Local qualitative shape from stereo without detailed correspondence

Extended Abstract

Shimon Edelman

Center for Biological Information Processing
MIT E25-201, Cambridge MA 02139
Internet: edelman@ai.mit.edu

Introduction

Schemes for the extraction of qualitative shape information from stereo tend to relegate the solution of the correspondence problem to a preprocessing stage and to assume that their input is provided in the form of a matched image pair – a disparity map [1]. The disparity map is a rich source of shape information, which allows the recovery of depth on a ratio scale [2] even when the camera geometry is unknown. Consequently, using the disparity map to extract qualitative rather than inexact quantitative shape information may be well-justified by considerations of stability and robustness [3], but appears to be wasteful in the sense that a rich and difficult to compute representation is transformed into a relatively terse one (Figure 1).

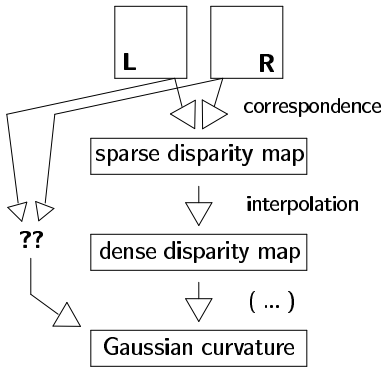

Figure 1: Can at least partial qualitative shape information be obtained from a stereo pair without going through a detailed disparity map?

Can at least partial qualitative shape information be obtained from a stereo pair without going through a detailed disparity map? A similar question arises in several approaches to model-based 3D object recognition, where object and model features have to be matched before the object's viewpoint transformation, used subsequently to align the object's image with the model, can be recovered (e.g., [4]). In object recognition the only alternative to detailed matching suggested so far is a class of methods that involve the computation of quantities that depend on entire objects rather than on their local features. One such method [5] exploits the observation that matrices formed by the 2D image-plane moments of rigid planar patch objects in 3D transform as tensors under a general viewpoint change (modeled, under orthographic projection, by an image-plane affine transformation). As an approach to recognition, the moment-based methods suffer from

an inherent sensitivity to occlusion and to imprecise or noisy segmentation. The idea behind the moment approach can be incorporated, however, into a qualitative vision module that provides information about the sign of the Gaussian curvature of surface patches through the use of regional rather than detailed correspondence (representation by the sign of the Gaussian curvature has been proposed, e.g., in [6,7]).

The idea

An arbitrary 3D viewpoint change can be modeled by an image-plane (2D) affine transformation if and only if the viewed object is planar. Let $\mathbf{p} = (x\ y\ z\ )^T$ and $\mathbf{p}' = (x'\ y'\ z'\ )^T$ be corresponding points in two views of the same object. Assume that $\mathbf{p}$ and $\mathbf{p}'$ are related by a 3D rotation $\mathbf{R}$ (the addition of 3D translation does not change the argument):

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix}$$

If the object is planar, that is, if for every point $\mathbf{p} = (x\ y\ z)^T$ the depth $z = Ax + By + C$, then the projections of $\mathbf{p}$ and $\mathbf{p}'$ are related by a 2D affine transformation:

$$\begin{aligned} x' &= r_{11}x + r_{12}y + r_{13}(Ax + By + C) = \\ &= (r_{11} + r_{13}A)x + (r_{12} + r_{13}B)y + r_{13}C \\ y' &= r_{21}x + r_{22}y + r_{23}(Ax + By + C) = \\ &= (r_{21} + r_{23}A)x + (r_{22} + r_{23}B)y + r_{23}C \end{aligned} \quad (1)$$

This transformation also relates the positions of the centroid (first-order moment tensor) of the planar patch as seen from the two viewpoints. Suppose that the patch is, in fact, not planar but hyperbolic (saddle-like). If the image of such a patch is divided into, say, four sectors with the origin at the centroid, some of the sectors will be in part closer to the two cameras than the tangent plane at the centroid, and others will be farther than the tangent plane. Moreover, the "near" and the "far" sectors will alternate as one goes around the patch centroid.[1] Consequently, the direction from the actual location of a sector's centroid towards its location as predicted by the "global" affine transform will change four times.[2] These changes can be detected by looking at the signs of the inner products of successive displacement vectors. The number of sign changes will be zero for elliptic or planar patches. Elliptic patches, however, can be detected by looking at the inside/outside difference in the signs of the relative depth values rather than at the sector-by-sector differences (see Figure 2B, right). A convex patch will have a positive (+) relative depth on the inside and a negative (−) one on the outside, and a concave patch – vice versa. Note that the depth is compared to that of a planar approximation to the surface, which is more or less parallel to the tangent plane at the center, but does not necessarily coincide with it (Figure 2A).

[1] More than four sectors may have to be looked at, e.g., if the parabolic lines at the center of the patch form an acute angle.

[2] The idea here is similar to Weinshall's [7], who related the number of sign changes around the center in the output of a simple operator to the sign of the Gaussian curvature of the surface.

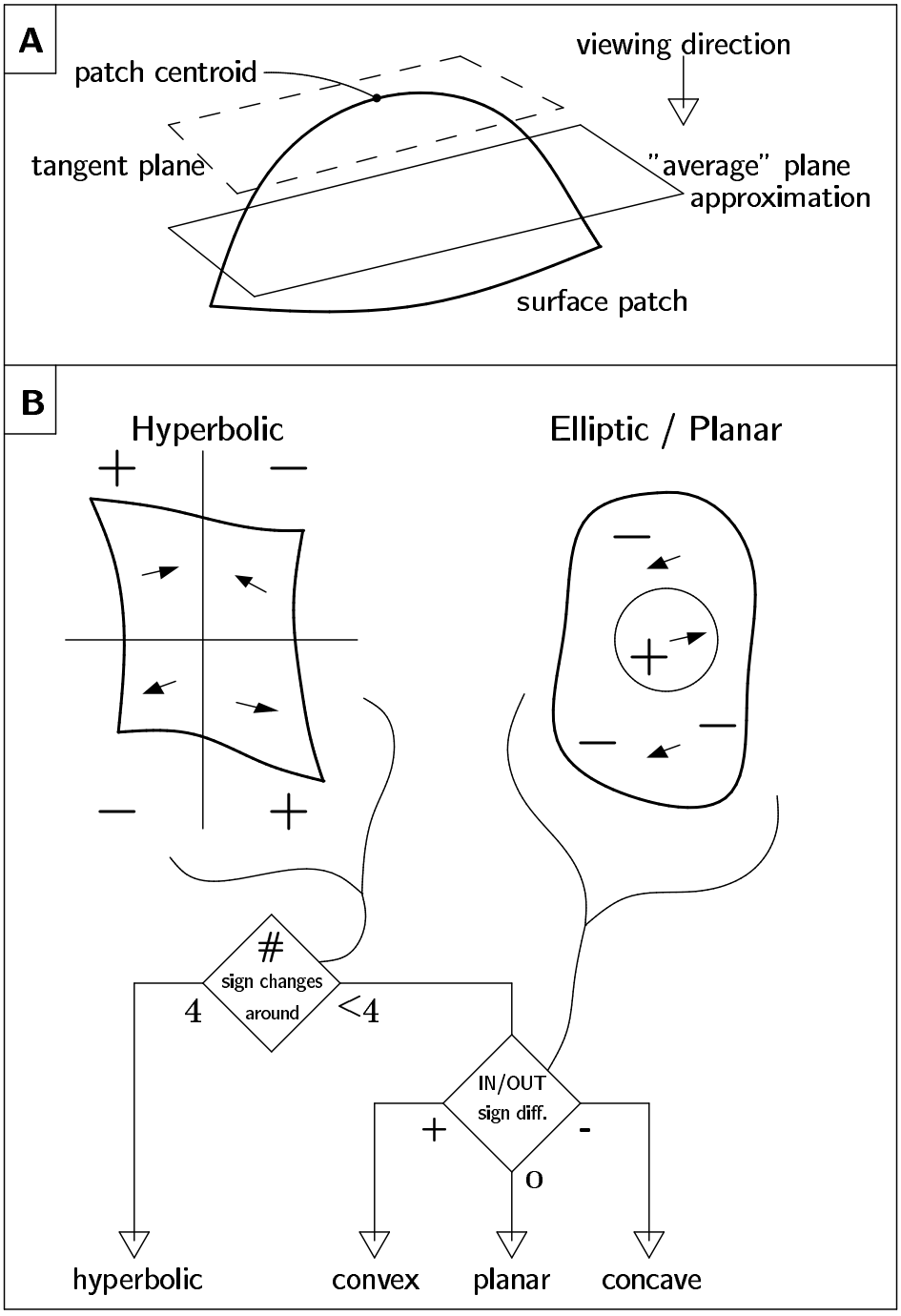

Figure 2: *(A)*, The "average" plane approximation to a convex patch. *(B)*, left: a hyperbolic patch is characterized by an alternation of the signs of depth values, where depth is computed relative to the "average" planar approximation. The arrows show the displacements of the sector centroids relative to the locations predicted by the global affine transform (see text). Right: an elliptic patch yields an inside/outside difference in the sign of the relative depth value rather than a sector-by-sector one. Bottom: the classification algorithm.

Underlying assumptions

The ability of the above method to come up with a correct qualitative description of the surface patch depends on several conditions:

- The "global" affine approximation to the transform relating the two input images must be known. One way to obtain such knowledge is through the use of the tensor method [5]. However, if the method is to be used in a binocular stereo setting, an approximate knowledge of the interocular distance would suffice.
- Since the method substitutes distance along the line of sight for the distance along the local normal to the surface, it works best when the two are close, i.e., when the slant and the tilt of the planar approximation to the surface are small. Alternatively, if the slant and the tilt are available independently (e.g., from a motion-based mechanism, or through a least-squares solution for A, B and C, given r_{ij} and several corresponding region centroids), the method can be modified to use them.
- While the method does not need feature to feature correspondence between the two frames, its performance depends on correct attribution of features to the sectors over which the centroids are computed (Figure 2B). Placing the origin of the reference system in each frame at the global centroid usually suffices for that purpose. Better yet, the origin may be placed at a prominent well-localized feature, if one can be identified in the two frames.

Implementation by receptive fields

As shown in Figure 2, the operation on which the present method is based can be carried out by a hard-wired mechanism such as a binocular "retinal" receptive field (RF) that can compute image centroids over its sub-fields and compare the results from the two images. The classification of the surface at each location can then be computed by combining the output of a "saddle detector" with that of an "egg detector" (Figure 2B).

Examples

Experiments with images of surface patches produced by a solid modeling system showed that the simple classification method outlined above works well both under orthographic and perspective projections and tolerates global surface tilt and slant that is comparable to largest difference in

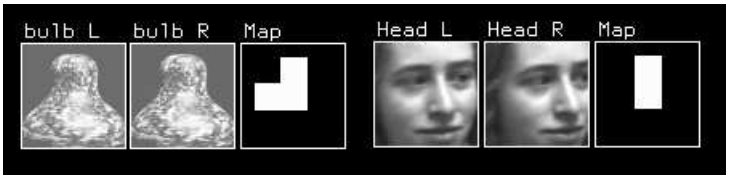

Figure 3: Left: A synthetic stereo pair, showing a textured pear-shaped object, and the output of a hyperbolic patch detector (white = hyperbolic). The pear's neck has been correctly classified as most likely to be hyperbolic. Right: The same operator applied to two frames of a motion sequence of a rotating human head. The bridge of the nose has been correctly classified as hyperbolic.

local normal orientation over the patch (that is, no part of the patch is allowed to be tangent to the line of sight because of the global tilt/slant) . The method can also be applied to raw (gray-level) synthetic images (Figure 3).

Summary

I have described a simple method that, under certain assumptions, is capable of extracting crude qualitative shape information (namely, the sign of the Gaussian curvature) from stereo without detailed correspondence. The method is based on centroid computation, an operation that can be implemented by a receptive field, and can be applied directly to gray-level images. Although the method is attractively straightforward and can serve as a part of a fast low-level qualitative shape module, it would probably be of little use in systems that can afford the computational effort of stereo matching and reconstruction. The question regarding the possibility of extracting more complex qualitative shape features without correspondence appears, at present, to have no definite answer.

References

[1] D. Weinshall. Application of qualitative depth and shape from stereo. In *Proceedings of the 2nd International Conference on Computer Vision*, pages 144–148, Tarpon Springs, FL, 1988. IEEE, Washington, DC.

[2] W. B. Thompson and J. K. Kearney. Inexact vision. In *Workshop on motion, representation and analysis*, pages 15–22, 1986.

[3] J. J. Koenderink and A. J. van Doorn. Depth and shape from differential perspective in the presence of bending deformations. *Journal of the Optical Society of America*, 3:242–249, 1986.

[4] S. Ullman. Aligning pictorial descriptions: an approach to object recognition. *Cognition*, 32:193–254, 1989.

[5] D. Cyganski and J. A. Orr. Application of tensor theory to object recognition and orientation determination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 7:662–673, 1985.

[6] P. J. Besl and R. C. Jain. Invariant surface characteristics for 3D object recognition in range images. *Computer Vision, Graphics, and Image Processing*, 33:33–80, 1986.

[7] D. Weinshall. Direct computation of 3D shape and motion invariants. A.I. Memo No. 1131, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, May 1989.