Binocular and Multiple View Stereo Using Tensor Voting

1. Research Team

Project Leader: Prof. Gérard Medioni, *Computer Science*

Graduate Students: Philippos Mordohai

Industrial Partner(s): Geometrix Inc.

2. Statement of Project Goals

We propose to treat the problem of binocular and multiple view stereo from a perceptual organization perspective using the Tensor Voting framework we have developed [4]. Under our approach feature matching and structure reconstruction are addressed within the same framework. In order to handle noise, lack of image features, and discontinuities, we adopt a tensor representation for the data and tensor voting for information propagation.

3. Project Role in Support of IMSC Strategic Plan

Our ability to generate models from images can be very useful to other projects at IMSC. It provides models to research related to graphics, animation, and haptics without having to resort to expensive and cumbersome laser range scanners. Other possible applications of our research in conjunction with other IMSC research projects include teleconferencing and immersipresence, e-commerce, scene analysis and cartography.

4. Discussion of Methodology Used

The premise of shape from stereo comes from the fact that, in a set of two or more images of a static scene, world points appear on the images at different disparities depending on their distance from the cameras. Establishing pixel correspondences on real images, though, is far from trivial. Projective and photometric distortion, sensor noise, occlusion, lack of texture, and repetitive patterns make matching the most difficult stage of a stereo algorithm. Significant progress has been made in this area, but the correspondence problem has not been solved. Most of the remaining difficulties are caused by occlusion and lack of texture.

We propose an approach that addresses stereo within a perceptual organization framework, taking into account both binocular and monocular sources of information. Both geometric and color information from the scene are used for grouping, complementing each other's strengths. We begin by generating matching hypotheses for every pixel. A variety of matching techniques can be integrated, thus allowing us to combine their particular advantages. Correct matches are detected based on the support they receive from their neighboring candidate matches in 3-D, after tensor voting. They are grouped into smooth surfaces, the projections of which on the images serve as the reliable set of matches. Segmentation using geometric properties is arguably the most significant contribution of our research. The inferred reliable set of matches guides the

generation of disparity hypotheses for the unmatched pixels. The match for such a pixel is selected among a set of candidates as the one that is a good continuation of the surface, and also compatible with the observed color distribution of the surface in both images. Thus, information is propagated from more too less reliable pixels considering both geometric and color information. We present results on standard stereo pairs that would rank first at the time of submission in the Middlebury College Stereo Evaluation webpage (http://www.middlebury.edu/stereo/).

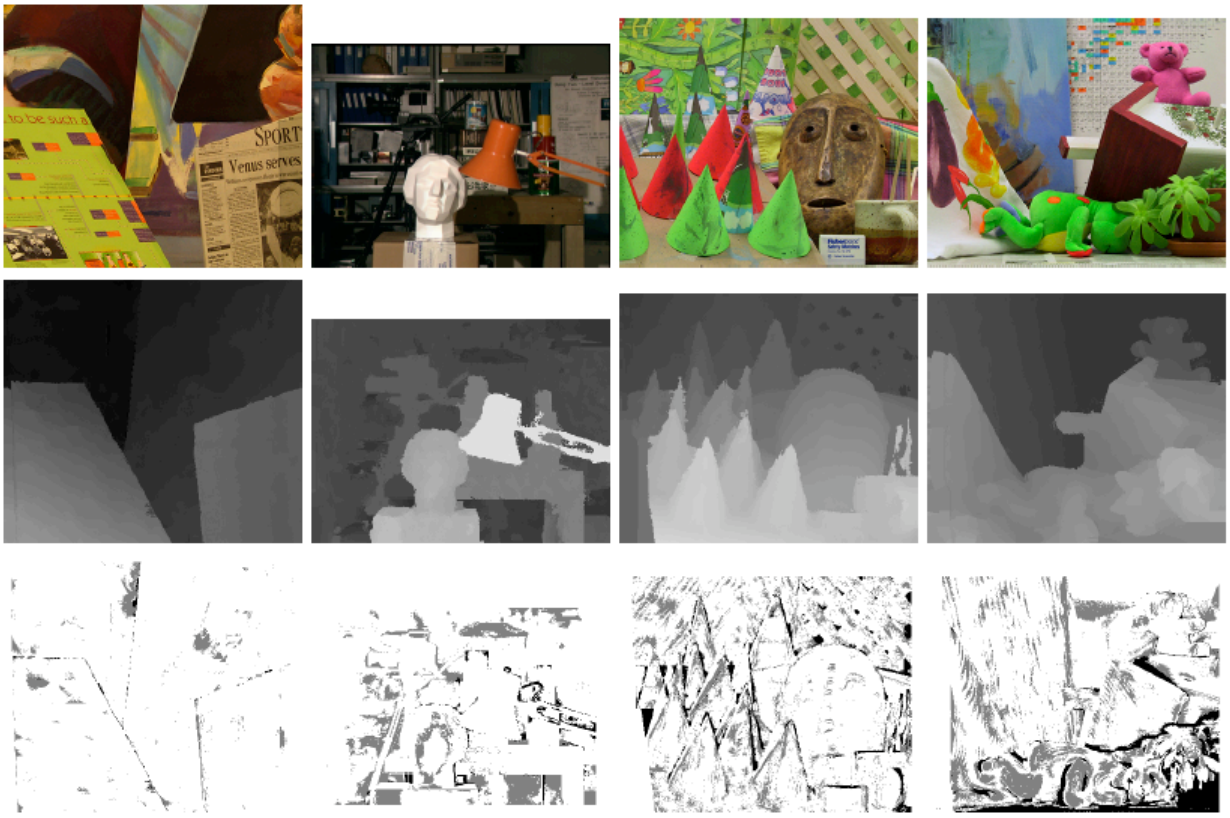

Figure 1: Input images (top row), disparity maps (middle row) and error maps (bottom row) for the Venus, Tsukuba, cones and teddy datasets

Figure 1 shows the left image of the stereo pair, the final disparity map and the error map for four of the datasets of the evaluation. Gray in the error maps corresponds to disparity errors between one half and one disparity level, while black corresponds to errors above one disparity level.

5. Short Description of Achievements in Previous Years

During the initial phase of this research, the Tensor Voting framework was developed and tested. Techniques for extracting regions, surfaces and curves from noisy three-dimensional datasets were developed [2] and stereovision was an excellent area to apply them. The problem of stereo reconstruction is a challenging and demanding one, thus aiding the development of our methodology, while, on the other hand, the Tensor Voting framework is well suited for the various problems that arise in the processing of stereo data. After our original binocular stereo

approach was published [1][3], we focused our attention on achieving performance comparable to the state-of-the-art in the binocular case and on developing a leading approach to multiple view stereo.

In 2001, we further confirmed the validity of our approach by applying it to classical stereo pairs and achieving good results. We have enforced constraints that are specific to the problem at hand via the Tensor Voting framework. In the case of multiple view stereo, the concept of consistency of matches was introduced by increasing the initial saliency of matches confirmed by more than two views. These modifications proved very effective as demonstrated in [5]. As far as the Tensor Voting framework is concerned, there were significant steps of progress, the most effective of which so far was the introduction of first-order tensor information that complements the second-order information and accommodates very accurate discontinuity and boundary detection. Boundaries play a major role in perceptual organization and data visualization, and they are the foundation of the multiple-scale approach we plan to use.

In 2002, our preliminary approach to binocular stereo was published in the most highly esteemed journal in the field, the IEEE Transactions on Pattern Analysis and Machine Intelligence [3]. Later, a paper on multiple view stereo was presented in the International Conference of Pattern Recognition sponsored by the IEEE and the IAPR [6]. Finally, we were invited to contribute a chapter, containing both details of the Tensor Voting framework and results on stereo, to a book titled "Handbook of Computational Geometry for Pattern Recognition, Computer Vision, Neurocomputing and Robotics".

5a. Detail of Accomplishments During the Past Year

In 2003, our first order augmentation to the tensor voting framework was accepted for publication in the most highly esteemed journal in the field, the IEEE Transactions on Pattern Analysis and Machine Intelligence [13]. Also, a paper on our approach to binocular stereo, as described in the previous section, was submitted and accepted for an oral presentation in the European Conference on Computer Vision, 2004 [12]. Finally, we prepared and submitted chapters describing our work for three scientific books:

- G. Medioni, P. Mordohai, and M. Nicolescu, "The Tensor Voting Framework", to appear in the Handbook of Computational Geometry for Pattern Recognition, Computer Vision, Neurocomputing and Robotics, E. Bayro-Corrochano (editor), Springer-Verlag, currently in press, 2004 [7]
- G. Medioni and P. Mordohai, "The Tensor Voting Framework", to appear in Emerging Topics in Computer Vision, S.B. Kang and G. Medioni (editors), Prentice Hall, currently in press, 2004 [8]
- G. Medioni and P. Mordohai, "Saliency in Computer Vision", to appear in Neurobiology of Attention, L. Itti, G. Rees, and J. Tsotsos (editors), Elsevier Science, 2004 [9]

6. Other Relevant Work Being Conducted and How this Project is Different

Due to lack of space we refer interested readers to [10] for a comprehensive review of binocular stereo algorithms. A multitude of approaches have been used including dynamic programming, non-linear optimization, relaxation labeling, clustering, graph cuts, stochastic optimization, non-

linear diffusion, cooperative stereo and adaptive windows. Each method has different advantages and can be applied to broader or more limited types of scenes according to the assumptions made. Good results have been obtained by all the afore-mentioned techniques, especially the ones who explicitly consider occlusion and lack of texture and attempt to correct the errors caused by them.

We have presented a novel stereo algorithm that addresses the limitations of binocular matching by incorporating monocular information. Careful initial matching and the use of smoothness, with respect to both surface orientation and color, complement each other to derive more information from the stereo pair. Textured pixels are typically resolved by binocular matching, while untextured ones by the smooth extension of neighboring surfaces guided by color similarity. Arguably the most significant contribution is the segmentation into layers based on geometric properties and not appearance. We claim that this is advantageous over other methods that use color-based segmentation, since it utilizes the already computed disparities, which are powerful cues that provide very reliable initial estimates for the color distribution of layers.

Other contributions include the initial matching stage that allows the integration of any matching technique without any modification to subsequent modules. Information propagation in 3-D via tensor voting eliminates interference between adjacent pixels from different world surfaces and makes saliency the criterion for the disambiguation of matches. The proposed color similarity model works very well, despite its simplicity, because, locally, similar colors tend to belong to the same layer. The choice of a local non-parametric color representation allows us to handle surfaces with heterogeneous and varying color distributions, such as the ones in the Venus dataset, on which image segmentation may be hard. Finally, the implicit assumption that scene surfaces are frontoparallel is only made in the initial matching stage, when all pixels in a small window are assumed to have the same disparity. After this point, the surfaces are never assumed to be anything other than continuous. The ordering constraint that is violated by certain scene configurations is also not used.

As far as multiple view stereo is concerned, volumetric methods have gained increased popularity and have achieved very good results. Surveys of these methods were published in [11]. Volumetric methods adopt a ray-space representation of the scene and pose multiple view stereo as the computation of the visual hull of the scene. Even though they are currently the state of the art, they suffer from some inherent limitations. First and foremost among them is the inability to handle concavities in the scene, while the general placement of cameras further increases the computational cost [11]. Volumetric methods are by nature iterative, require careful initialization of the scene model, and cannot recover from erroneous initializations. Perceptual methods operate at a higher level than pixels and voxels exploiting relationships between elements of the same scene or image structure. A perceptual grouping framework enables the propagation of information among tokens and imposes global constraints, such as smoothness, therefore, it can utilize information from the images more efficiently. The limitations of current approaches based on perceptual organization are that either their representation is inadequate for generalization to multiple views, or that computational complexity increases exponentially with the number of views.

Overall, we have shown how a fundamental computer vision problem can be addressed within the tensor voting framework. The development and integration of modules that interface between the images and the framework, and enforce the problem-specific constraints is all that is necessary. No modifications need to be made to the voting procedure and very similar modules can be used for other vision problems, such as visual motion analysis. Very good results have been achieved in both cases with the main difference being the dimensionality of the space where organization takes place (3-D for stereo and 4-D for motion analysis).

7. Plan for the Next Year

During the next year we intend to investigate further the use of monocular, image-based cues in our stereo algorithm. The rather simple scheme for color consistency enforcement we proposed in [12] performs very well. However, there are more monocular cues one could consider, such as the shape of the occlusion contours. Our focus will be on ensuring that all the elements added to the approach are consistent with the general principles of the Tensor Voting Framework and that they can be generalized to more than two images.

We also intend to investigate the application of Tensor Voting at multiple scales, as many real scenes contain objects that are perceived at different scales. This is due to variations in size and level of detail as well as variations in texture density. A multiple scale processing scheme is necessary for processing such scenes. It must be capable of detecting and preserving fine details, not allowing larger or richer in texture regions to dominate smaller or less textures ones, achieving good continuation of structures, and filling in missing data. Instead of compromising the quality of the results by selecting a single scale of analysis for the entire scene, which would be sub-optimal in most regions, we intend to develop an approach that automatically infers the smallest scale that can preserve the finer details, then proceeds with larger scales to ensure good continuation in regions with sparse or missing data.

8. Expected Milestones and Deliverables

We intend to submit our methodology and results to conferences and journals for review and publication. Since our methodology and results have been published, we will submit them to the Stereo Vision Research Page of Middlebury College (http://www.middlebury.edu/stereo/), which is a repository of a large number (more than 25) of approaches to stereovision that includes a quantitative evaluation on the same standardized datasets.

Our industrial partner, Geometrix Inc., may incorporate parts of our research in their 3-D reconstruction products. Since our algorithm is now capable of producing high quality depth maps from a rather large class of input images, its usefulness to the industry has increased.

In terms of fundamental research, our future work will be along two major axes. First, on the augmentation of the tensor voting framework which will make it more efficient and capable of conveying more information. Second, on the proper incorporation of monocular cues to complement the binocular ones in order to overcome occlusion and lack of texture. We also aim at a Tensor Voting framework that will operate at multiple scales and will be able to handle extremely difficult cases of perceptual grouping some of which arise from complicated stereo

pairs. We anticipate that the latter goal will require the most time and effort, but will stand as a significant milestone of our research.

9. Member Company Benefits

This approach, once successfully implemented, represents a significant improvement over the existing state-of-the-art. It can be incorporated directly into one of the company's product lines, which involves the creation of 3-D models from images. The focus of our future research will be on addressing complex stereovision scenarios for which no satisfactory solutions exist in the literature, as for instance the reconstruction of very thin objects such as hair.

10. References

[1] M.S. Lee and G. Medioni, "Inferring Segmented Surface Description from Stereo Data", Proc. CVPR, pp. 346-352, 1998.

[2] M.S. Lee and G. Medioni, "Grouping ., -, ->, O-, into Regions, Curves, and Junctions", CVIU, vol. 76, no. 1, pp. 54-69, 1999.

[3] M.S. Lee, G. Medioni, and P. Mordohai, "Inference of Segmented Overlapping Surfaces from Binocular Stereo", IEEE Trans. on PAMI, vol. 24, no. 6, pp. 824-837, June 2002.

[4] G. Medioni, M.S. Lee, and C.K Tang, "A Computational Framework for Feature Extraction and Segmentation", Elsevier, 2000.

[5] P. Mordohai, G. Medioni, and M.S. Lee, "Inference of Segmented Overlapping Surfaces from Binocular and Multiple-View Stereo", 3rd Workshop on Perceptual Organization in Computer Vision, 2001.

[6] P. Mordohai and G. Medioni, "Perceptual Grouping for Multiple View Spereo using Tensor Voting", International Conference on Pattern Recognition, vol. 3, pp. 639-644, Quebec City, Canada, August 2002.

[7] G. Medioni, P. Mordohai, and M. Nicolescu, "The Tensor Voting Framework", to appear in the Handbook of Computational Geometry for Pattern Recognition, Computer Vision, Neurocomputing and Robotics, Eduardo Bayro-Corrochano (editor), Springer-Verlag, currently in press, 2004.

[8] G. Medioni and P. Mordohai, "The Tensor Voting Framework", to appear in Emerging Topics in Computer Vision, S.B. Kang and G. Medioni (editors), Prentice Hall, currently in press, 2004.

[9] G. Medioni and P. Mordohai, "Saliency in Computer Vision", to appear in Neurobiology of Attention, L. Itti, G. Rees, and J. Tsotsos (editors), Elsevier Science, 2004.

[10] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," IJCV, vol. 47, no. 1-3, pp. 7–42, April 2002.

[11] G. Slabaugh, B. Culbertson, T. Malzbender, and R. Schafer, "A Survey of Methods for Volumetric Scene Reconstruction from Photographs", Proc. International Workshop on Volume Graphics, 2001.

[12] P. Mordohai and G. Medioni, "Stereo using Monocular Cues within the Tensor Voting Framework", accepted to the European Conference on Computer Vision, 2004.

[13] W.S. Tong, C.K. Tang, P. Mordohai, and G. Medioni, "First Order Augmentations to Tensor Voting for Boundary Inference and Multiscale Analysis in 3-D", to appear in IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004.