STEREOVISION: A MODEL OF HUMAN STEREOPSIS

Jens Månsson

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

jens.mansson@fil.lu.se

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/throughs), 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 *accomodation* 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 accomodation 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

Lund University Cognitive Studies — LUCS 64. ISSN 1101 — 8453. © 1998 by the author.

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*.

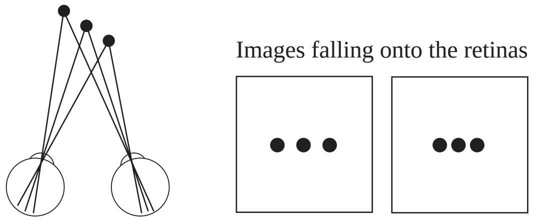

Figure 1. Due to the diference in perspective, the images of the dots will fall onto slightly different locations in the two retinas.

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 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.

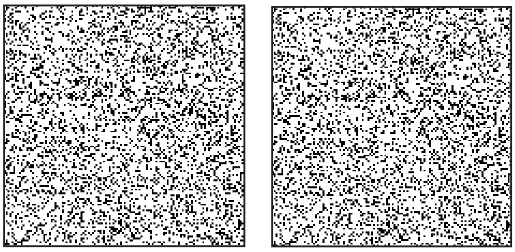

Figure 2. A random-dot stereogram contains no monocular depth-cues. The 3-D structure hidden in the stereogram can only be percieved when the images are binocularly fused in a stereoscope or by crossing the eyes.

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. Finally, 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.

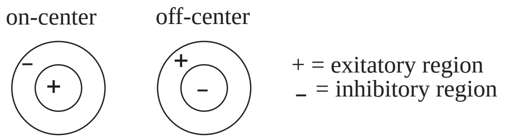

Figure 3. Receptive field mapping of the retinal ganglion cells.

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'.

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).

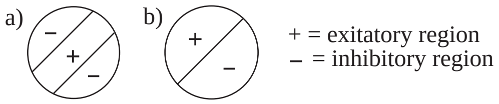

Figure 4. Receptive fields of two typical simple cells.

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?

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 $\nabla^2 G$, 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

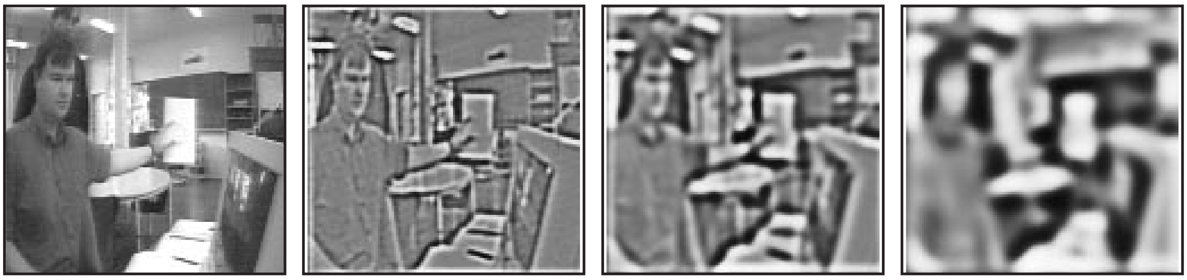

Figure 5. (a) Showing an image (128x128 pixels) and the results after having convoluted the image with the ∇^2G-operator. The space constant σ has the values of 1, 2 and 4 pixels in (b), (c) and (d) respectively.

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.

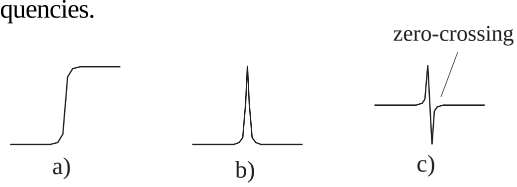

Figure 6. A change in light intensity (a) will rise to a peak (b) in its first derivative, and to a zero-crossing (c) in its second derivative.

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 “throughs” in the convoluted images should be matched as well. In this context, peaks and throughs refers to the maximum and minimum values in the convoluted image (Fig. 6c).

No doubt, the information corresponding to peaks/throughs 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/throughs, 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 termsof 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.

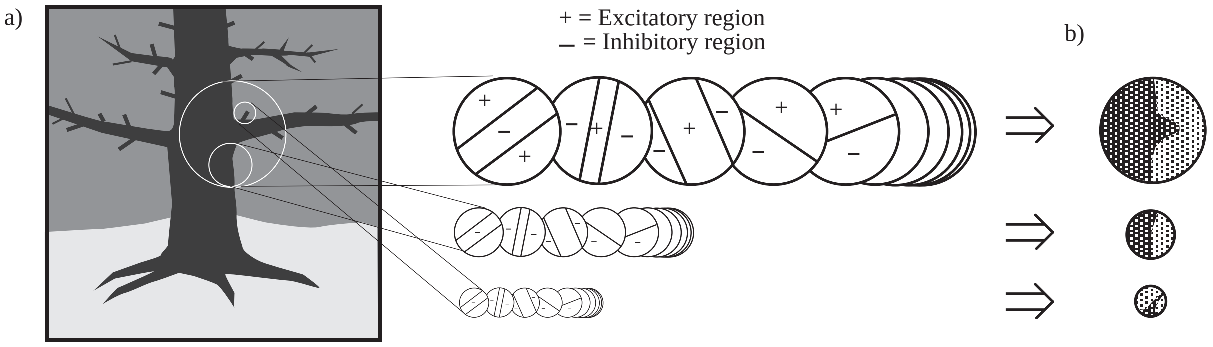

Figure 7. (a) Schematic organisation of the suggested groups of simple and complex cells, showing the sizes and compositions of these cells' receptive fields. 7 (b) is supposed to illustrate how the overall configuration of contrast, within the receptive fields, could be reconstructed from the "superimposed"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

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 interpreted 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.

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.

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.

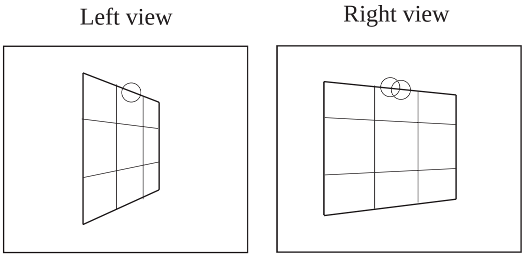

Figure 8. Due to the orientation of the surface, and the difference in perspective, the light from the marked edge will be projected onto regions of different sizes in the two retinas.

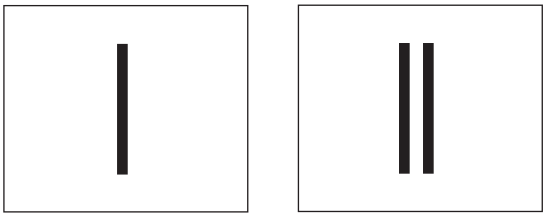

Figure 9. Illustrating *Panum's limiting case*. The bar in the left image can be matched with either of the two bars in the right image. When fused the experienced percept is that of two separate bars, hovering at different depths.

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 relative 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 togheter 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 computational 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.

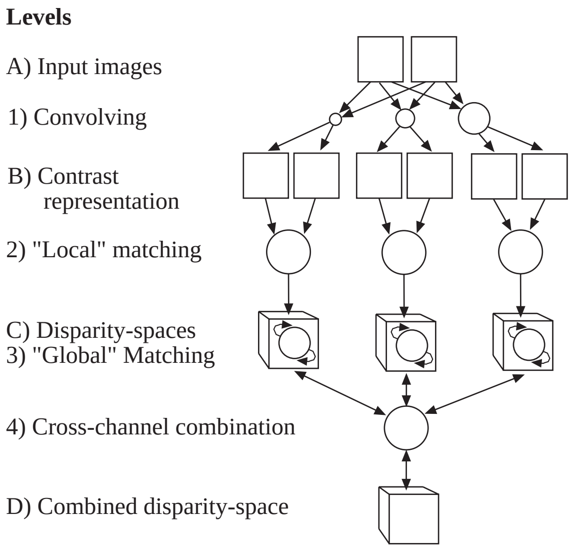

Figure 10. Schematic overview of the different levels of representation and processing. Representational states are shown as squares/cubes and are labeled with letters (A–D). Processing stages are displayed as circles and are labeled with numbers (1–4). (A) The input stereogram. (1) Each image is convoluted 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”).

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.

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.

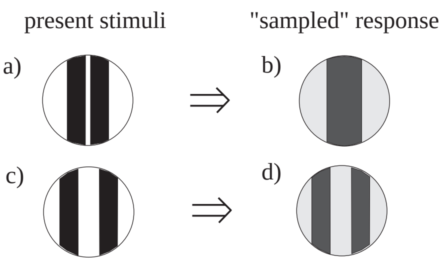

Figure 11. (a & c) Present stimuli within the receptive field of a group of simple and complex cells. (b & d) Showing the (assumed) "sampled" response.

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.

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

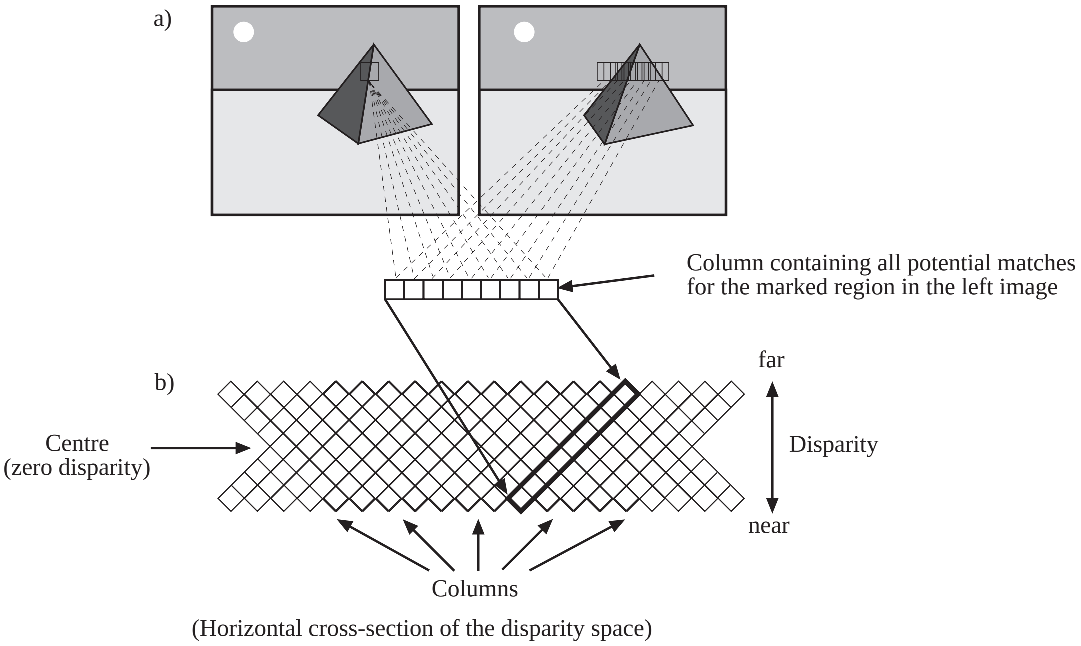
a)
Column containing all potential matches
for the marked region in the left image
b)
far
Disparity
near
Centre
(zero disparity)
Columns
(Horizontal cross-section of the disparity space)

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.

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.

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.

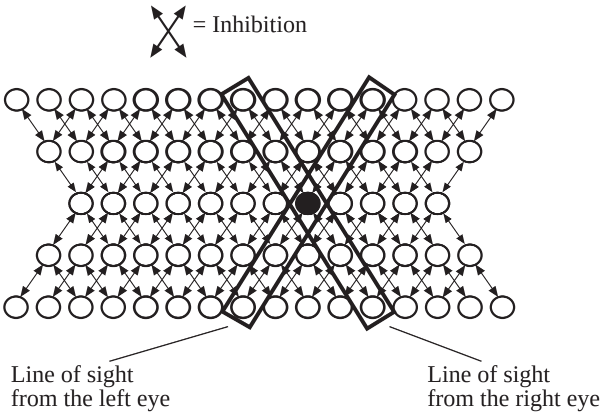

Figure 13. Horizontal cross-section of a disparity-space. The constraint of uniqueness is implemented by letting all cells, along the two lines of sight, inhibit each other.

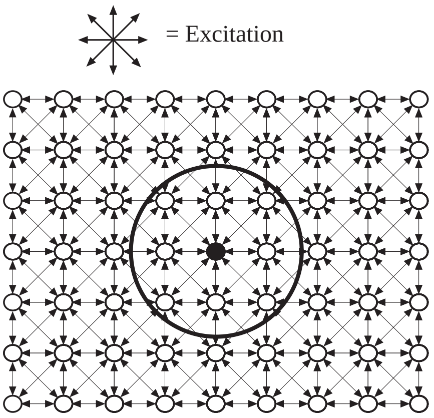

Figure 14. Vertical cross-section of a disparity-space. The constraint of continuity is implemented by letting all active cells excite the cells, in neighbouring columns, that representing similar binocular disparity.

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, ($C_{x,y}$) {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 centered at (x,y). More formally, the normalized convoluted value, NC, at point (x,y) are given by the following equation,

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

where $r=4\sigma$ 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 centered within a 10 pixel range of the pixel (x_R,y_R) in the right image, which has the same relative position as the center of the left region (x_L=x_R, y_L=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 centered at (x_L,y_L) with a region in the right image centered at (x_R,y_R) is formally described by the following equation,

$$C = \frac{1}{(2r+1)^2} \sum_{x=-r}^{r} \sum_{y=-r}^{r} \Big[sign(L_{x,y} \cdot R_{x,y}) \cdot$$

$$\cdot \min\left(\frac{|L_{x,y}|}{|R_{x,y}|}, \frac{|R_{x,y}|}{|L_{x,y}|}\right) \cdot W(\min(|L_{x,y}|, |R_{x,y}|)) \Big],$$

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 centered around (x_L,y_L), and the right region centered around (x_R,y_R) respectively. The W(x) function

$$W(x) = (2.0 - e^{\frac{1}{c_1 + (1.0 + c_2 |x|)^{c_3}}})^{c_4}$$

returns a value beteween 0.0 and 1.0 that is proportional to the strengt of the weakest of the two contrast values. The constants c1, c2, c3 and c4 have the values of 0.4427, 5.0, 3.0 and 1.5 respectively. 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 intervall [–1.0, 1.0], but since only the positive values are of interest all negative values are set to zero.

Note: Despite the complex appearance of the W(x)-function, all the function does is to amplify the contrast value so that the resulting correlation values will be more evenly spread over the interval [0.0, 1.0]. I later found that this function could be approximated by a much simpler one;

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

(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 recieve is determined by three factors: the current degree of activation, the strengt of inhibiting “cells” lying along the same two lines of sight, and the strengt of exciting neighbouring “cells” representing similar disparity. The following functions descibes how the new activation value (NA) is computed for a node in a disparity space:,

$$NA\,(CA, P, N) = CA + Excitation(P) - Inhibition(N),$$

where CA 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

$$Excitation(P) = 1.0 - \frac{1.0}{1.0 + \frac{P}{c}}$$

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

$$Inhibition(N) = 1.0 - \frac{1.0}{(1.0 + N)^c}$$

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 (level 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 processings 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 different 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).

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 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

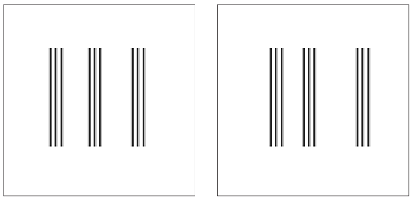

Figure 15. Input stereogram.

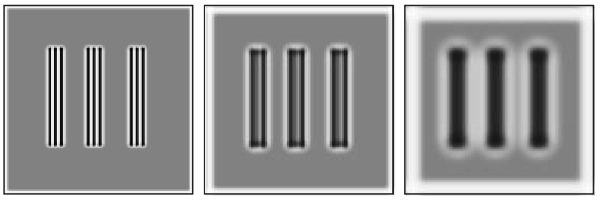

Figure 16. Result of the convolutions. If only the information within the finest channel (a) was considered, each of the thinner lines could be matched with any of the three thin lines, in the corresponding group, in the other image. However, in the coarser channels (b and particularly c) there is no 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.

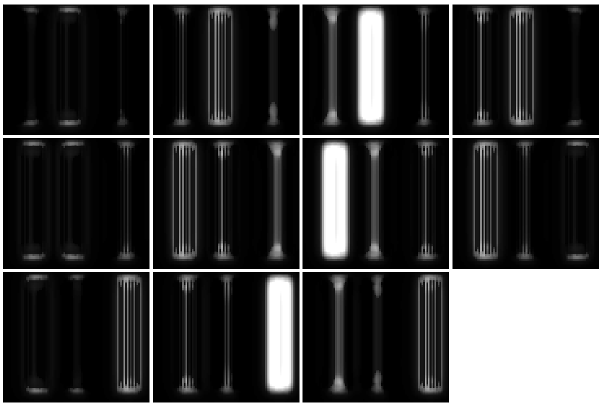

Figure 17. Each image above shows the activity within a vertical cross-section of the CDS. The brighter areas indicate high activity (potential matches). Depth increases from left to right, and from top to bottom.

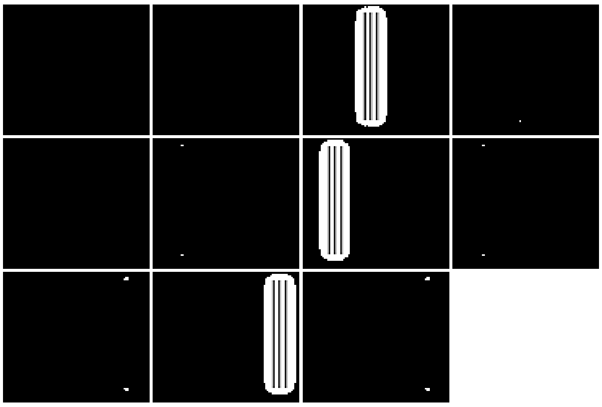

Figure 18. Activity within the CDS after 3 iterations. The original (left) image of the stereogram has been mapped ontop of areas that still are active.

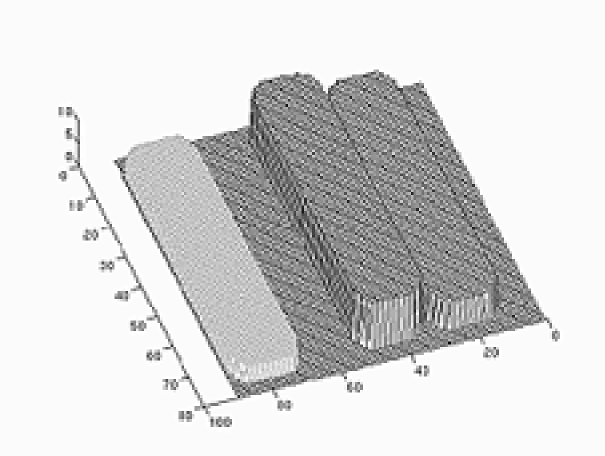

Figure 19. Wire-diagram of the disparity (activity) within the CDS after 3 iterations.7

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.

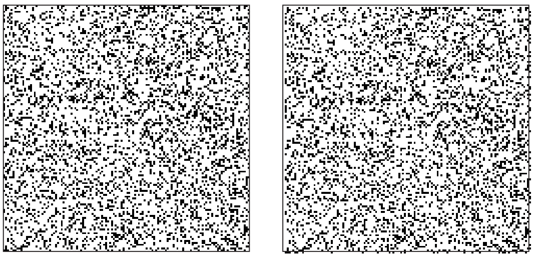

Figure 20. Random-dot stereogram with a density of 25% black dots.

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.

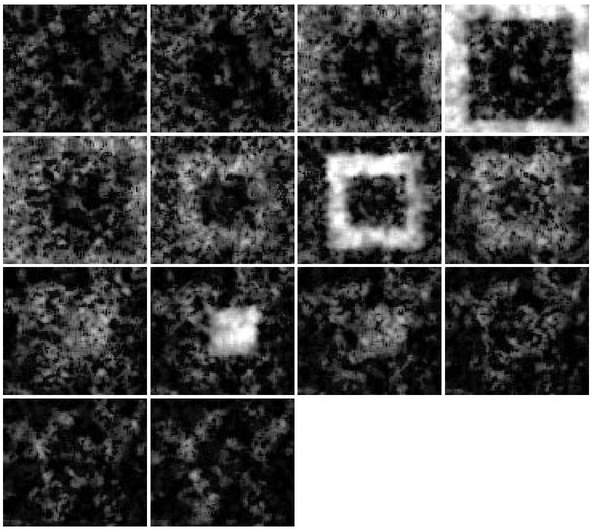

Figure 21. Activity within the CDS after the initial matching procedure.

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.

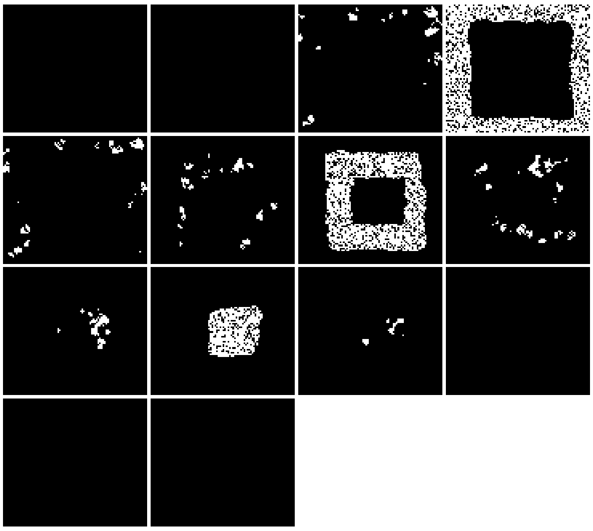

Figure 22. Remaining activity after 5 iterations (with the original left image mapped ontop).

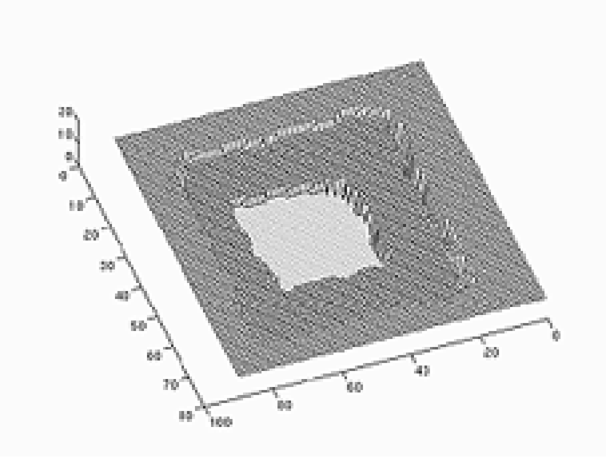

Figure 23. Wire-diagram of disparity (activity) within the CDS after 5 iterations.

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 satisfiably even if a substantial amount of "noise" (uncorrelated information) is added to the stereogram, or if the individual images are slightly shifted vertically.

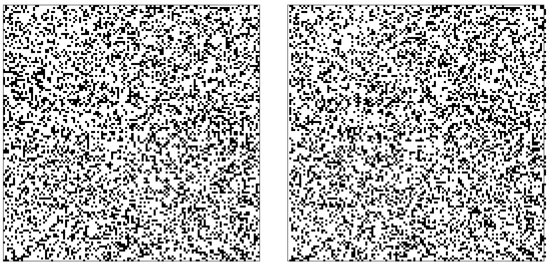

Figure 24. Random-dot stereogram where only 75% of the dots are correlated.

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).

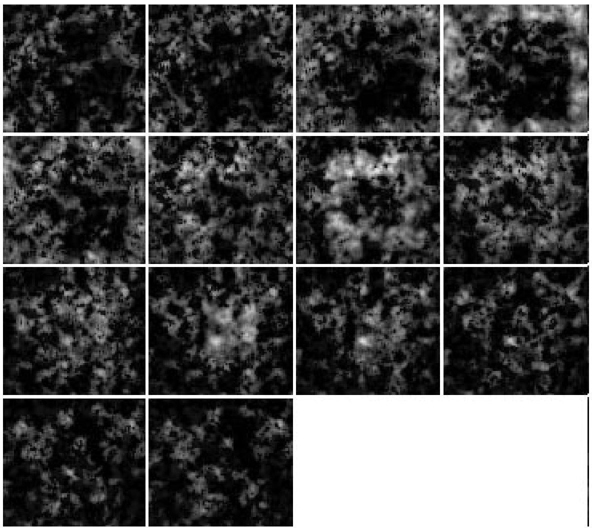

Figure 25. Initial matching.

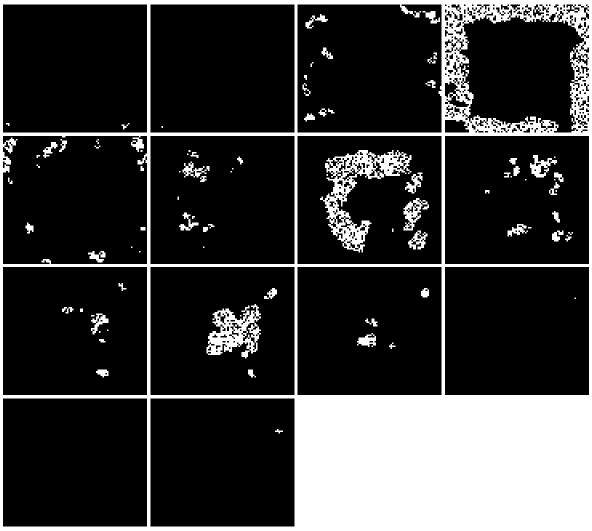

Figure 26. Activity after 7 iterations (left image mapped ontop).

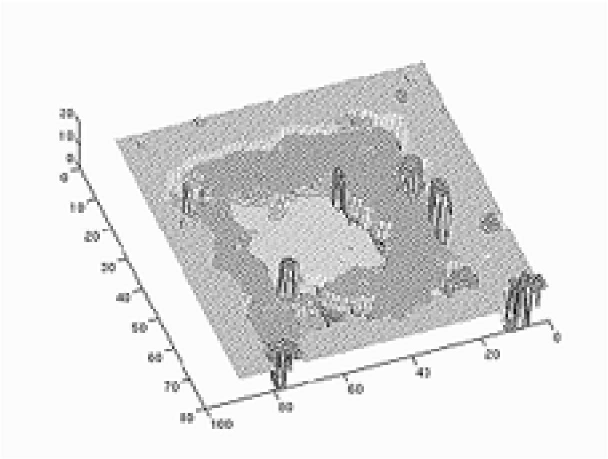

Figure 27. Active nodes after 7 iterations. Although the three planes are somewhat distorted, due to remaining false matches, they are still clearly distinguishable.

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).

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.

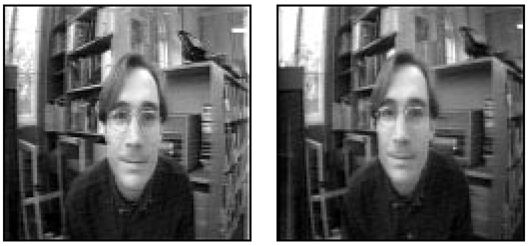

Figure 28. The author(s).

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.

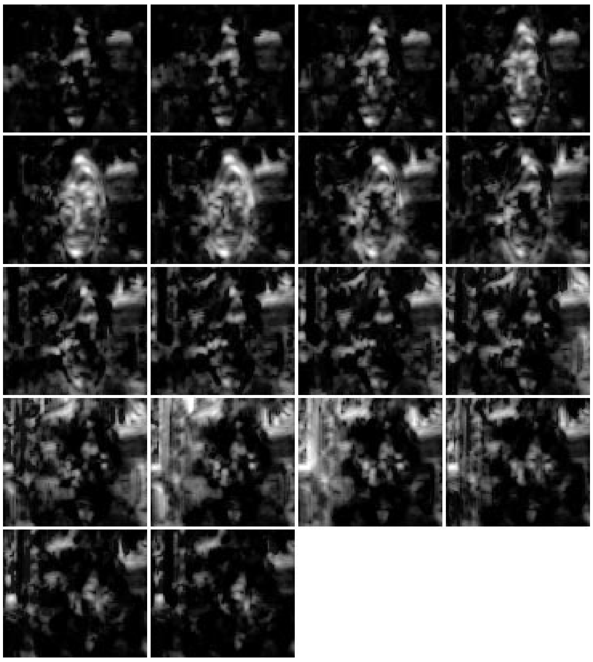

Figure 29. Initial matching result.

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.

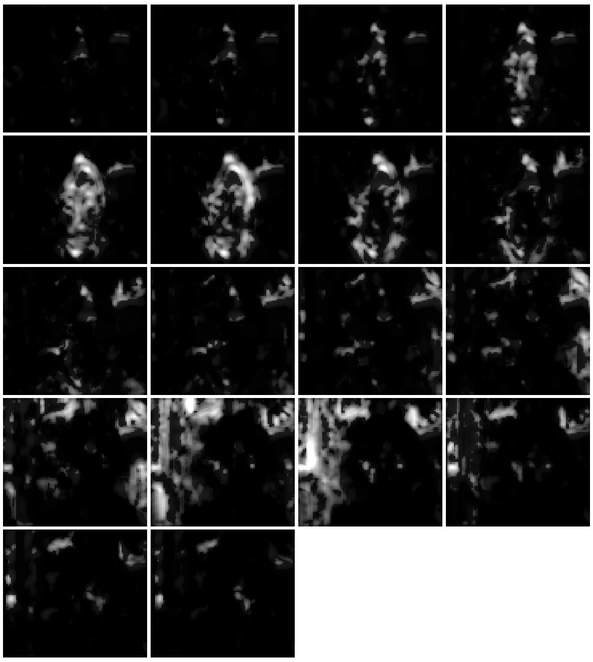

Figure 30. Activity after the first iteration.

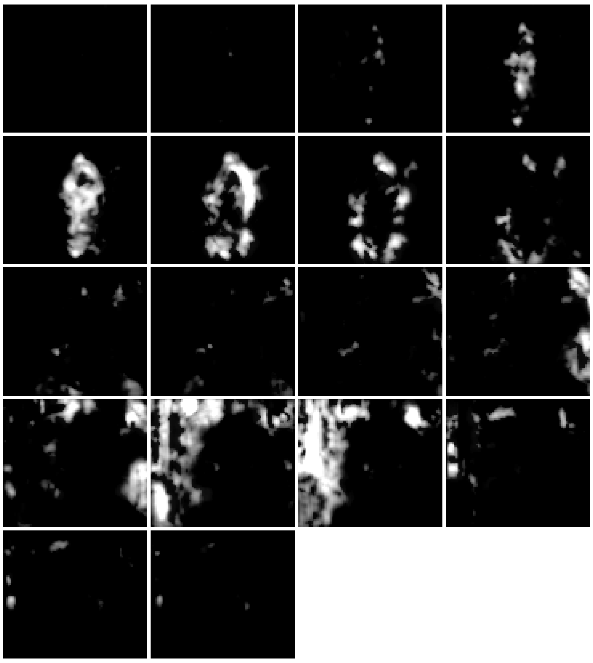

Figure 31. Third iteration.

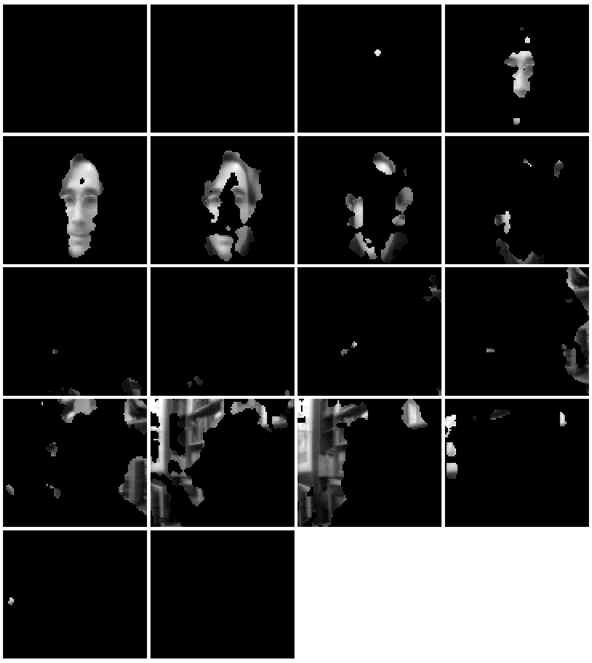

Figure 32. Activity after 7 iterations (with the original left image mapped ontop of the most active regions).

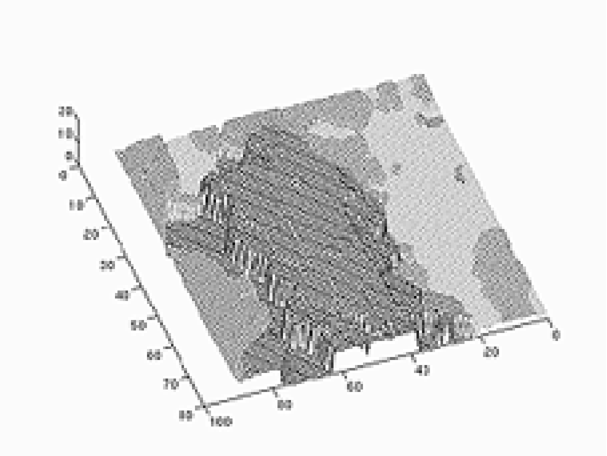

Figure 33. Disparity-map after 7 iterations.

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).

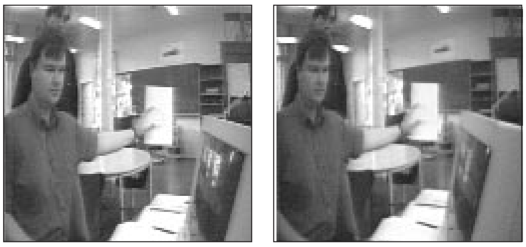

Figure 34. Input (my tutor).

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 seeen 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.

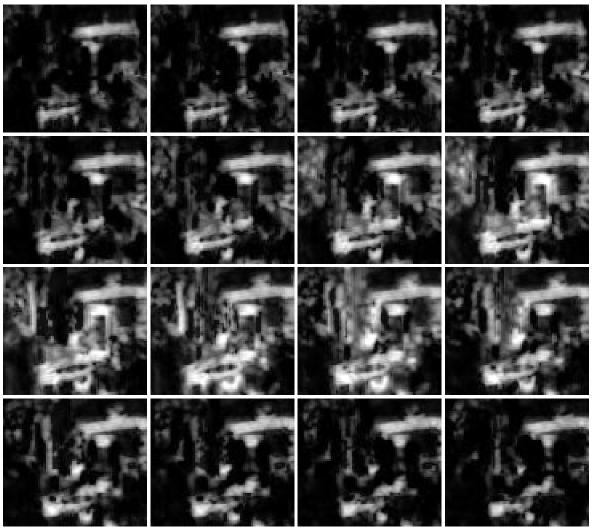

Figure 35. Activity after the initial matching.

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

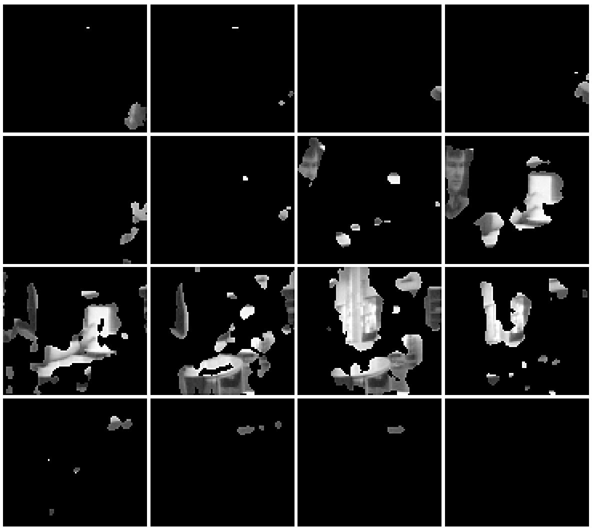

Figure 36. Active nodes after 7 iterations.

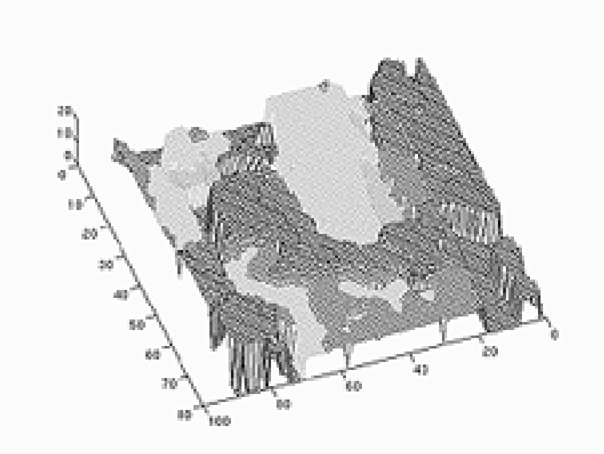

Figure 37. Disparity-map after 7 iterations.

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

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 important (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 *accomodation* 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 accomodation 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 accomodation 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 accomodation 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 continuos 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 neurone’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-neurone” 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 continuos 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 continuos 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.