On Integration of Vision Modules*

Sharath Pankanti Anil K. Jain
Pattern Recognition and Image Proc. Lab.
Michigan State University
East Lansing, MI 48824

Mihran Tuceryan
European Computer-Industry Research Centre
München, Germany

Abstract

Individual cues from visual modules are fallible and often ambiguous. As a result, only integrated vision systems can be expected to give a reliable performance in practice. The design of such systems is challenging since each vision module works under different and possibly conflicting sets of assumptions. We have proposed and implemented a multiresolution system which integrates perceptual grouping, segmentation, stereo, shape from shading, and line labelling modules. The output of the integrated system is shown to be relatively insensitive to the constraints imposed by the individual modules.

1 Introduction

An important research issue in computer vision is whether the 3D structure of the scene could be reliably recovered from sensed image(s) in less assuming and more realistic situations. The intensity images convey 3D information in several different ways. Shape-from-X modules have been identified and shown to be capable of conveying shape of the object in constrained environments. Several other modules such as perceptual grouping have been demonstrated to be indirectly helpful in the depth recovery process. Each of these modules in itself has been known to have certain limitations. Quite often, the domains of applicability of each module are either disjoint or have a very little overlap. In a general situation, one does not expect a reliable recovery of depth information using a single module alone. A robust vision system should have a seamless integration of a number of different modules for obtaining correct depth information. Since each module works with a different set of (possibly conflicting) requirements (constraints), designing a synergistic integration of these modules is a challenging task.

Most of the shape-from-X modules are motivated by Marr's paradigm of modular design of a vision system [10]. However, more recent research studies [6] have shown that the assumption of the independence of vision modules is a gross simplification of the reality. It is now believed that the human vision involves a complex bottom-up and top-down control flow. Ideally, these processes could be modeled as a connection matrix with each state variable interacting with every other variable, thereby eliminating the concept of a vision module. To design such a system would be an ambitious goal; to maintain and extend such a system would be even more difficult.

A reasonable implementation of an integrated vision system would involve models of each module and models of their *interaction*. We believe that integration strategies based on explicit information exchange between the modules is a first step towards building more robust and tractable vision systems. This approach also emphasizes cooperation and *resonance* between the individual modules – which many researchers believe to be the key to the effectiveness of human vision system [5, 6]. Earlier efforts in integration assumed that the estimates obtained by the individual modules are reasonably accurate and hence adopted a feedforward strategy. However, an integrated system based on feedforward strategy alone is *critically* dependent on the performance of individual modules. Our effort here is to demonstrate the efficacy of a feedback-feedforward strategy for integration.

Integration models proposed in the literature are based on Bayesian [16], MRF [11], lattice-theoretic [7], game-theoretic [3], regularization [15], and energy minimization [2] formulations. It is reasonable to assume that each strategy is best suited for a certain type of interaction. The success of these interaction models critically depends on a number of user-specified parameters. It is hoped that the dynamics of the system with feedback obviates these elaborate models of interaction among the modules and replaces them with simpler interaction schemes facilitating the implementation of large integrated systems.

Another important issue to be considered in integration is that of the *emergent* behavior of the system. The vision modules often can not individually solve their assigned tasks unless they are given a specific model of the reality [1]. When one considers such an implementation of a single vision module, each simplifying assumption re-

*This work is supported by the NSF research grant IRI-9103143. This paper appeared in the *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 94)* , Seattle, WA, 1994.

quired by the module drastically reduces the scope of its application to real images. However, in an integrated system, the availability of additional cues lets the complete system transcend beyond the scope of fragile individual modules proscribed by their assumptions. Thus each of the modules could have failures, but the overall output of the integrated system is reasonably accurate. We will demonstrate that scope of the proposed integrated system goes beyond the nominal assumptions under which the individual modules are designed.

This paper discusses an implementation of the integration of perceptual organization, stereo, shape from shading, and line labelling modules. These modules were chosen primarily because of their importance in low-level vision. Also, they are known to interact with each other and are complementary in their strengths. Our strategy for integration can be extended to include additional modules. The overall block diagram of the proposed system is shown in Figure 1.

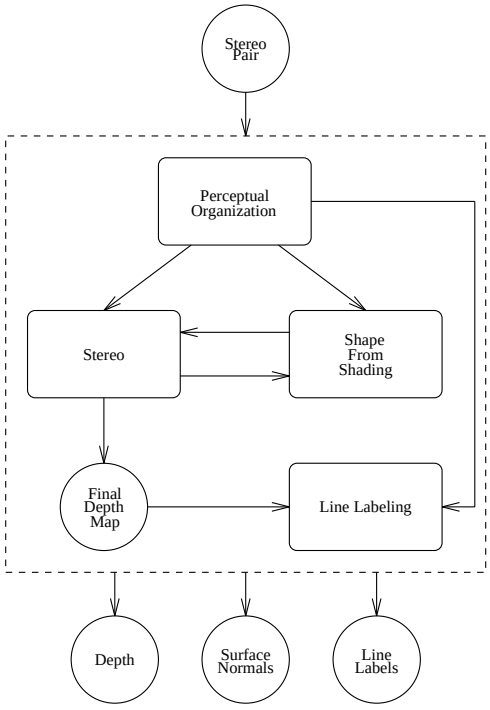


Figure 1: Overall Block Diagram.

In the next section we describe the proposed integration scheme. In Section 3 we will discuss our experiments and results. We will conclude with a summary of our experimental results and discussion.

2 Integration of Modules

In this section, we will discuss individual vision modules being integrated and our formulation of interaction among the modules.

Perceptual Organization Module: The objective of perceptual organization module is to "segment" the given image into regions with coherent photometric properties. In practice, poor imaging conditions, insufficient contrast, noise sensitivity of the selected attributes, and artifacts of segmentation operators all conspire to produce an imperfect segmentation.

We have used outputs of Canny edge operator and split and merge segmentation [14] as our edge-based and region-based segmentations, respectively. These two representations are combined in two phases. In the first phase, we select a subset of boundaries which are supported by both the segmentations. The boundaries thus obtained are tangentially extended based on the presence of boundaries in region-based *or* edge-based segmentations. In the second phase, the boundary terminations and corners in the representation obtained from the previous stage are linked based on *Gestalt* criteria (proximity, collinearity, cotermination) and the presence of gradient across the linking edge in the region-based segmentation. Instead of considering all possible linking edges, only the edges connecting Voronoi neighbors are considered [13].

Shape From Shading: We use the shape from shading algorithm developed by Oliensis and Dupuis [12]. Given an intensity image and the depth values at singular points in the image, their method of reconstructing the characteristic strips is relatively insensitive to noise.

We have extended this algorithm to piecewise constant albedo surfaces. The extension is based on segmenting the image into regions with constant albedo and treating them autonomously. The singular points are now detected in individual regions and the propagation of the depth values is prevented across the region boundaries. Such a treatment, when implemented in an isolated module, is plagued with erroneous depth recovery due to the following reasons: (i) the stereo module, needed for initial depth values at singular points, itself might provide erroneous depths; (ii) the Lambertian model may not be an accurate model for the image surfaces; and (iii) interreflections and specularities may also cause inaccurate reconstruction.

Stereo Module: Stereo is a robust estimator of depth at an acute visual angle, particularly in image regions with a significant variation in intensity.

Given a pair of stereo images, the problem is to find corresponding physical points in the two images and then compute the depth at (possibly) each location in the image. We use the multi-resolution stereo matcher proposed by Weng *et al.* [18] which uses a gradient-descent technique to maximize the correlation of certain attributes be-

tween the stereo pair. In regions with small changes in intensity, there is not sufficient information available for a gradient-descent technique to drive the disparities in the correct direction. Fortunately, these are the regions where the shape from shading performs well.

Interaction between Shape from Shading and Stereo Modules: According to Bulthoff *et al.* [4], the depth conveyed by disparity overrides the information conveyed by shading. They also have accumulated evidence to support the conjecture that in scenes with sparse zero crossings, shape from shading interpolates depth in the area between two zero crossings. Several attempts have been made to integrate the shape from shading and stereo modules. All of these earlier integration schemes assumed a single constant albedo over the surfaces comprising the entire scene. Further, they do not include any strategy for two-way interaction between the stereo and shape from shading nor does the shading guide the computation of stereo disparities.

In our approach, we have a more general model of the surface reflectance of the scene and a more reliable and robust strategy of treating the scenes which deviate from the underlying assumptions. Specifically, the feedback loop between the shape from shading and stereo modules is designed to counter the limitations of both the modules. When there is sparse texture in the scene, the shape from shading module will interpolate the surface depths between the (sparse) zero crossings. On the other hand, when the reconstruction offered by shape from shading is in error, the gradient descent iteration of the stereo module will attempt to force the resultant reconstruction towards the true depth value. This synergistic cooperation is the essence of the efficient treatment of their individual errors.

We use disparity correction function obtained from depth recovered by the shape from shading to *iteratively* drive the disparities in the correct direction. The disparity correction function operates on the following principle. Given a disparity map at $(n-1)^{th}$ iteration, D^{n-1}, and depth maps predicted by shape from shading module, D_{ss}, and stereo module, D_{st}, the integration module computes a corrected disparity map, D^n. Inputs from both the modules can not be combined directly and we need to extract relevant information from each module to update the disparity map at each resolution. Let us call the corrected disparity map contributed by the stereo depth map to be $\tilde{D}_{st}$. The corrected depth map provided by shape from shading module will be called $\tilde{D}_{ss}$. A location in the image is defined to be *distinctive* if intensity gradient at that location is sufficiently large. $\tilde{D}_{st}(i,j)$ at location (i,j) is determined by linearly interpolating the depth values at its nearest distinctive epipolar neighbors in the region, (i,k) and (i,l). The numerical value of $\tilde{D}_{ss}(i,j)$ at each location is given by integration of surface normals from (i,k) to (i,j) and from (i,l) to (i,j). The corrected disparity at pixel (i,j) at iteration n is given by

$$\begin{aligned} D^n(i,j) &= f(D^{n-1}, D^n_{st}, D^n_{ss}). \\ &= \tilde{D}^n_{st}(i,j) + \alpha \tilde{D}^n_{ss}(i,j). \end{aligned}$$

Parameter α is the coupling coefficient between the shape from shading module and the stereo module. Notice that correction is not based on any precise calibration, but is set to an arbitrary monotonic function of the depth depending on the value of α. In practice, we have seen that the performance of the system does not critically depend on the value of α as long as it is sufficiently small ($\alpha \leq 0.01$).

The flow of information from the stereo module to shape from shading module is relatively straightforward. The depths at the singular points are initialized to the corresponding depth values predicted by the stereo module. The concavity or convexity of the surface at the singular points is also estimated from the depth map obtained from the stereo module.

Interaction between the depth modules and Segmentation Module: Regions formed by the perceptual boundaries are filled-in with the appropriate features (like color, brightness, texture, etc.) and the perceptual boundaries act as feature barriers [6]. Many vision researchers have been using intensity gradient as a deterrent to smoothing across regions [18], but perceptual boundaries have not been used frequently for this purpose.

In shape from shading module we have prevented the propagation of depth values across the perceptual boundary. Similarly, in stereo module, the smoothness constraint is not enforced across perceptual boundaries. This has significantly reduced the blurring of the sharp depth boundaries in the recovered depth map. We also set the depth values at perceptual boundaries to the values predicted by the stereo module – since the reliability of the stereo module is best in these regions.

Line Labelling: A boundary detected in an image could be a result of a number of different physical events such as discontinuity in illumination (shadow), a change in surface albedo, surface markings, discontinuity in surface orientation, discontinuity in depth, and self occlusion. The line labelling module labels the boundaries detected by the segmentation module into various different categories.

The line labelling module is unique in terms of the geometrical constraints it imposes on the 3D interpretation. Although the line labelling problem has been rigorously studied for a limited object domain, it is plagued with innumerable implementation problems due primarily to noisy boundaries.

The boundaries detected by the segmentation mod-

ule are parsed into a graph represented by vertices and arcs. Arcs representing albedo edges, shadow edges, and surface markings are separated from the rest of the boundaries based on the output of the two depth modules by measuring the depth gradient across each segmented boundary and comparing it with a threshold value. We now describe classification of rest of the arcs into occluding (limb) edges and surface normal discontinuities.

Detection of Limb Edges: Ideally, the surface normal at the limb edges should be perpendicular to the viewing direction. However, the accuracy of the surface normal direction estimated from the depth modules is limited and the detected limb edges would not be reliable if we depend on this information alone. For this reason, we base our decision on the trends offered by an ensemble of the surface normals in the vicinity of a boundary. Let us call the angle defined by the surface normal (n_i), the viewing direction, and the boundary normal (N) as θ. Usually, these angles decrease monotonically as we move in the direction of boundary normal pointing inside the region enclosed by the limb edge. For non-limb edges, this trend should be considerably less significant. We have devised the following test for determining the statistical significance of this trend.

Suppose we are investigating "limbness" of a point b on the boundary B and the two estimated normals of the boundary are in the directions N and $-N$. Let $n_1, n_2, n_3, \cdots, n_M$ be the M surface normals sampled in the direction N at distances $d_1, d_2, d_3, \cdots, d_M$. We assume that the sampling is guided by segmentation to avoid samples coming from different surfaces of the scene. Construct a separate rank ordering of the angles θ_i and distances (breaking any existing ties randomly) to obtain two rank sequences r_i and s_i. We compute the Spearman's statistic, S, for measuring rank-order correlation: $S = (\sum_i ((r_i - \overline{r})(s_i - \overline{s})))/(\sqrt{(r_i - \overline{r})^2}\sqrt{(s_i - \overline{s})^2})$. The significance of a negative value of S is tested by computing $R = S\sqrt{(M-2)/(1-S^2)}$ and comparing it to a threshold with a size of 0.05. This test is supplemented by the average starting angle test. This test evaluates whether arithmetic mean of the first few samples of θ_i is larger than certain threshold, t. Both of these tests are repeated for the samples of surface normals taken in the direction of the other boundary normal ($-N$). A point is assessed to be part of a limb edge if at least one set of surface normals passes both the trend test as well as the average starting angle test. If a significant fraction (70%) of the points belonging to an arc pass the test, the entire arc is considered to be a limb.

Malik's Labelling Algorithm: Given a graph representation of the boundaries in the image and an ordered list of labels based on the classification provided by the limb detection module, we determine the most consistent label for each of the arcs and junctions in the graph. We use the line labelling module for curved objects proposed by Malik [8] who provided an explicit catalog of legal line labels for objects with C^3 surfaces.

Attempts in integrating shape from shading and line labelling for curved objects have been reported by Malik and Maydan [9]. They formulated the integration problem as an optimization problem to simultaneously recover surface orientation and line labels. Their cost function also includes a regularization term. However, their problem has been formulated for scenes composed of a single constant albedo and their experiments were limited to synthetic images. Trytten has attempted to label the line drawings obtained from perceptually grouped edgels [17]. Our approach is an improvement of the strategy proposed by Trytten [17] and exploits the powerful constraints exerted by the limb boundaries to disambiguate the line drawing interpretation.

The proposed algorithm assumes that the image is already segmented into background and foreground regions. Based on this knowledge, it separates the arcs bounding the regions into "inside" and "outside" arcs. The arcs are initially classified into "limb" and "non-limb" boundaries based on the output of the limb detection algorithm. The junctions are then labeled based on their degree, Malik's junction catalog, and the existing (if any) labels of their adjacent arcs. The ambiguities and inconsistencies in junction labels are removed by using following strategies (in the order described below): (i) If the length of relevant arcs is sufficiently small (a few pixels) and degree of the junction is 2, the junction is classified as a *phantom* junction; (ii) Angle and curvature measurements are considered in disambiguating various junction types; (iii) Finally, if the inconsistencies still exist, the initial labels of the arcs are reconsidered.

3 Experiments

Our experimental results will be discussed in the context of the quality of reconstruction obtained from the integrated system. The experiments are primarily designed to demonstrate the graceful deterioration in the performance of the system as the assumptions made in individual modules are violated. Our system was tested on a number of images of different synthetic and real scenes. We will now briefly describe our imaging setup before presenting our results.

All the images were captured by an inexpensive CCD camera (Panasonic GP-KR202, $f = 25\ cm$, maximum aperture). The images were subsequently gamma corrected with $\gamma = 2.0$ and normalized to 256 gray levels. The stand off was approximately $80\ cm$ and to obtain a stereo pair of images, the camera was either translated or translated and rotated. The translation was in the direction of *x–axis* and rotation was about *y–axis* (*z–axis* being ap-

proximately aligned with the optical axis). The imaging setup was not calibrated; all alignments, translations, and rotations were approximate and were not precisely measured/verified. The rotation of the camera was effected to bring disparity of the region of interest close to zero. The scene was illuminated with ambient light and a single incandescent light source located (30 cm) behind the camera (approximately in x-z plane) pointing in the direction $(0, 0, 1)$. A polarizing setup was used (when necessary) to reduce the specular component of the reflection.

All the reconstruction results are presented for the right image of the stereo pair and, for all the experiments, α was set to 0.001. Further, the shape from shading constraints were exploited only for the four finest image resolutions (64x64, 128x128, 256x256, and 512x512). Use of shape from shading module for lower resolution representations did not significantly improve the results.

Figures 2(a) and 2(b) show stereo images of an unglazed ceramic object (mushroom) and an object made of acrylic plastic (Y-shaped pipe). These images were captured without cross-polarized filters to allow the specular reflections on the pipe to be imaged. Notice the two specularities on the surface of the pipe. Figures 2(c) and 2(d) show the (relative) depth reconstruction obtained by the stereo module and the integrated system, respectively. The diffusion of disparities across the perceptual boundaries significantly blurs the depth map by the stereo module; this is prevented in the depth map obtained by the integrated system. Figures 2(e) and 2(f) show the orientation of the object surfaces for stereo and integrated system, respectively. The quality of the reconstructed depth map of the integrated system as shown in Figures 2(d) and 2(f) has largely remained insensitive to the specular reflections.

Figures 3(a) and 3(b) show stereo images of an unglazed ceramic object (*egg*) and a foam cup. Figures 3(c) and 3(e), respectively, show the depth reconstruction and surface needle map obtained from the isolated stereo system. Figures 3(d)and 3(f) show the corresponding representations for the integrated system. Very bright (low depth values) regions between the objects and to the far left in the image are due to occlusion. Notice that the quality of reconstruction of surfaces of (foam) cup in case of the stereo module alone (Figures 3(c) and 3(e)) is comparable to that of the integrated system (Figures 3(d) and 3(f)). The depth reconstruction provided by the integrated system is relatively more accurate except for the slight deterioration on the lower part of the egg. The surface of the egg has been incorrectly reconstructed by the isolated stereo module as seen in Figures 4(a) and 4(b). The quality of reconstruction of the surface of the cup is comparable in both systems due to presence of coarse texture on its surface.

Figures 5(a) and 5(b) show the outputs of the line labelling algorithm proposed by Trytten [17] and our algorithm, respectively for Mushroom and Pipe image. Note that all the *curvilinear-L* junctions are correctly labeled by our algorithm. Note also that the labelling of the *T-junctions* is a reasonable approximation of the physical reality, given the quality of output generated by the segmentation module.

We have quantitatively evaluated the surface reconstruction from the integrated system using synthetic images. Synthetic stereo images were generated from range images by assuming the positions of cameras and baseline similar to the imaging setup described above. An additive *i.i.d.* Gaussian noise (2%) was then added to the left and right images separately. Using the known ground truth, error in final reconstruction was measured using a squared difference objective function. Compared to the system using the stereo module alone, 5-8% improvement was observed in the estimation of depth and surface orientation using the integrated system.

4 Conclusions

A reliable vision system should consider all visual cues to obtain a meaningful and unambiguous interpretation of the input scene. However, the information provided by each visual cue is often based on a different set of assumptions. This raises several important research issues in solving the integration problem: (i) What is the most reliable information provided by each visual cue?, (ii) How to design an integrated system which can be easily maintained and extended?, and (iii) How to integrate vision modules so that system performance does not critically depend on individual modules? In this paper we have made an attempt to explore these issues by integrating the following four modules: perceptual organization, shape from shading, stereo, and line labelling.

We have proposed and implemented an integration framework emphasizing *interaction* and information exchange among the four vision modules. We also demonstrate the consistent performance of the integrated system even in the adverse situations where one or more assumptions made by the individual modules are violated. The numerical accuracy of the recovered depth is assessed in case of synthetically generated data. We have also qualitatively evaluated our approach by reconstructing geons from the depth data obtained from the integrated system.

In general, the potential of a system relying on low-level modules has been grossly underestimated because of their individual vulnerabilities. Our results show that an integrated system comprising of several low-level modules can deliver adequate performance without significant information from the top-down, knowledge-based modules.

Acknowledgments We are grateful to John Oliensis and John Weng for making their shape from shading and stereo software available to us. We thank P. Tsai for providing synthetic images.

References

[1] Y. Aloimonos and D. Shulman. *Integration of Visual Modules: An Extension of the Marr Paradigm*. Academic Press, San Diego, CA, 1989.

[2] A. Blake and A. Zisserman. *Visual Reconstruction*. The MIT Press, Cambridge, Massachusetts, 1987.

[3] H.I. Bozma and J.S. Duncan. Integration of vision modules: A game-theoretic framework. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 501–507, Maui, Hawaii, 1991.

[4] H. H. Bulthoff and H. A. Mallot. Integration of depth modules: Stereo and shading. *Journal of Optical Society of America*, 5(10):1749–1758, October 1988.

[5] F. Crick. Function of the thalamic reticular complex: The searchlight hypothesis. In *Proceedings of the National Academy of Sciences*, pages 4586–4590, 1984.

[6] Stephen Grossberg, editor. *Neural Networks and Natural Intelligence*. The MIT Press, Cambridge, Massachusetts, 1988.

[7] A. Jepson and W. Richards. A lattice framework for integrating vision modules. *IEEE Transactions on Systems, Man, and Cybernetics*, 22:1087–1096, 1992.

[8] J. Malik. Interpreting line drawings of curved objects. *International Journal of Computer Vision*, 1:73–103, 1987.

[9] J. Malik and D. Maydan. Recovering three–dimensional shape from a single image of curved objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(6):555–566, June 1989.

[10] D. Marr. *Vision*. W.H. Freeman and Co., San Francisco, 1982.

[11] Sateesha G. Nadabar and Anil K. Jain. Edge Detection and Labeling by Fusion of Intensity and Range Images. *Proc. of SPIE Conf. on Applications of AI: Machine Vision and Robotics*, 1708:108–119, 1992.

[12] J. Oliensis and P. Dupuis. Direct method for reconstructing shape from shading. In *Proc. of SPIE conference on Geometric Methods*, volume 1570, pages 116–128, San Diego, California, July 1991.

[13] S. Pankanti, Anil K. Jain, and M. Tuceryan. On integration of vision modules. Technical Report TR–PRIP1, Michigan State University, February 1994.

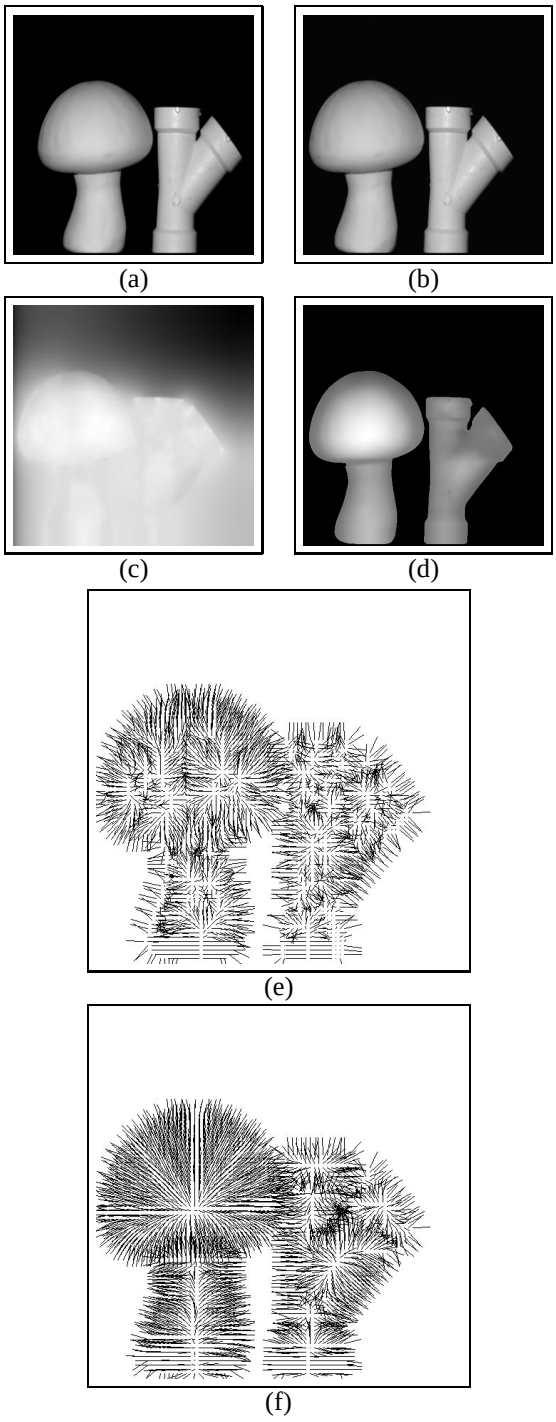


Figure 2: Mushroom and Pipe image (size 512x512): (a), (b) Left and Right stereo images; (c) Recovered depth from stereo alone; (d) Recovered depth from the integrated system; (e) Recovered surface normals from stereo alone; (f) Recovered surface normals from the integrated system.

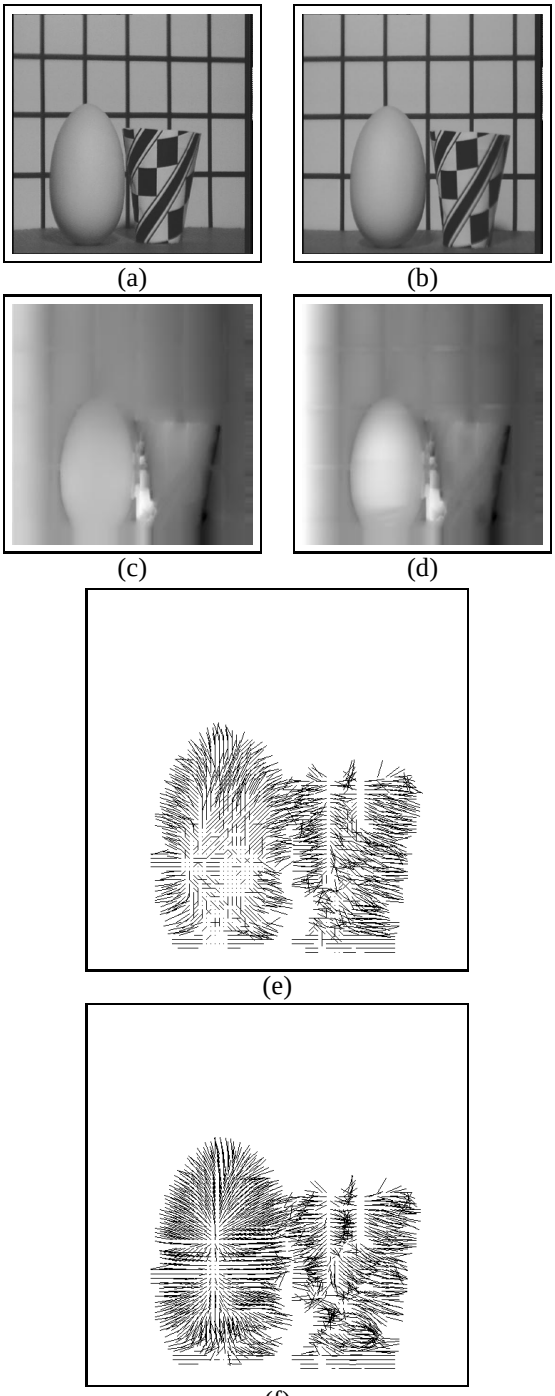


Figure 3: Egg and Cup image (size 512x512): (a), (b) Left and Right stereo images; (c) Recovered depth from stereo alone; (d) Recovered depth from the integrated system; (e) Recovered surface normals from stereo alone; (f) Recovered surface normals from the integrated system.

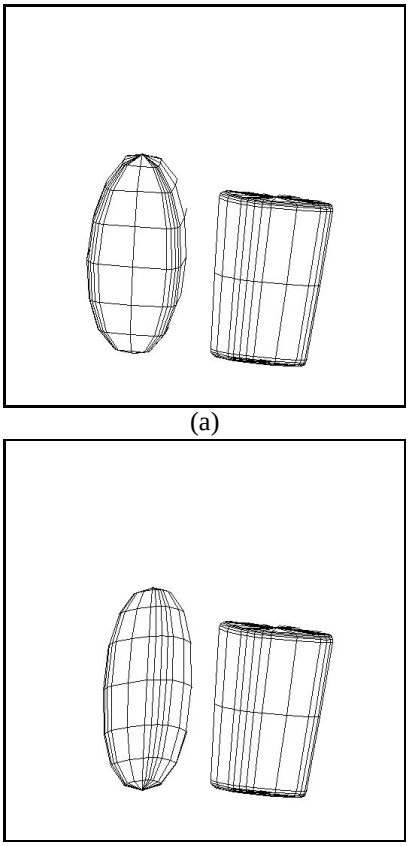


Figure 4: Recovery of 3D shapes for Egg and Cup scene (Fig. 3): (a) Recovered superquadrics from stereo alone; (b) Recovered superquadrics from the integrated system.

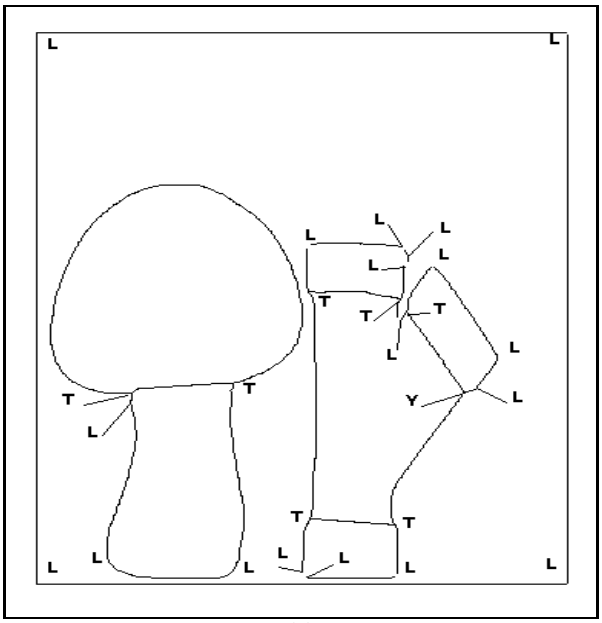
[14] T. Pavlidis. *Algorithms for Graphics and Image Processing*. Computer Science Press, Rockville, Maryland, 1982.

[15] T. Poggio, V. Torre, and C Koch. Computational vision and regularization theory. *Nature*, 317:638–643, 1985.

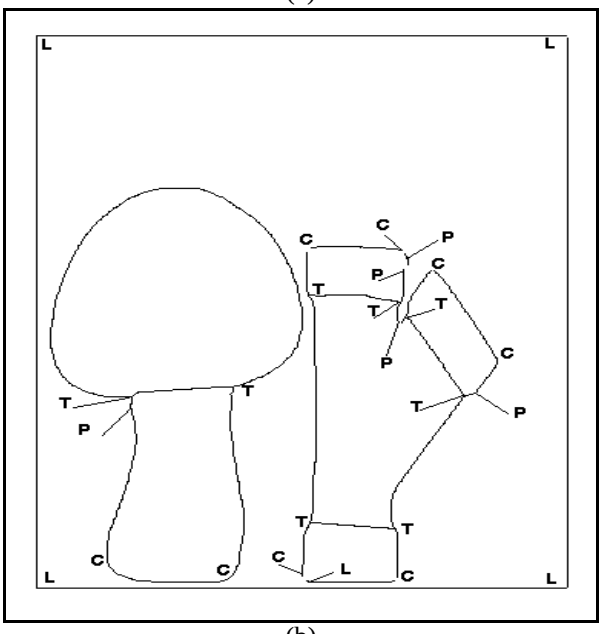
[16] S. Sarkar and K.L. Boyer. Perceptual organization using Bayesian networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 251–256, Champaign, Illinois, 1992.

[17] Deborah A. Trytten. *Integrating Diverse Perceptual Modules to Create a 2.5 Dimensional Sketch*. PhD thesis, Michigan State University, E. Lansing, Michigan, 1992.

[18] J. Weng, N. Ahuja, and T. S. Huang. Matching two perspective views. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(8):806–825, August 1992.



(a)



(b)

Figure 5: Junction labeling results for Mushroom and Pipe image (Figure 2): (a) using line labeling alone; (b) by the integrated system using the information provided by the shape from shading module; L, C, T, Y, A, and P denote L, curvature-L, T, Y, arrow, and phantom junctions, respectively.