Rethinking the Prior Model for Stereo

Hiroshi Ishikawa[1] and Davi Geiger[2]

[1] Department of Information and Biological Sciences,
Nagoya City University,
Nagoya 467-8501, Japan
hi@nsc.nagoya-cu.ac.jp
[2] Courant Institute of Mathematical Sciences,
New York University,
New York, NY 10012, USA
geiger@cs.nyu.edu

Abstract. Sometimes called the smoothing assumption, the prior model of a stereo matching algorithm is the algorithm's expectation on the surfaces in the world. Any stereo algorithm makes assumptions about the probability to see each surface that can be represented in its representation system. Although the past decade has seen much continued progress in stereo matching algorithms, the prior models used in them have not changed much in three decades: most algorithms still use a smoothing prior that minimizes some function of the difference of depths between neighboring sites, sometimes allowing for discontinuities.

However, one system seems to use a very different prior model from all other systems: *the human vision system.* In this paper, we first report the observations we made in examining human disparity interpolation using stereo pairs with sparse identifiable features. Then we mathematically analyze the implication of using current prior models and explain why the human system seems to use a model that is not only different but in a sense diametrically opposite from all current models. Finally, we propose two candidate models that reflect the behavior of human vision. Although the two models look very different, we show that they are closely related.

1 Introduction

The main task in low-level vision is to filter out as much as possible irrelevant information that clutter the input image. There, disambiguation is one of the central problems, since resolving ambiguity eliminates great amount of later processing. Ambiguity arises because input images to a vision system usually do not contain enough information to determine the scene. Thus the vision system must have a prior knowledge on the kinds of scenes that it is likely to encounter in order to choose among possible interpretation of given data. In the case of stereo matching, where the correspondences between locations in the two or more images are determined and the depths are recovered from their disparity, ambiguity arising from such factors as noise, periodicity, and large regions of constant intensity makes it impossible in general to identify all locations in the two images with certainty. Thus, any stereo algorithm must have a way to resolve ambiguities and interpolate missing data. In the Bayesian formalism of stereo vision,

A. Leonardis, H. Bischof, and A. Pinz (Eds.): ECCV 2006, Part III, LNCS 3953, pp. 526–537, 2006.

this is given by the prior model. The prior model of a stereo matching algorithm is the algorithm's expectation on the surfaces in the world, where it makes assumptions about the probability to see each surface that can be represented in its representation system.

The prior model is an ingredient of stereo matching reasonably separate from other aspects of the process: whether a stereo system uses Dynamic Programming or Graph Cut or Belief Propagation, it explicitly or implicitly assumes a prior; and it is also usually independent of image formation model, which affects the selection of features in the images to match and the cost function to compare them. Also, in some algorithms, it is less obvious than in others to discern the prior models they utilize, especially when the smoothing assumption is implicit as in most local, window-based algorithms. In some cases, it is intricately entwined with the image formation model, as in the case where a discontinuity in disparity is encouraged at locations where there are intensity edges. As far as we could determine, however, the prior models that are used in stereo matching algorithms have not changed much in three decades. Computational models (Marr and Poggio[15, 16]; Grimson[8]; Poggio and Poggio[19]; Pollard, Mayhew, and Frisby[18]; Gillam and Borsting[7]; Ayache[1]; Belhumeur and Mumford[3]; Jones and Malik[10]; Faugeras[5]; Geiger, Ladendorf, and Yuille[6]; Belhumeur[2]) have generally used as the criterion some form of smoothness in terms of dense information such as the depth and its derivatives, sometimes allowing for discontinuities; among them, the most common by far is the minimization of the square difference of disparities between neighboring pixels, which encourage front-parallel surfaces.

Perhaps that most of the citations above are at least ten years old is indicative of the neglect the problem of prior model selection has suffered. The latest crop of algorithms, using Graph Cut (Roy and Cox[21, 20]; Ishikawa and Geiger[9]; Boykov et al.[4]; Kolmogorov and Zabih[12],) did not improve on the prior models, concentrating on the optimization. The excellent recent survey of stereo algorithms by Scharstein and Szeliski[22] does not classify the algorithms in their taxonomy by prior models– rightly, because there are not much difference in this respect among them.

Of course, by itself it might mean that the selection was exactly correct the very first time. However, it appears that there is a glaring exception to the widespread use of smoothing / front-parallel criterion as the prior model: *the human vision system.* In this paper, we first report the observations we made in examining human disparity interpolation using stereo pairs with sparse identifiable features. Then we mathematically analyze the implication of using current prior models and explain why the human system seems to use a model that is not only different but in a sense diametrically opposite from all current models. Finally, we propose two candidate models that reflect the behavior of human vision. Although the two models look very different, we show that they are closely related.

2 Experiment and Analysis

We used a stereogram, shown in Fig.1a, of textureless surface patches with explicit luminance contours to investigate the prior model the human vision uses. In these displays, there are very few features that can be depended upon when matching the points. The only distinguishing feature is the intensity edges on the circumference of the shape,

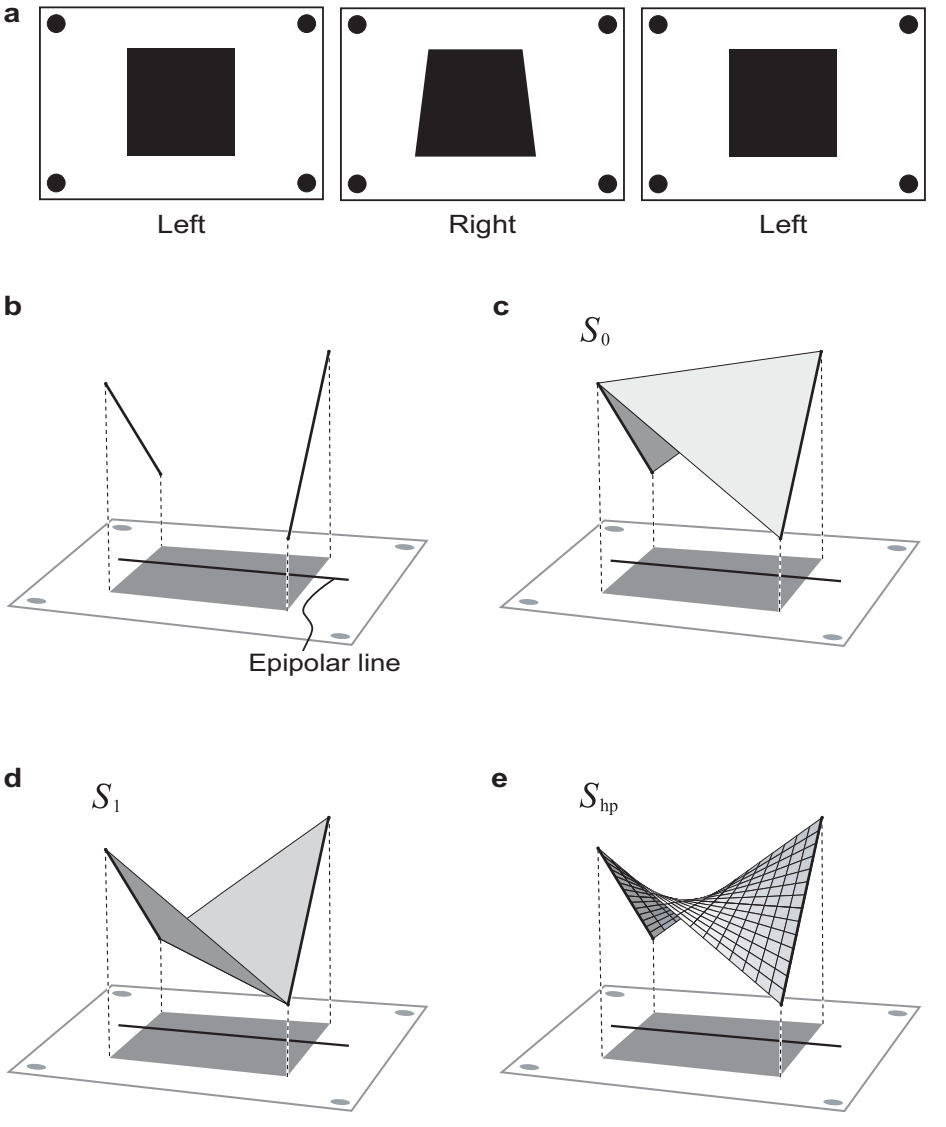

Fig. 1. The stereogram and possible surfaces. **(a)** A stereo pair. When the right images are cross-fused (or the left two images are fused divergently,) a three-dimensional surface is perceived. **(b)** The thick lines represent the disparity values unambiguously obtainable from local feature. **(c), (d)** The human brain tend to perceive either of these two. However, no current algorithm has this behavior. **(e)** Algorithms that seek to minimize gradient give a "soap bubble" surface like the one shown here. Models in which the prior on epipolar lines are independent line-by-line also give this solution.

where the discontinuous change in luminance occurs. There are no other cues that are ordinarily present, such as surface shade and partial occlusions (Gillam and Borsting[7]; Nakayama and Shimojo[17]; Malik[13]). Matching the edges gives the depth information illustrated in Fig. 1b. Everywhere else, each location in one image can perfectly match to a variety of locations in the other. This corresponds to the fact that any perfectly black surface spanning the two segments in Fig. 1b looks exactly the same.

Nevertheless, the perception human observers report is much less ambiguous. We examined shape judgments by human observers from the interpolated stereoscopic disparity. The details of the experiment can be found in the appendix. Most observers who viewed the stereogram reported the percept of one of the two surfaces shown in Fig. 1c and d, which we call S_0 and S_1. This result is in stark contrast to the smooth surface S_{hp}, shown in Fig. 1e, that is predicted by most extant computational models of stereo, as we explain in the following subsections.

2.1 One-Dimensional Models

First of all, any 1D interpolating model would predict the ruled surface S_{hp}. The three-dimensional geometry of image formation dictates the possible pairs of points in the image that can match each other (Fig. 2a). A point in a 3D scene and the two focal points determine a plane in the space. The projecting rays from the point through the focal points onto the retinae must lie on this plane. Thus, when the correspondence is not known, it can at least be said that a feature on one image can match only those locations on the other image that lie on the plane determined by the point and the two foci. Such possible matching points form a line called the epipolar line. Imagine a plane rotating around the line connecting the two foci: it sweeps the retinae, defining a set of corresponding epipolar lines. Geometrically, only points on the corresponding epipolar lines can match to each other. Thus, in theory, the stereopsis can be a 1D process that

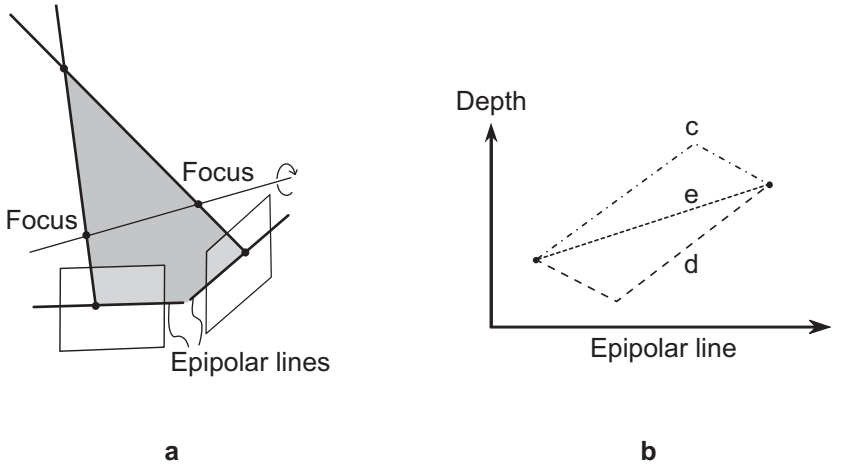

Fig. 2. **(a)** The geometry of stereopsis. Only points on the corresponding epipolar lines can match to each other. **(b)** The cross sections for solutions in Fig. 1c,d, and e on an epipolar line.

matches the locations on the two images line by line. One may be lead to postulate that the interpolation is also done one-dimensionally. However, the experiment shows that is not what is done in human perception. If the interpolation is done one-dimensionally on each epipolar line as Fig. 2b shows, the perceived surfaces S_0 and S_1 have forms that cannot be readily explained (**c** and **d** in Fig. 2b). Since the sole depth data given on each epipolar line are at the two endpoints, the only reasonable 1D interpolation is to connect the two points by a straight line, as indicated as **e** in Fig. 2b. As a whole, the lines give the smooth surface $S_{\rm hp}$. Thus, theories that only use one-dimensional information do not predict the surfaces usually seen by human observers.

2.2 Gradient Minimization Models

In some computational models of stereopsis, epipolar lines are not independent. It would be useful to have an interaction between the matching on different epipolar lines even just for the sake of robustness in the presence of noise. Most current theories model the matching by a depth surface that gives a dense map of the depth at each point in the view. Mathematically, a depth surface S is typically represented by specifying the value of the depth $d^S_{i,j}$ at each of dense sample points, which usually are laid out as an equally-spaced grid $X = (i, j)$. In such models, distinguishing feature such as intensity edges can give a strong evidence of matches, determining the depth value $d^S_{i,j}$ at some of the sample points. Ever since Marr and Poggio[15, 16] and Pollard, Mayhew, and Frisby[18], most computational models of stereopsis used a weak smoothing scheme that in effect predicts a surface S that minimizes the total change in depth:

$$E(S) = \sum_X \{(d^S_{i+1,j} - d^S_{i,j})^2 + (d^S_{i,j+1} - d^S_{i,j})^2\}, \tag{1}$$

which approximates the total depth gradient $\int |\nabla d^S|^2$. Here, the sum evaluates how much the depth value changes from one sample point to the next. One can see that the value $E(S)$ is minimum when the surface is flat and the depths $d^S_{i,j}$ at all points are equal, in which case $E(S) = 0$. When some data points have definite depth values that come from the matching, which is usually the case, these models predict a depth surface S that minimizes $E(S)$ under the constraint that it has definite values where they are known. Or, the data from the feature matching are evaluated as another "energy" function and the sum of the two is minimized. This can be considered as giving a probability to each possible surface. If the surface has the depth value that is strongly supported by the matching, it would have a higher probability; other than that, the surface has higher probability when the sum (1) is smaller. In the Bayesian formulation of stereopsis (Szeliski[23]; Belhumeur and Mumford[3]; Belhumeur[2]), this "energy" corresponds to a negative logarithm of the prior probability distribution, which gives an a priori probability for possible surfaces. It represents the model's idea of what surfaces are more likely in the absence of data. In the case of the image pair in Fig. 1a, the edges determine the depth at the two intensity edges that can be matched, as illustrated in Fig. 1b. At other sample points, however, there is not enough data to decide what depth to give to the point. This is why the model must have some disambiguating process.

How would such models react to the stereo pair in Fig. 1a? The answer is that all current theories predict a surface similar to $S_{\rm hp}$, rather than the most perceived surfaces

S_0 and S_1. This is because the gradient modulus $|\nabla d^S|$, at all points, is larger for S_0 and S_1 than for $S_{\rm hp}$. This can be easily seen by simple calculation as follows.

Let $2l$ be the side of the square and $2h$ the height (the difference of the maximum and the minimum depth) of the surface. We set up a coordinate system where the four corners of the square have the coordinates $(x, y) = (\pm l, \pm l)$. Of the definite depths determined by matching the intensity edges, we assume that the two corners (l, l) and $(-l, -l)$ have the depth h and the other two have the depth $-h$. Thus, the boundary condition is the two line segments, shown in Fig. 1b, determined by the equations $x = l, d = \frac{h}{l}y, -l \le y \le l$ and $x = -l, d = -\frac{h}{l}y, -l \le y \le l$. Then, the depth and the depth gradient for the surfaces S_0 and S_1 are as follows:

$$S_0: \quad d^{S_0}(x,y) = \begin{cases} \frac{h}{l}(x-y)+h & (x \le y) \\ \frac{h}{l}(-x+y)+h & (x \ge y) \end{cases}$$
$$\nabla d^{S_0}(x,y) = \begin{cases} (\frac{h}{l}, -\frac{h}{l}) & (x < y), \\ (-\frac{h}{l}, \frac{h}{l}) & (x > y) \end{cases}$$

$$S_1: \quad d^{S_1}(x,y) = \begin{cases} \frac{h}{l}(x+y)-h & (x \ge -y) \\ \frac{h}{l}(-x-y)-h & (x \le -y) \end{cases}$$
$$\nabla d^{S_1}(x,y) = \begin{cases} (\frac{h}{l}, \frac{h}{l}) & (x > -y), \\ (-\frac{h}{l}, -\frac{h}{l}) & (x < -y) \end{cases}$$

Thus, we obtain $\sqrt{2}\frac{h}{l}$ as the gradient modulus for S_0 and S_1 everywhere on the square, except on the diagonal where it is not defined. On the other hand, the depth and its gradient for $S_{\rm hp}$ at point (x, y) are defined by

$$S_{\rm hp}: \quad d^{S_{\rm hp}}(x,y) = \frac{h}{l^2}xy$$
$$\nabla d^{S_{\rm hp}}(x,y) = (\frac{h}{l^2}y, \frac{h}{l^2}x)$$

Thus the gradient modulus for $S_{\rm hp}$ at point (x, y) is $\frac{h}{l^2}\sqrt{x^2+y^2}$, which is smaller than $\sqrt{2}\frac{h}{l}$ wherever $x^2 + y^2$ is smaller than $2l^2$, which is the case inside the square. In fact, this observation rules out not only the energy (1) but also any energy that is the sum of an increasing function of the gradient modulus, which is to say most models.

2.3 Convex Models

To rule out the rest of the current prior models (and more), we can consider a functional of the form

$$E(S) = \sum_X f(\delta d^S), \tag{2}$$

where δd^S represents the derivative of some order of the depth function. For instance, the first-order case is the gradient such as in (1). The derivative δd^S, which in general is

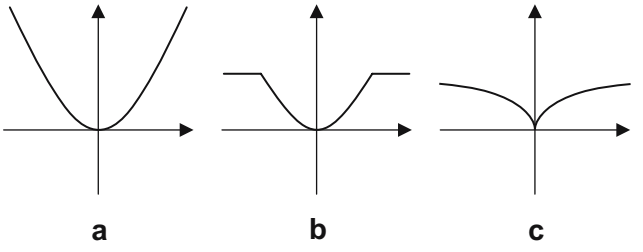

Fig. 3. Convex and non-convex functions of the magnitude. **(a)** A convex function. **(b)** A convex function with a cut-off value. **(c)** A concave function.

a vector, can be of any order, or a combination of several derivatives of different orders. Then, for a real number u between 0 and 1, we define a surface S_u that interpolates the two surfaces:

$$d^{S_u} = (1-u)d^{S_0} + ud^{S_1} \qquad (0 \le u \le 1).$$

We assume that f is a convex function of the derivative. In general, a function $f(x)$ that has the property

$$f((1-u)x_0 + ux_1) \le (1-u)f(x_0) + uf(x_1), \quad (0 \le u \le 1)$$

is said to be convex (see Fig. 3a.) If f is convex, then

$$E(S_u) \le (1-u)E(S_0) + uE(S_1) \le \max\{E(S_0), E(S_1)\}$$

implies that any linear interpolation of the two surfaces has the value of $E(S)$ that is at least as small as the larger of the values for the two surfaces. Moreover, if the energy is symmetric with respect to the sign inversion of depth, it would give $E(S_0) = E(S_1)$; and if the energy is strictly convex, the extremes S_0 and S_1 would be maxima among all the interpolated surfaces, not minima. All convex theories of which we are aware satisfy the latter two conditions. We conclude that the perceived surfaces are not predicted by any theory that uses the minimization of the energy function of the form (2) with convex f for disambiguation. Most current theories, including thin plate and harmonic, employ a convex energy functional as their prior, when seen in this representation. The minimization problem of the continuous version of (1) (called the Dirichlet integral) has the hyperbolic paraboloid $S_{\rm hp}$ as the solution.

2.4 Non-convex Models

Note that d^S's total sum is determined by the boundary condition. In order to minimize a sum $f(x) + f(y)$ of a convex function $f(x)$ while keeping $x + y$ constant, the value should be distributed as much as possible. Thus convex energy functions such as (1) tend to round the corners and smooth the surface. What, then, about functions that are not convex? More recent theories of stereopsis use sophisticated priors that model discontinuity in depth and slope. One such model (Belhumeur[2]) minimizes the second derivative of depth, except for certain locus where it gives up and allows discontinuities in slope, or a crease, making it non-convex. In effect, it uses a function $f(x)$ such as

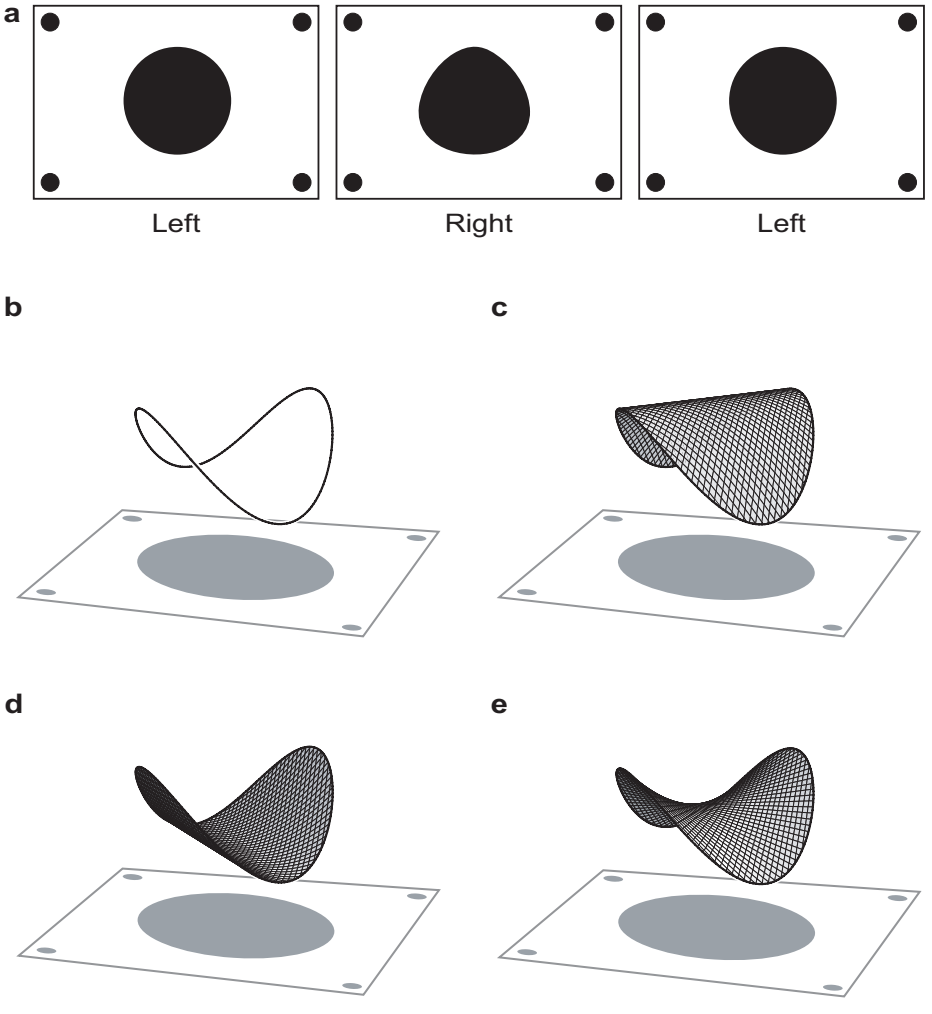

Fig. 4. Another stereogram further rules out possible theories. (a) Another stereo pair. (b) The unambiguous wire frame. (c), (d), and (e) are all possible surfaces that agree with the boundary condition. The human brain perceives either (c) or (d). Even algorithms that use non-convex functionals to allow discontinuities in depth and slope cannot have the solution (c) and (d) without having (e) as a better solution.

shown in Fig. 3b, which is still convex in low-modulus region but with a cut-off value beyond which the function value stays constant. This model actually can predict S_0 and S_1, with right parameter values, since both surfaces have zero second derivatives except at the crease, where the curvature can be as high as needed without any impact on the functional more than the cut-off value.

However, this model fails to predict the outcome on another stereogram, shown in Fig. 4a. Most observers reported a percept of one of the surfaces in either Fig. 4c

or d. Assume that the non-convex energy above predicts the outcome. That there are no creases in the two solutions indicates that the curvature stays below the cut-off value everywhere. Since the function is convex in this domain, and because any interpolation of the two solutions would also have no creases, the same argument as the convex case applies. It follows that any interpolation of the two surfaces would have lower function value than the higher of the two.

Going even further, we can think of using concave functions, such as shown in Fig. 3c. This is akin to minimizing $\sqrt{x} + \sqrt{y}$ while keeping $x + y$ constant, and tend to concentrate the value to fewer variables. If we use such a function $f(x)$ in (2) with a second-order δd^S, it would try to concentrate the second derivative at fewer points, and always predict a creased, piecewise-flat surface, never a smooth surface as shown in Fig. 4c or d.

2.5 Discussion

Thus, it seems no prior model in current computational model predicts the same surfaces as the human brain does. The experiment shows a clear tendency that the human vision has towards the opposite direction than such theories predict, leading us to the conclusion that the current computational theories of stereopsis are not very similar to the disambiguation by the human brain. The prior model that human vision seems to use is diametrically opposite to the current models in the sense that it predicts the *extreme* surfaces according to the energy functional representation of these models.

Is it possible that minimum disparity gradient or some similar models are used in human vision except when overridden by a strong prior preference for some special features? For instance, in the case of Fig. 1a, the observers' percepts might be biased towards S_0 and S_1 and away from $S_{\rm hp}$ because of the linear contours and sharp corners of the black square, since normally straight boundary contour edges derive from polygonal objects while curved contour edges derive from curved surfaces. However, the second experiment shows that even in the case where there are no straight edges, the percepts tend to be those of the extreme surfaces, rather than the hyperbolic paraboloid. Three of the observers were shown the round shape in Fig. 4a first (thus no bias because of the other pair) and still reported the percept of either convex or concave shape. Note that the same computational models are excluded by the second experiment alone, by the same argument as above. Thus, even if there is a bias towards linear surfaces for linear contours and sharp corners, it is not enough to explain the observation, nor does it change our conclusion.

Also, it has been demonstrated (Mamassian and Landy[14]) that human perception prefers elliptic (egg shaped) to hyperbolic (saddle shaped). Since the prediction of current theories is hyperbolic, the observed departure from it may be because of this bias. However, note that all the surfaces that are preferred are parabolic, i.e., neither elliptic nor hyperbolic. This is remarkable since the parabolic case constitutes a set of measure zero in the space of all possible local shapes. Because of this, it is hard to argue that any tendency or bias toward elliptic brought the percept exactly to that rare position.

3 Alternatives

The consideration of parabolic nature of the surfaces that are perceived by the human vision leads to a model that reflects this respect of human vision. Namely, the Gaussian curvature of the four preferred surfaces in the two experiments is zero everywhere it is defined. Zero Gaussian curvature is a characteristic of parabolic points. Surfaces with zero Gaussian curvature are developable, meaning they can be made by rolling and bending a piece of paper. In other words, one possibility is that the human vision system tries to fit a paper on the boundary wire frame (the sparse frame that represent definite depth data shown in Fig. 1b and Fig. 4b in the case of the experiments).

Thus, minimizing the total sum of the absolute value or square of Gaussian curvature, for example, may predict the surfaces similar to those that are perceived by humans. Such a functional would be neither convex nor concave. It is also nonlinear, which means that the solutions depend on the starting location; that makes the analysis of such a problem nontrivial, which is why we said it *may* predict the surfaces.

From a very different point of view, it is also noteworthy that in both of the examples the two surfaces most perceived by the human brain are the front and back of the convex hull of the boundary wire frame. A set in a space is called convex when any line segment that connects two of its points is also contained in it. The convex hull of a set of points is the minimal convex set containing all the points. In the case of Fig. 1b, the convex hull is the tetrahedron defined by the four endpoints of the line segments with definite depth data. This leads us to another model: a model that predicts surfaces that are a face of the convex hull of the depth points that are determined by matching.

Now, although these two models are very different, it turns out that they are closely related. That is, the Gaussian curvature of the surface of the convex hull of a set (such as the boundary wire frame) at a point that does not belong to the original set is zero, wherever it is defined. We are not aware of any mention of this fact in the literature; so we present it here as a theorem:

Theorem. *Let A be a set in the three-dimensional Euclidean space, B its convex hull, and p a point in $\partial B \setminus A$, where ∂B denotes the boundary of B. Assume that a neighborhood of p in ∂B is a smooth surface. Then the Gaussian curvature of ∂B at p is zero.*

Proof. Since p is in the convex hull of A, there are finite number of points $q_1, \ldots, q_n$ in A and positive numbers $a_1, \ldots, a_n$ such that $p = \sum_{i=1}^{n} a_i q_i$ and $\sum_{i=1}^{n} a_i = 1$, where q_i's are all distinct and $n \geq 2$, since p is not in A. Also, since p is on the boundary ∂B of a convex set B, all points of B are in the same half space H whose boundary ∂H is the tangent plane of ∂B at p. Since p is on the plane ∂H and all q_i's are in H, it follows that all q_i's are on ∂H because a_i's are all positive. (To see this, imagine a coordinate system in which p is at the origin and ∂H is the x-y plane; consider the z coordinates of q_i's, which we can assume are all on or above the x-y plane; if any of them had a positive z coordinate, so would p.) Consider the convex hull C of $\{q_1, \ldots, q_n\}$. Then C is in B, since B is convex. It is also in ∂H, since all q_i's are. Thus, it follows $C \subset \partial B$ since $C \subset B \cap \partial H$. Since $n \geq 2$ implies that C is not a point, the plane ∂H is tangent to the surface ∂B around p along at least a line segment. Thus the Gaussian curvature of ∂B at p is zero. □

This guarantees that, in a situation as in the experiments where there are points with definite depths and those with no information at all, we can take the convex hull of the points with depth information and take one of its faces to obtain a surface with minimum Gaussian curvature.

The minimization of Gaussian curvature seems a more familiar course for machine vision, while the convex-hull model gives some intuitive reason to think why these models might work better: the convex hull has the "simplest" 3D shape that is compatible with the data, much in the way the Kanizsa triangle (Kanizsa[11]) is the simplest 2D shape that explains incomplete contour information; and in the real world, most surfaces are in fact faces of some body; so it makes sense to try to interpolate the surfaces as such.

4 Conclusion

In this paper, we have reported the observations we made in examining human disparity interpolation using stereo pairs with sparse identifiable features. A mathematical analysis revealed that the prior models used in current algorithms don't have the same behavior as the human vision system; rather, they work in a diametrically opposite way. In discussing the implications of the findings, we have also proposed two quite different candidate models that reflect the behavior of human vision, and discussed the relations between them.

Acknowledgement. Hiroshi Ishikawa thanks the Suzuki Foundation and the Research Foundation for the Electrotechnology of Chubu for their support. Davi Geiger thanks NSF for the support through the ITR grant number 0114391.

References

1. N. Ayache. *Artificial Vision for Mobile Robots*. MIT Press. Cambridge, MA. 1991.
2. P. N. Belhumeur. "A Bayesian approach to binocular stereopsis". *Int. J. Comput. Vision* 19, pp. 237–262, 1996.
3. P. N. Belhumeur and D. Mumford."A Bayesian treatment of the stereo correspondence problem using half-occluded regions". In: *Proc. CVPR '92*, pp.506–512, 1992.
4. Y. Boykov, O. Veksler, R. Zabih. "Fast approximate energy minimization via graph cuts." *IEEE T. PAMI* 23, pp. 1222-1239, 2001.
5. O. Faugeras. *Three-Dimensional Computer Vision*. MIT Press. Cambridge, MA. 1993.
6. D. Geiger, B. Ladendorf, and A. Yuille. "Occlusions and binocular stereo". *Int. J. Comput. Vision* 14, pp. 211–226, 1995.
7. B. Gillam and E. Borsting. "The role of monocular regions in stereoscopic displays". *Perception* 17, pp. 603–608, 1988.
8. W. E. Grimson. *From Images to Surfaces*. MIT Press. Cambridge, MA. 1981.
9. H. Ishikawa and D. Geiger. "Occlusions, discontinuities, and epipolar lines in stereo." In *Fifth European Conference on Computer Vision*, Freiburg, Germany. 232–248, 1998.
10. J. Jones and J. Malik. *Image Vision Comput.* 10, pp. 699–708, 1992.
11. G. Kanizsa. *Organization in Vision*. Praeger. New York. 1979.
12. V. Kolmogorov and R. Zabih. "Computing Visual Correspondence with Occlusions via Graph Cuts." In *ICCV2001*, Vancouver, Canada. pp. 508–515.
13. J. Malik. "On Binocularly viewed occlusion Junctions". In: *Fourth European Conference on Computer Vision, vol.1*, pp. 167–174, 1996.

14. P. Mamassian and M. S. Landy. "Observer biases in the 3D interpretation of line drawings." *Vision Research* 38, pp. 2817-2832, 1998.
15. D. Marr and T. Poggio. "Cooperative computation of stereo disparity". *Science* 194, pp. 283–287, 1976.
16. D. Marr and T. Poggio. "A computational theory of human stereo vision". *Proc. R. Soc. Lond. B* 204, pp. 301–328, 1979.
17. K. Nakayama and S. Shimojo. "Da Vinci stereopsis: depth and subjective occluding contours from unpaired image points". *Vision Research* 30, pp. 1811–1825, 1990.
18. S. B. Pollard, J. E. W. Mayhew, and J. P. Frisby. "PMF: A stereo correspondence algorithm using a disparity gradient". *Perception*, 14, pp. 449–470, 1985.
19. G. Poggio and T. Poggio. "The Analysis of Stereopsis". *Annu. Rev. Neurosci.* 7, pp. 379–412, 1984.
20. S. Roy. Stereo without epipolar lines : A maximum-flow formulation. *Int. J. Comput. Vision* 34, pp. 147–162, 1999.
21. S. Roy and I. Cox. A maximum-flow formulation of the N-camera stereo correspondence problem. In *International Conference on Computer Vision*, Bombai, India. pp. 492–499, 1998.
22. D. Scharstein and R. Szeliski. "A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms". *Int. J. Computer Vision* 47, pp. 7–42, 2002.
23. R. Szeliski. "A Bayesian modelling of uncertainty in low-level vision". Kluwer Academic Press. Boston, MA. 1989.

Appendix. Methods

Seven observers naïve to the purpose of the experiment viewed the stereoscopic images in Fig. 1a and Fig. 4a. The images were presented on a 17-inch CRT monitor at a viewing distance of 1.5m through liquid crystal shutter goggles, which switch between opaque and transparent at 100Hz, synchronized to the monitor so that alternate frames can be presented to the left and right eyes, allowing stereoscopic displays. Images contained the black shape shown in the figures, the height of which was 10cm on the monitor surface. Four of the observers first viewed the image in Fig. 1a, and then Fig. 4a; the rest viewed the images in the reverse order. In each viewing, the observer was asked to describe what was perceived after 15 seconds; and then was asked to choose from the three pictures in Fig. 1c-e (when Fig. 1a is shown) or Fig. 4c-e. There was no discrepancy between what they described and what they chose. A few stereo pairs of color pictures were shown to each viewer prior to the experiment in order to ascertain that the observer is capable of binocular stereo perception. Only one of the observers reported the percept of a saddle-type shape (Fig. 1e). Other six viewers reported the percept of either convex (Fig. 1c) or concave (Fig. 1d) shape. One reported the percept of both of the convex and concave shapes.

Viewer	#1	#2	#3	#4	#5	#6	#7
Fig. 1a	convex	saddle	concave	concave	convex	convex	both
Fig. 4a	concave	saddle	concave	concave	convex	convex	convex
Which first?	Fig. 1a	Fig. 1a	Fig. 1a	Fig. 1a	Fig. 4a	Fig. 4a	Fig. 4a