture stereopsis, Nature, vol 317, p 527-530

Anderson, B., 1994 , The role of partial occlusion in stereopsis, Nature, v 367, p 365-368

Nakayama, K., 1996, Binocular visual surface perception, Proc. Natl. Acad. Sci USA, v 93, p 634-639

Anderson, B. L., 1998, Stereovision: beyond disparity computations, Trends in Cognitive Sciences, v 2 no 6, p 214-222

Ramachandran, V. S., 1986, Capture of stereopsis and apparent motion by illusory contours, Perception & Psychophysics, vol 39, no 5, 361-373

Häkkinen, J. & Nyman, G., 2001, Phantom surface capture stereopsis, Vision Research v41, p 187-199

Gilliam, B. & Nakayama, K., 1999, Quantitative depth for a phantom surface can be based on cyclopean occlusion cues alone, Vis. Res. v 39, p 109-112

Marr, D. & Poggio, T., 1976, Cooperative computation of stereo disparity. Science, 194, 283-287

Pollard, S.B., Mayhew, J.E.W. & Frisby, J.P., 1985, PMF: A stereo correspondence algorithm using a disparity gradient limit. Perception, v14, 449-470

Prazdny, K., 1985, Detection of binocular disparities. Biol. Cybern. 52, 93-99

Gillam, B. & Borsting, E., 1988, The role of monocular regions in stereoscopic displays, Perception, v 17, p 603-608

Gillam, B. & Nakayama, K., 2002, Subjective contours at line terminations depend on scene layout analysis, not image processing, Journal of Experimental Psychology: Human Perception and Performance, v 28, no 1, p 43-53

Kanizsa, G., 1974, Contours without gradients or cognitive contours? Italian Journal of Psychology, 1, 93-112

Månsson, J., 2002, The Uniqueness constraint revisited: A symmetric near-far inhibitory mechanism producing ordered binocular matches. Lund University Cognitive Studies 95

Weinshall, D., 1991, Seeing "Ghost" planes in stereo vision. Vision Research, V 31, No 10, pp 1731-1748