Allerton Conference, Urbana-Champaign, Illinois, September 1994

Stereo Vision via Markov Models

Davi Geiger[1]
Courant Institute, NYU
251 Mercer Street
New York, NY 10012

Laurence T. Maloney
Psychology, NYU
6 Washington Place
New York, NY 10003

Krishna Kumaran
Department of Physics
Rutgers University
Piscataway, NJ 0885

Abstract

Binocular stereo is the process of obtaining depth information from two simultaneous views of a scene. The performance of computational stereo algorithms remains distinctly inferior to human performance, and at the same time, many phenomena in human stereo vision (transparency, Da Vinci stereopsis, subjective contours) have no natural place in these algorithms.

Through the framework of hidden Markov models we formulate a stereo algorithm that match along epipolar lines (one dimensional). Our aim is to introduce, in a simple and yet general manner, constraints that can account for these psychophysical phenomena. The first result is a distinctive neighborhood structure for our model and these various properties of stereo, such as occlusions, disparity discontinuity, transparency, are accounted. In order for us to fully address the problem of subjective contours the model is extended to two dimensions and a Markov random field model emerges. Finally we discuss the primitives of stereo, proposing differential geometric features, and studying their relation to occlusions, discontinuities and transparency. Initial experiments demonstrate the power of our formulation.[2]

1 Introduction and background

A biological visual system or computer vision system that uses binocular stereo transforms the disparities between features in the two images into information about the shapes and locations of objects in the scene. In the 1960's and later, experience with the Julesz random dot stereogram led to the development of several computational theories of disparity processing in biological vision including those of Sperling[19], Julesz [9]; Marr and Poggio[11] [12]; Pollard, Mayhew and Frisby[18]; Grimson[8]; Baker and Binford [2]; Kanade and Okutomi[10]. The performance of binocular stereo algorithms, including those cited above, has proven to be markedly inferior to that of biological vision systems.

These algorithms also neglect sources of stereo information corresponding to transparency, occlusion, and subjective contour phenomena in human vision. In many scenes, there are portions of the scene visible to one camera/eye that are not visible to the other (half-occlusion regions). The experiments of Nakayama and Shimojo

[2]Part of the work done when visiting the Isaac Newton Institute of Mathematics, Cambridge, UK.

[15] suggest that these regions play an influential role in human binocular stereo vision. Geiger, Ladendorf and Yuille [7] and independently Belhumeur and Mumford [3] have recently proposed models that address the relation between discontinuities and half-occlusions (termed by [3]). Much workremains to be done in this area. The phenomenon of transparency is marginalized by typical computational theories of binocular stereo, even though disparity cues alone can give the human observer a strong impression of transparency[16]. Interactions among epipolar lines are commonly discarded or underused. Indeed, there are no specific performance ciriteria proposed for such interactions (other than "looks good"). Epipolar lines are corresponding lines in the left and right images.

1.1 Our approach

Our main intents are to develop a robust, efficient, and simple stereo algorithm that outperforms existing algorithms and to provide alternative explanations for a variety of recent psychophysical experiments and suggest new experiments. We are developing a theoretical framework for binocular stereo in which occlusion, transparency, and subjective contours are treated in a unified way (all are aspects of the same problem). It is a Bayesian approach starting with differential geometric primitives designed to detection the kinds of discontinuities that signal occlusion and transparency phenomena. These features are paired bipartite windows (*D-windows*). The hidden Markov model (which includes prior information) guides the reconstruction of a surface that best matches these primitives. In the hidden Markov model, our observations are the features extracted from the stereo pairs. The states are given by the possible disparity values and the transitions are guided by the prior surface model we impose, piecewise smooth surfaces with constraints on transparency and occlusions. When interactions among epipolar lines are included then a Markov random field model naturally emerges.

Our presentation can be sketched as follows:

1. Stereo primitives ('D-windows') and candidate correspondences.
2. Occlusion, transparency, and subjective contours via a hidden Markov model.
    a. Half-occlusions to discontinuities. We start building the hidden Markov model for surface reconstruction considering our previous work that relates half-occlusions to discontinuities.
    b. Transparency. We discuss criteria to complete (fill in) the disparity information on half-occluded (unmatched) areas, taking into account Metelli's rule [13], and how it may affect the full recovery of depth.
    c. Subjective contours. We complete the model by introducing epipolar line interactions and generalizing it to a two dimensional Markov random field model. We argue that appropriate modeling (we are here defining performance criteria) of the epipolar line interactions, also known as figural continuity, yields the formation of subjective contours.
3. Initial experimental results targeted to 2.

2 Bayesian Reconstruction: A hidden Markov model

The Bayesian approach assumes that we can express the probability of the scene S given input data I by a distribution $P(S|I)$ which, by Bayes' theorem, can be written as $P(I|S)P(S)/Z$, in terms of the *imaging model* $P(I|S)$, the prior model $P(S)$ and a normalization constant Z. In addition, we can impose *hard constraints* on the possible solutions S^*. We assume that the optimal interpretation S^* is obtained by the *maximum a posteriori* estimate, $S^* = ARG\{MAX_S\}P(S|I)$. The theory can be described either in terms of solving a matching problem or in terms of surface reconstruction. It gives a compromise between accurate fitting of the data and conformance to a prior model of surfaces.

2.1 Observations or input probabilities

For each candidate match where a feature at l, from the left eye, matches a feature at r, from the right eye, a disparity $d = r - l$ is considered. In the matching space this information corresponds to paving the space with data accounting for the strength of the match. The matching space is a two-dimensional space where the axes are given by the epipolar lines of the left and right images and each element of the space, $M_{l,r}$, decides whether a feature at pixel l, in the left image, matches a feature at pixel r, in the right images (see figure 1). A solution for the matching problem, a disparity map, along an epipolar line is represented by a path in the matching space.

Primitives: D-windows. The stereo primitives we employ are small windows in each of the two stereo images. The windows are halves of elliptical disks (ellipses), shaped like a filled-in 'D' and we will term them *D-windows*. A D-window is specified by giving the center of its elliptical boundary, the two axes , and the angle (zero to 2π) of the larger axis (bisecting the D-window). For any D-window there is a *complementary D-window* that is the other half of the elliptical disk (Figure 2b). D-windows are transformed into D-windows under affine (linear) transformation. . We wish to find those affine correspondences that represent 'good' matches between the contents of the D-window in the left image (under affine transformation) and the corresponding D-window in the right image. For each "good" affine of D-windows correspondence there is a complementary D-window measure, providing two sets of parameters. It is often the case that the parameters are the same. At certain points, there is no acceptable affine mapping of the complementary windows. This suggests there is a half-occlusion region. The last possibility is that there is an affine mapping of the two complementary windows, but the set of affine parameters is markedly different. This suggests that the orientation of the surface in the scene has changed discontinuously, but that no occlusion has occurred. This information will also prove useful in reconstructing the three-dimensional structure of the scene.

As in [11] [7], we use a matching process $M_{l,r}$ that is 1 if, a feature at pixel l in the left eye matches a feature at pixel r in the right eye, and is 0 otherwise. We define the probability of observing a pair of inputs, $D^{L,\pi}$ and $D^{R,\pi'}$, given the matching process M, by

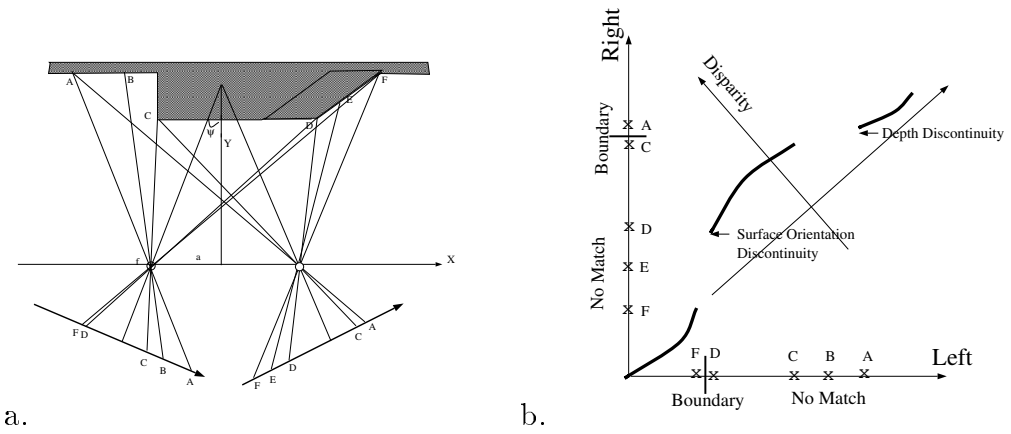

Figure 1: *(a) A polyhedron (shaded area) with self occluding regions and with a discontinuity in the surface-orientation at feature D and a depth discontinuity at feature C. (b) A matching space for the image of the ramp above. It has elements $M_{l,r}$ that decide if a feature at pixel l, in the left epipolar line, matches to a feature at pixel r, in the right epipolar line.*

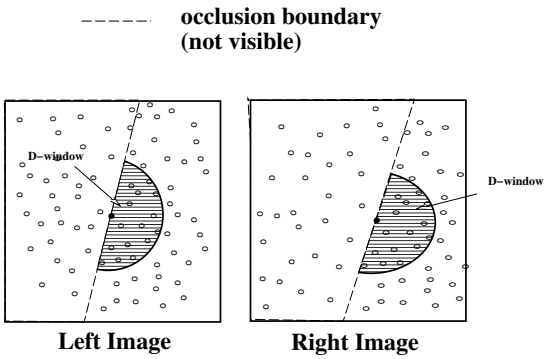

Figure 1a

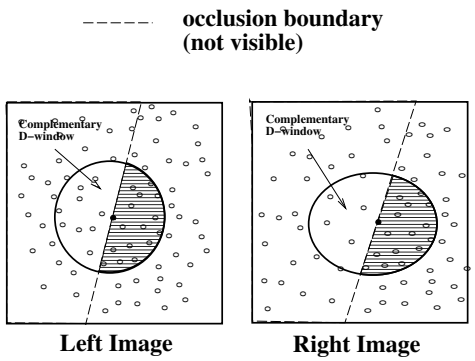

Figure 1b

Figure 2: D-windows and its transformations. a. exhibits an D-window that is the image of the D-window on the left. They are, of course, half-elliptical in shape, and the other half of the ellipse is the *complementary D-window* (b).

$$P_{input}(D^{L,\pi}, D^{R,\pi'}|M) = e^{-\sum_{l,r}\left[M_{l,r}\|D_l^{L,\pi} - D_r^{R,\pi'}\| + \epsilon(1-M_{l,r})\right]}/C_1 \quad (1)$$

where $l = 0, ..., N-1$ and $r = 0, ..., N-1$ are indices that scan the left and right images along the epipolar lines. The ϵ term pays a penalty for unmatched points, where $M_{l,r} = 0$, with ϵ being a positive parameter to be estimated. C_1 is a normalization constant. In order to prohibit multiple matches from occuring we impose a *uniqueness* constraint:

$$\sum_{l=0}^{N-1} M_{l,r} = 0, 1 \quad \forall r \in (0, ..., N-1) \quad \text{and} \quad \sum_{r=0}^{N-1} M_{l,r} = 0, 1 \quad \forall l \in (0, ..., N-1)\,. \quad (2)$$

Notice that this uniqueness guarantees that there is at most one match per feature, but also permits unmatched features to exist. We define a disparity field for the left eye, d^L, and another for the right eye, d^R, as

$$d_l^L(M) = \sum_{r=0}^{N-1} M_{l,r}(r-l)\,, \quad \text{and} \quad d_r^R(M) = \sum_{l=0}^{N-1} M_{l,r}(r-l)\,.$$

where d^L and d^R are defined only if a match occurs. This definition leads to integer values for the disparity field (if no continuous values of $M_{l,r}$ are considered).

2.2 Half-occlusions and discontinuities

The initial prior surface model is for surfaces that are smooth but also allows for large diparity changes to occur without high penalties. We suggest to use the potential

$$\begin{aligned} U(d^L, d^R(M)) &= \mu[\sum_l Sm|d_l^L - d_{l-1}^L| + \sum_r Sm|d_r^R - d_{r-1}^R|\,] \\ &= \mu \sum_l \sum_r M_{r,l}[\sum_{r'=0}^{N-1} M_{r',l-1} Sm(|r-r'-1|) + \sum_{l'=0}^{N-1} M_{r-1,l'} Sm(|l-l'-1|)\,] \end{aligned}$$

where μ is a constant, and $Sm(x)$ is a monotonicly increasing smooth function, such that after some value of x, it takes negative derivative. This implies that the function increases less and less as x increases. The function $\sqrt{(}x)$ is a good candidate for prior and we adopt in this paper. We emphasize that the choice of prior is motivated by the class of stimuli on which the system is designed to work. The occlusion constraint [7],

Proposition 1 (Occlusion constraint) *A discontinuity in one eye, along the epipolar line, corresponds to an occlusion in the other eye. Moreover, the prior cost for an occlusion in the left and right eye must be the same.*

can be formulated as a *monotonicity constraint*. Plot all the matched pairs in the matching space. We require that for any matched pair (r, l) at least one of neighbouring points at $r+1$ (in the right image) or $l+1$ (in the left eye) must also be matched. Join neighbouring matched pairs by straight line segments to form a curve. The *monotonicity constraint* only allows matching such that this curve is monotonic when consider as a function either of l or r. This constraint, which is a variant of the ordering constraint, can be easily applied by modifying the potential to

$$U(M) = \mu \sum_l \sum_r M_{r,l} \left(\sum_{r'=0}^{r-1} M_{r',l-1} \sqrt{|r-r'-1|} + \sum_{l'=0}^{l-1} M_{r-1,l'} \sqrt{|l-l'-1|} \right)$$

and assigning a Gibbs probability distribution for the prior and combining it with (1), using Bayes' theorem, we obtain

$$P_{stereo}(M| \parallel D_l^{L,\pi} - D_r^{R,\pi'} \parallel) = \frac{e^{-\epsilon N^2}}{Z}$$

$$\prod_{r=0}^{N-1} \prod_{l=0}^{N-1} e^{-M_{l,r} \left[\|D_l^{L,\pi} - D_r^{R,\pi'}\| - \epsilon + \mu \left(\sum_{r'<r} M_{r',l-1} \sqrt{|r-r'-1|} + \sum_{l'<l} M_{r-1,l'} \sqrt{|l-l'-1|} \right) \right]} \tag{3}$$

2.3 Filling in half-occlusions and transparency

Filling in the disparity (surface) for the half-occluded regions may affect the overall solution of the correpsondence. We introduce a new matching-occlusion-transparency field, T_{lr}, which is 1 or 0. It is 0 if neither pixel, l or pixel r, is occluded. It is 1 only if the disparity estimation is $r-l$ and one of them are occluded thus, we have never the case $T_{lr} = M_{lr} = 1$. In our theory transparency can only occur when $T_{lr} = 1$.

The first observation is that a feature in the half-occluded area is likely to have the same disparity as a similar neighbor feature (eventually one neighbor feature nearby will not be half-occluded to provide the disparity information). The neighbor with larger similarity provides the information to be given to the occluded pixels. This is the fill in constraint.

Transparency: Always, one of the possible solutions considers the occluded surface to be in front of the matched one ($M_{lr} = 1$). Then, a phenomena of transparency must occur or else how come occluded features are perceived in front ? In these cases, we must consider transparent requirements and we recall Metelli's transparent rules [13] to include in the energy function. Thus, $Met^L(l, r, l', l'')$ is 1 if $I^L_{l''}$ could be in front of $I^L_{l''+l'-l+1}$ such that $I^R_{r-l+l''}$ would be seen in the right eyeand is zero otherwise. Here, $l - l'$ is the size of the half-occluded region (i.e. $M_{lr} = 1$ and $M_{l',r-1} = 1$) and l" is some pixel in the occluded region. Analogously $Met^R(r, l, r', r'')$ is 1 or zero, in the right image.

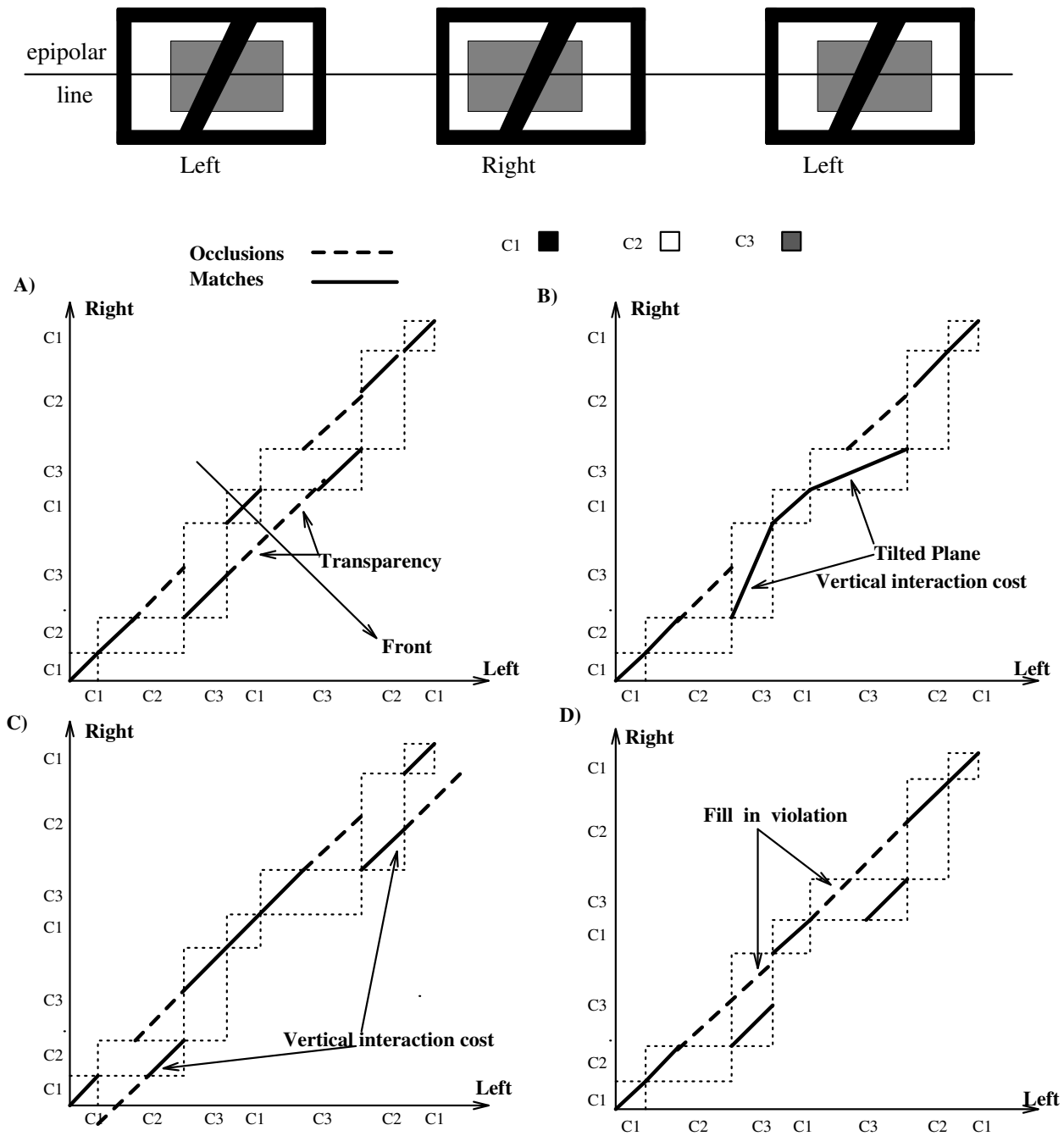

Figure 3: *Transparency from stereo (After Nakayama and Shimojo, 1992). We analyse four situations. Notice that the solution at top (and bottom) epipolar lines gives zero disparity everywhere. The dotted boxes represent regions of possible matches (acceptable similarity measures). A. Actual solution. The transparency is a consequence of the filling in process. Note that the filling in process has limits of diffusion that can be tested via psychophysical experiments. B. The tilted plane solution, possible along epipolar lines, but with high cost when vertical interactions are added since, different epipolar lines will give different tilted angles (if the black inclined bar were vertical the tilted solution would be also plausible). C. Again a possible epipolar solution (with the same cost as A.), but becomes costly when vertical interactions are added (because of region C2, white). D. Not a possible solution due to the fill in constraint (fill in with similar neighbors disparity).*

Another possibility of a tilted plane must be included (See figure 3 B.). We propose to add the energy term

$$
\begin{aligned}
U^{T,L}_{r,l,l'} &= \sum_{l''=l'+1}^{l-1} \Big\{ T_{l'',r-1+l''-l'} \parallel D^{L,\pi}_{l''} - D^{L,\pi'}_{l'} \parallel + \\
&\quad [\,(2 - Met^L(l,r,l',l''))\, T_{l'',r+l''-l} + T_{l'',r}] \parallel D^{L,\pi}_{l''} - D^{L,\pi'}_{l} \parallel \Big\} \\
U^{T,R}_{r,l,r'} &= \sum_{r''=r'+1}^{r-1} \Big\{ T_{l-1+r''-r',r''} \parallel D^{R,\pi}_{r'} - D^{R,\pi'}_{r''} \parallel + \\
&\quad [\,(2 - Met^R(r,l,r',r''))\, T_{l+r''-r,r''} + T_{l,r''}] \parallel D^{R,\pi}_{r} - D^{R,\pi'}_{r''} \parallel \Big\}.
\end{aligned}
$$

Note that for each r'' and l'', only three cases are considered for one solution (T=1): (i) The first one gives a regular occlusion (ii) gives a transparent "occlusion" and (iii) gives a tilted surface (see figure 3). An improvement over this cost for the disparity estimation at occluded regions is via the $D-windows$ information since, their parameters provide information about the surface. We plan to explore this possibility theoretically and experimentally.

We can use a dynamic programming algorithm [4] to solve for the optimal path due to the local neighbor interactions. One of the first works on stereo using dynamic programming was by Baker and Binford [2] and later ([17]).

2.4 Epipolar line interactions and subjective contours

Epipolar line interaction play an important role on the formation of a depth map from a stereo pair. Many theories of stereo address this issue. We propose that epipolar interaction's models to be measured on their capacity of predicting subjective surfaces/contours (as a true positive case). Moreover, some transparent phenomena are fully understood (because of disambiguation) when epipolar line interaction take place (see figure 3 B. and C.).

We observe that subjective contours are always perceived with a perception of a surface in front of another one (occluding another one). The stereo process yields the perception of depth surfaces and thus, cause strong subjective contours sensations. In the example of the stereo Kanizsa triangle (see figure 4), there are epipolar lines with zero disparity, e.g. in the middle of the figure. We conjecture that the depth perception is consequence of the epipolar line interaction that enforces piecewise smooth depth maps. We are considering adding a vertical interaction energy term of the form

$$
U^{ep} = \sum_e \Big\{ \sum_r (1 - \gamma e h^R_r) \sqrt{d^{R,e}_r - d^{R,e-1}_r} + \sum_l (1 - \gamma e h^L_l) \sqrt{d^{L,e}_l - d^{L,e-1}_l} \Big\},
$$

where eh^R_r and eh^L_l are the intensity edge maps of the right and left image, i.e. they are 1 if there is an edge and zero otherwise, γ is an attenuation factor smaller (or

equal) to 1. This term will effectively block the epipolar interactions when an intensity edge separate the epipolar lines. The model becomes a Markov random field, in two dimensions, and we can no longer use dynamic programming (unless approximations between these interactions are made, e.g. mean field approximation).

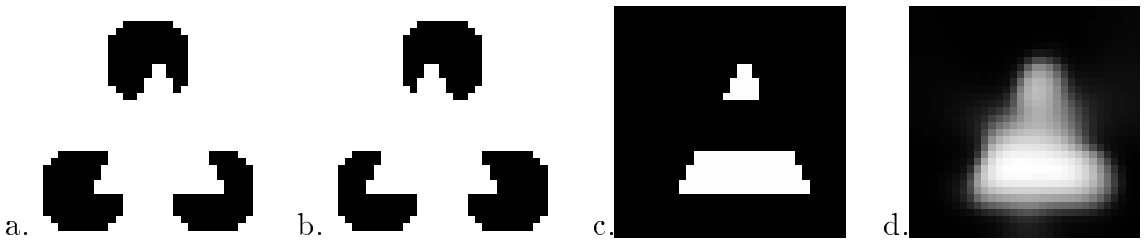

Figure 4: *(a) and (b) 32 by 32 pixels of the Left and right Kanisza triangle. (c) solution along epipolar lines (using a bias term towards zero disparity) (d) reconstruction of disparity from sparse data (assuming information availble from (c) just at intensity edges.*

2.5 Pilot experiment

The algorithm for (3), without using the filling in process and the epipolar line interaction, has already been tested on real images with positive results (see [7]). We have so far conducted pilot experiments with black and white images to test these increments. The features are single pixels. However, along vertical interactions we have implemented a simpler approach. We simply run the algorithm along epipolar lines and selected the information at intensity edges (where is more reliable). Then a reconstruction of 2D, from these sparse data (disparity data at intensity edges) provides a full recover of subjective surfaces. We could alternatively use the hidden markov model machinery to estimate the confidence for each estimate, but further analysis needs to be done. Figure 4 is a typical pilot experiment using the Kanizsa triangle and the results are supportive of the model.

Acknowledgements: D.G. would like to thank the hospitality and environment of the Isaac Newton Institute. In particular, discussions with A. Blake and D. Mumford and J. Mayhew were extremely helpful.

References

[1] N. Ayache. *Artificial Vision for Mobile Robots* The MIT Press. Cambridge, Mass 02142. 1991.

[2] H.H. Baker and T. O. Binford. Depth from edge and intensity based stereo. In *Proceedings IJCAI*, pages 631–636, Vancouver, 1981.

[3] P.N. Belhumeur and D. Mumford. A bayesian treatment of the stereo correspondence problem using half-occluded regions. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition*, 1992.

[4] R. E. Bellman. *Dynamic Programming.* Princeton University Press, 1957.

[5] A. Blake and A. Zisserman. *Visual Reconstruction.* MIT Press, Cambridge, Mass, 1987.

[6] D. Buckley and J. Frisby and J. Mayhew. Integration of stereo and texture cues in the formation of discontinuities during three-dimensional surface interpolation. *Perception*, 18-5:563–588, 1989.

[7] D. Geiger and B. Ladendorf and A. Yuille". Binocular stereo with occlusion. *2nd ECCV-Lecture Notes on Computer Science*, Springer-Verlag, May 1992.

[8] W. E. L. Grimson. *From Images to Surfaces.* MIT Press, Cambridge, Mass., 1981.

[9] B. Julesz. *Foundations of Cyclopean Perception.* The University of Chicago Press, Chicago, 1971.

[10] T. Kanade and M. Okutomi. A stereo matching algorithm with an adaptive window: theory and experiments. In *Proc. Image Understanding Workshop DARPA*, PA, September 1990.

[11] D. Marr and T. Poggio. Cooperative computation of stereo disparity. *Science*, 194:283–287, 1976.

[12] D. Marr and T. Poggio. A computational theory of human stereo vision. *Proceedings of the Royal Society of London B*, 204:301–328, 1979.

[13] F. Metelli. The perception of transparency. *Scientific American*, 230 (90):91-98, April 1974.

[14] D. Mumford and J. Shah. Boundary detection by minimizing functionals, i. In *Proc. IEEE Conf. on Computer Vision & Pattern Recognition*, San Francisco, CA, 1985.

[15] K. Nakayama and S. Shimojo. Da vinci stereopsis: depth and subjective occluding contours from unpaired image points. *Vision Research*, 30:1811–1825, 1990.

[16] K. Nakayama and S. Shimojo". Experiencing and Perceiving Visual Surfaces. *Science*, 257:1357–1363, 1992.

[17] Y. Ohta and T. Kanade. Stereo by intra- and inter-scanline search using dynamic programming. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-7(2):139–154, 1985.

[18] S. B. Pollard, J. E. W. Mayhew, and J. P. Frisby. Disparity gradients and stereo correspondences. *Perception*, 1987.

[19] G. Sperling. Binocular vision: A physical and a neural theory. *American Journal of Psychology*, 83:461–534, 19670.