An Experimental Comparison of Stereo Algorithms

Richard Szeliski[1] and Ramin Zabih[2]

[1] Microsoft Research, Redmond, WA 98052-6399, szeliski@microsoft.com
[2] Cornell University, Ithaca, NY 14853-7501, rdz@cs.cornell.edu

Abstract. While many algorithms for computing stereo correspondence have been proposed, there has been very little work on experimentally evaluating algorithm performance, especially using real (rather than synthetic) imagery. In this paper we propose an experimental comparison of several different stereo algorithms. We use real imagery, and explore two different methodologies, with different strengths and weaknesses. Our first methodology is based upon manual computation of dense ground truth. Here we make use of a two stereo pairs: one of these, from the University of Tsukuba, contains mostly fronto-parallel surfaces; while the other, which we built, is a simple scene with a slanted surface. Our second methodology uses the notion of prediction error, which is the ability of a disparity map to predict an (unseen) third image, taken from a known camera position with respect to the input pair. We present results for both correlation-style stereo algorithms and techniques based on global methods such as energy minimization. Our experiments suggest that the two methodologies give qualitatively consistent results. Source images and additional materials, such as the implementations of various algorithms, are available on the web from http://www.research.microsoft.com/~szeliski/stereo.

1 Introduction

The accurate computation of stereo depth is an important problem in early vision, and is vital for many visual tasks. A large number of algorithms have been proposed in the literature (see [8,10] for literature surveys). However, the state of the art in evaluating stereo methods is quite poor. Most papers do not provide quantitative comparisons of their methods with previous approaches. When such comparisons are done, they are almost inevitably restricted to synthetic imagery. (However, see [13] for a case where real imagery was used to compare a hierarchical area-based method with a hierarchical scanline matching algorithm.)

The goal of this paper is to rectify this situation, by providing a quantitative experimental methodology and comparison among a variety of different methods using real imagery. There are a number of reasons why such a comparison is valuable. Obviously, it allows us to measure progress in our field and motivates us to develop better algorithms. It allows us to carefully analyze algorithm characteristics and to improve overall performance by focusing on sub-components. It allows

B. Triggs, A. Zisserman, R. Szeliski (Eds.): Vision Algorithms'99, LNCS 1883, pp. 1–19, 2000.

us to ensure that algorithm performance is not unduly sensitive to the setting of "magic parameters". Furthermore, it enables us to design or tailor algorithms for specific applications, by tuning these algorithms to problem-dependent cost or fidelity metrics and to sample data sets.

We are particularly interested in using these experiments to obtain a deeper understanding of the behavior of various algorithms. To that end, we focus on image phenomena that are well-known to cause difficulty for stereo algorithms, such as depth discontinuities and low-texture regions.

Our work can be viewed as an attempt to do for stereo what Barron *et al.*'s comparative analysis of motion algorithms [3] accomplished for motion. The motion community benefited significantly from that paper; many subsequent papers have made use of these sequences. However, Barron *et al.* rely exclusively on synthetic data for their numerical comparisons; even the well-known "Yosemite" sequence is computer-generated. As a consequence, it is unclear how well their results apply to real imagery.

This paper is organized as follows. We begin by describing our two evaluation methodologies and the imagery we used. In section 3 we describe the stereo algorithms that we compare and give some implementation details. Section 4 gives experimental results from our investigations. We close with a discussion of some extensions that we are currently investigating.

2 Evaluation Methodologies

We are currently studying and comparing two different evaluation methodologies: comparison with ground truth depth maps, and the measurement of novel view prediction errors.

2.1 Data Sets

The primary data set that we used is a multi-image stereo set from the University of Tsukuba, where every pixel in the central *reference* image has been labeled by hand with its correct disparity. The image we use for stereo matching and the ground truth depth map are shown in figure 1. Note that the scene is fairly fronto-planar, and that the ground truth contains a small number of integer-valued disparities.

The most important limitation of the Tsukuba imagery is the lack of slanted surfaces. We therefore created a simple scene containing a slanted surface. The scene, together with the ground truth, are shown in figure 2. The objects in the scene are covered with paper that has fairly high-texture pictures on it. In addition, the scene geometry is quite simple. Additional details about this imagery can be found at the web site for this paper,
`http://www.research.microsoft.com/~szeliski/stereo`.

2.2 Comparison with Ground Truth

The ground truth images are smaller than the input images; we handle this by ignoring the borders (i.e., we only compute error statistics at pixels which are

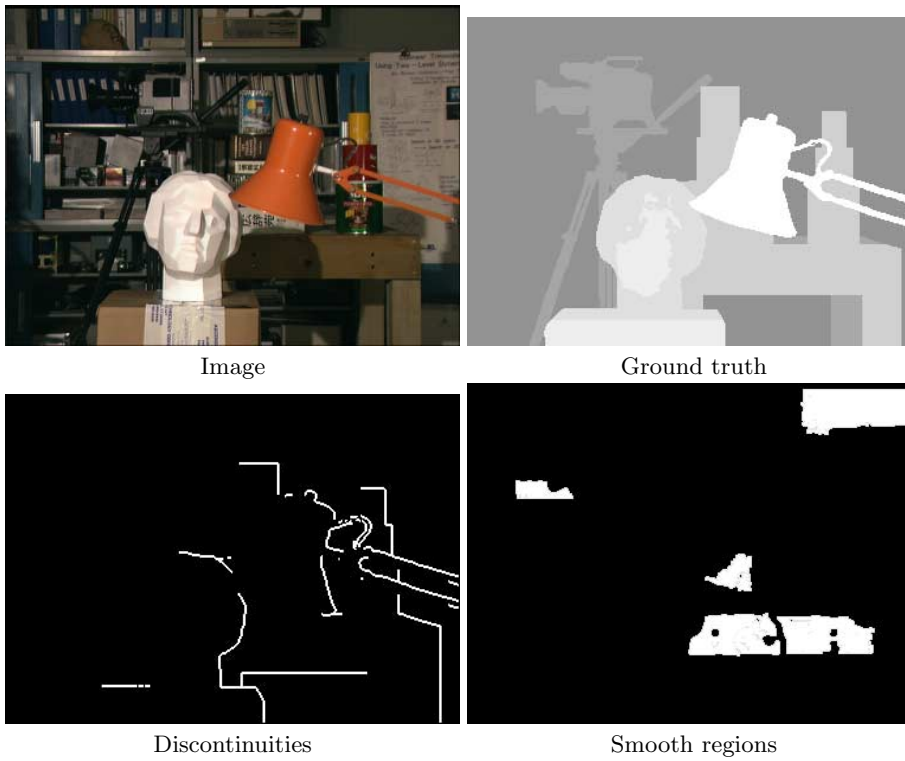


Fig. 1. Imagery from the University of Tsukuba

given a label in the ground truth). Discarding borders is particularly helpful for correlation-based methods, since their output near the border is not well-defined.

The interesting regions in the Tsukuba imagery include:

- Specular surfaces, including the gray shelves in the upper left of the image, the orange lamp, and the white statue of a face. Specularities cause difficulty in computing depth, due to the reflected motion of the light source.
- Textureless regions, including the wall at the top right corner and the deeply shadowed area beneath the table. Textureless regions are locally ambiguous, which is a challenge for stereo algorithms.
- Depth discontinuities, at the borders of all the objects. It is difficult to compute depth at discontinuities, for a variety of reasons. It is especially difficult for thin objects, such as the orange lamp handle.
- Occluded pixels, near some of the object borders. Ideally, a stereo algorithm should detect and report occlusions; in practice, many algorithms do not do this, and in fact tend to give incorrect answers at unoccluded pixels near the occlusions.

Our goal is to analyze the effectiveness of different methods in these different regions. We have used the ground truth to determine the depth discontinuities and the occluded pixels. A pixel is a depth discontinuity if any of its

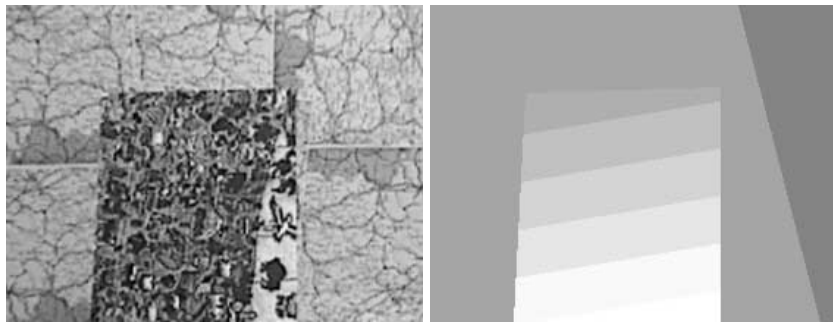


Image Ground truth

Fig. 2. Imagery from Microsoft Research.

(4-connected) neighbors has a disparity that differs by more than 1 from its disparity.[1] A pixel is occluded if according to the ground truth it is not visible in both images.

While the Tsukuba imagery contains textureless regions, there is no natural way to determine these from the ground truth. Instead, we looked for large regions where the image gradient was small. We found five such regions, which are shown in figure 1. These regions correspond to major features in the scene; for example, one is the lamp shade, while two come from shadowed regions under the table.

We have computed error statistics for the depth discontinuities and for the textureless regions separately from our statistics for the other pixels. Since the methods we wish to compare include algorithms that do not detect occlusions, we have ignored the occluded pixels for the purpose of our statistics. Our statistics count the number of pixels whose disparity differs from the ground truth by more than ± 1. This makes sense because the true disparities are usually fractional. Therefore, having an estimate which differs from the true value by a tiny amount could be counted as an error if we required exact matches. In fact, we also computed exact matches, and obtained quite similar overall results.

2.3 Comparison Using Prediction Error

An alternative approach to measuring the quality of a stereo algorithm is to test how well it predicts novel views.[2] This is a particularly appropriate test when the stereo results are to be used for image-based rendering, but it is also useful for other tasks such as motion-compensated prediction in video coding and frame rate conversion. This approach parallels methodologies that are prevalent

[1] Neighboring pixels that are part of a sloped surface can easily differ by 1 pixel, but should not be counted as discontinuities.

[2] A third possibility is to run the stereo algorithm on several different pairs within a multi-image data set, and to compute the *self-consistency* between different disparity estimates [16]. We will discuss this option in more detail in section 5.

in many other research areas, such as speech recognition, machine learning, and information retrieval. In these disciplines, it is traditional to partition data into *training* data that is used to tune up the system, and *test* data that is used to evaluate it. In statistics, it is common to leave some data out during model fitting to prevent over-fitting (*cross-validation*).

To apply this methodology to stereo matching, we compute the depth map using a pair of images selected from a larger, multi-image stereo dataset, and then measure how well the original *reference* image plus depth map predicts the remaining views. The University of Tsukuba data set is an example of such a multi-image data set: the two images in figure 1 are part of a 5×5 grid of views of this scene. Many other examples of multi-image data sets exist, including the *Yosemite* fly-by, the *SRI Trees* set, the *NASA (Coke can)* set, the MPEG-4 *flower garden* data set, and many of the data sets in the CMU Image Database (`http://www.ius.cs.cmu.edu/idb`). Most of these data sets do not have an associated ground truth depth map, and yet all of them can be used to evaluate stereo algorithms if prediction error is used as a metric.

When developing a prediction error metric, we must specify two different components: an algorithm for predicting novel views, and a metric for determining how well the actual and predicted images match. A more detailed description of these issues can be found in our framework paper, which lays the foundations for prediction error as a quality metric [20].

In terms of view prediction, we have a choice of two basic algorithms. We can generate novel views from a color/depth image using *forward warping* [22], which involves moving the source pixels to the destination image and potentially filling in gaps. Alternatively, we can use *inverse warping* to pull pixels from the new (unseen) views back into the coordinate frame of the original reference image. This is easier to implement, since no decision has to be made as to which gaps need to be filled.

Unfortunately, inverse warping will produce erroneous results at pixels which are occluded (invisible) in the novel views, unless a separate occlusion (or visibility) map is computed for each novel view. Without this occlusion masking, certain stereo algorithms actually outperform the ground truth in terms of prediction error, since they try to match occluded pixels to *some* other pixel of the same color. In our experiments, therefore, we do not include occluded pixels in the computation of prediction error.

The simplest error metric that can be used in an L_2 (root mean square) distance between the pixel values. It is also possible to use a robust measure, which downweights large error, and to count the number of outliers [17]. Another possibility is to compute the per-pixel *residual motion* between the predicted and real image, and to compensate one of the two images by this motion to obtain a *compensated* RMS or robust error measure [20]. For simplicity, we use the raw (uncompensated and un-robustified) RMS error.

3 Algorithms

Loosely speaking, methods for computing dense stereo depth can be divided into two classes.[3] The first class of methods allow every pixel to independently select its disparity, typically by analyzing the intensities in a fixed rectangular window. These methods use statistical methods to compare the two windows, and are usually based on correlation. The second class of methods relies on global methods, and typically find the depth map that minimizes some function, called the energy or the objective function. These methods generally use an iterative optimization technique, such as simulated annealing.

3.1 Local Methods Based on Correlation

We implemented a number of standard correlation-based methods that use fixed-size square windows. We define the *radius* of a square whose side length is $2r+1$ to be r. The methods we chose were:

- Correlation using the L_2 and L_1 distance. The L_2 distance is the simplest correlation-based method, while the L_1 distance is more robust.
- Robust correlation using M-estimation with a truncated quadratic [5].
- Robust correlation using Least Median Squares [17].

3.2 Global Methods

Most global methods are based on energy minimization, so the major variable is the choice of energy function. Some stereo methods minimize a 1-dimensional energy function independently along each scanline [1,4,15]. This minimization can be done efficiently via dynamic programming [14]. More recent work has enforced some consistency between adjacent scanlines. We have found that one of these methods, MLMHV [9], performs quite well in practice, so we have included it in our study.

The most natural energy functions for stereo are two-dimensional, and contain a data term and a smoothness term. The data term is typically of the form $\sum_p [I(p) - I'(p + d(p))]^2$, where d is the depth map, p ranges over pixels, and I and I' are the input images. For our initial experiments, we have chosen a simple smoothness term which behaves well for largely front-planar imagery (such as that shown in figure 1). This energy function is the *Potts energy*, and is simply the number of adjacent pixels with different disparities.

In the energy minimization framework, it is difficult to determine whether an algorithm fails due to the choice of energy function or due to the optimization method. This is especially true because minimizing the energy functions that arise in early vision is almost inevitably NP-hard [21]. By selecting a single energy function for our initial experiments, we can control for this variable.

[3] There are a number of stereo methods that compute a sparse depth map. We do not consider these methods for two reasons. First, a dense output is required for a number of interesting applications, such as view synthesis. Second, a sparse depth map makes it difficult to identify statistically significant differences between algorithms.

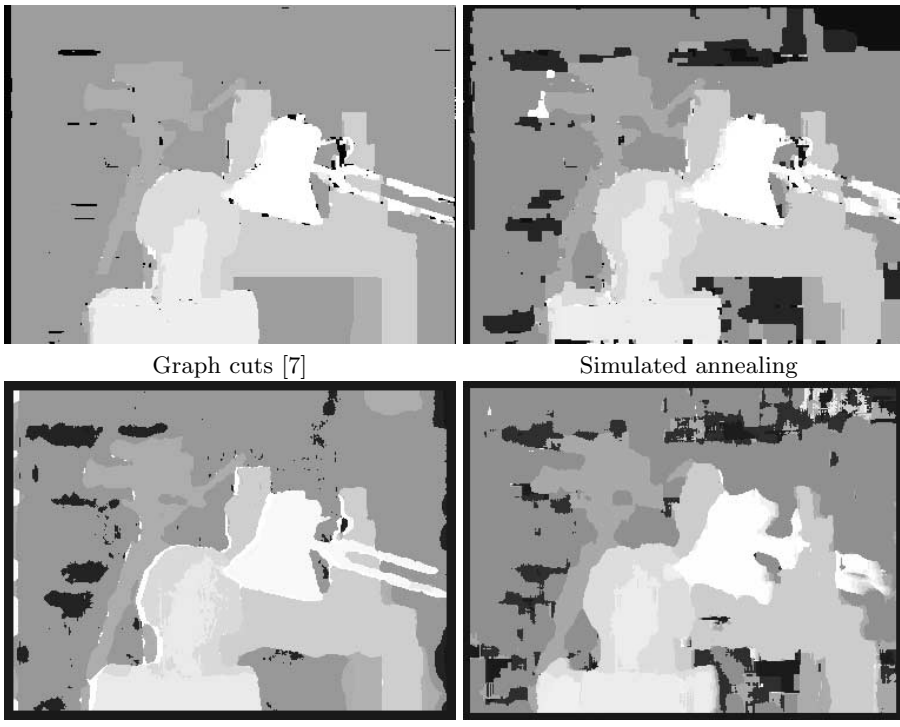


Fig. 3. Results for several algorithms, on the imagery of figure 1

We used three methods to minimize the energy:

- Simulated annealing [12,2] is the most common energy minimization technique. Following [6], we experimented with several different annealing schedules; our data is from the one that performed best.
- Graph cuts are a combinatorial optimization technique that can be used to minimize a number of different energy functions [7]. Other algorithms based on graph cuts are given in [18,15].
- Mean field methods replace the stochastic update rules of simulated annealing with deterministic rules based either on the behavior of the *mean* or *mode* disparity at each pixel [11], or the local *distribution* of probabilities across disparity [19]. We present results for the latter algorithm.

The final method we experimented with is a global method that is not based on energy minimization. This algorithm, due to Zitnick and Kanade [23], is a cooperative method in the style of the Marr-Poggio algorithm. A particularly interesting feature of the algorithm is that it enforces various physical constraints, such as uniqueness. The uniqueness constraint states that a non-occluded pixel in one image should map to a unique pixel in the other image.

# errors ($> \pm 1$)	Total pixels	Discontinuities	Low-texture areas
Image	84,863	1,926	6,037
L_1 distance	10,457	1,025	967
Annealing	4,244	720	765
Zitnick/Kanade [23]	2,191	749	60
Graph cuts [7]	1,591	572	0

Fig. 4. Errors on ground truth data, from the results shown in figure 3.

4 Experimental Results

We have run all the mentioned algorithms on the Tsukuba imagery, and used both ground truth and prediction error to analyze the results. In addition, we have run the correlation-based algorithms on the Microsoft Research imagery.

4.1 Results on the Tsukuba Imagery

Figure 3 shows the depth maps computed by three different algorithms. Figures 5–7 show the performance of various algorithms using the ground truth performance methodology. Figure 5 shows the performance of correlation-based methods as a function of window size. Figure 6 shows the performance of two global optimization methods.

Figure 7 summarizes the overall performance of the best versions of different methods. The graph cuts algorithm has the best performance, and makes no errors in the low-textured areas shown in figure 1. Simulated annealing, mean-field estimation, M-estimation, and MLMHV [9] seem to have comparable performance. Note that the differences in overall performance between methods cannot be explained simply by their performance at discontinuities or in low-textured areas. For example, consider the data for annealing and for graph cuts, shown in figure 4. There is a substantial difference in performance at discontinuities and in textureless regions, but most errors occur in other portions of the image.

4.2 Analysis of Ground-Truth Data

Our data contains some unexpected results. First of all, it is interesting that the different correlation-based methods are so similar in terms of their performance. In addition, there was surprisingly little variation as a function of window size, once the windows were sufficiently large. Finally, the overall performance of the correlation-based methods was disappointing, especially near discontinuities. Note that an algorithm that assigned every pixel a random disparity would be within ± 1 of the approximately 20% of the time, and thus correct under our definition.

It is commonly believed that it is important for matching algorithms to gracefully handle outliers. In terms of statistical robustness, the L_2 distance is the worst, followed by the L_1 distance. M-estimation is better still, and least median squares is best of all. There is some support for this argument in our

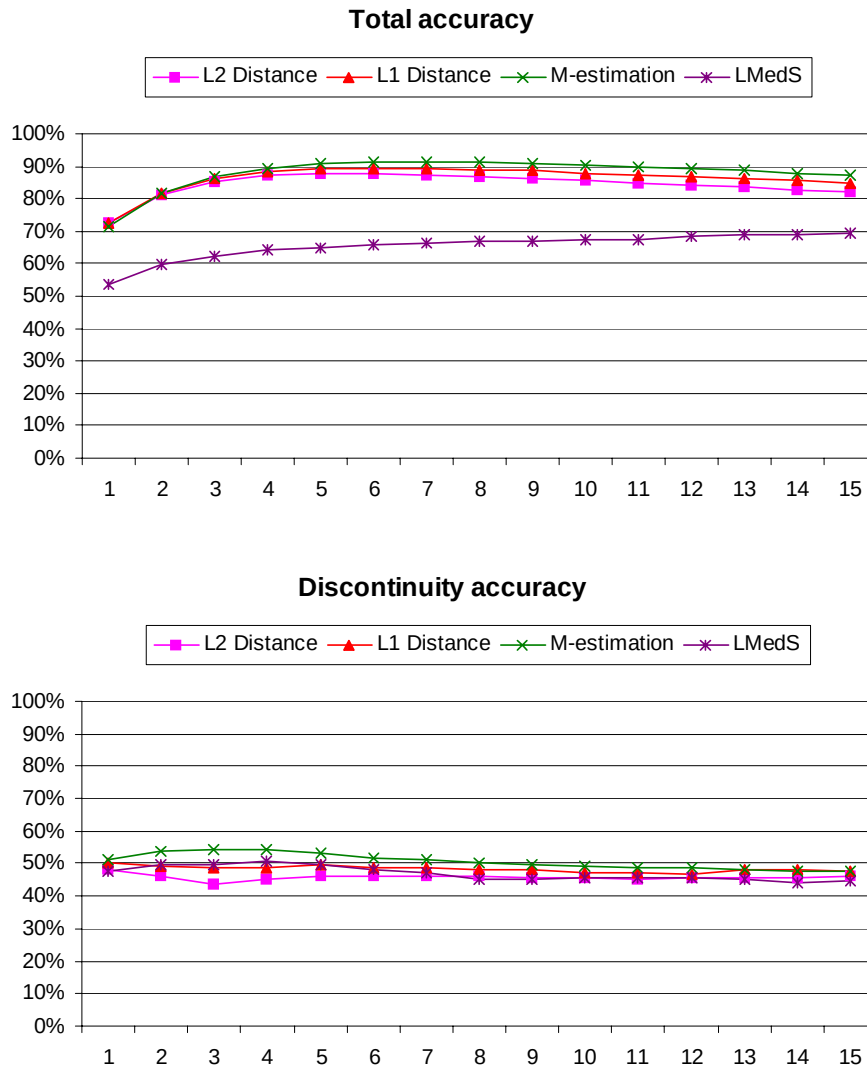


Fig. 5. Performance of standard correlation-based methods as a function of window radius, using the ground truth of figure 1. The graph at top shows errors for all pixels (discarding those that are occlusions according to the ground truth). The graph at bottom only considers pixels that are discontinuities according to the ground truth.

data, but it is not clear cut. The L_1 distance has a small advantage over the L_2 distance, and M-estimation has a slightly larger advantage over the L_1 distance. Least median squares does quite badly (although to the naked eye it looks fairly good, especially with small windows). The regions where it makes the most

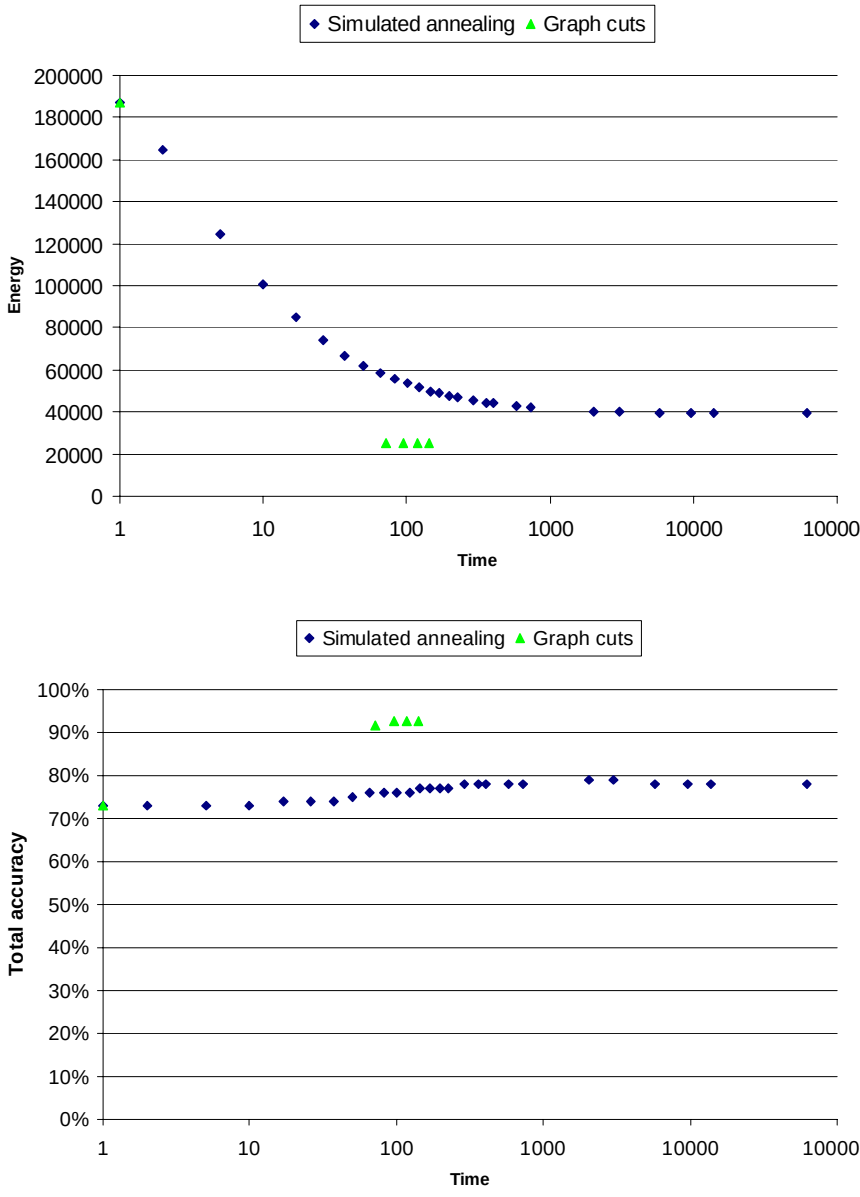


Fig. 6. Performance of global optimization methods as a function of running time, using the Potts model energy on the imagery of figure 1.

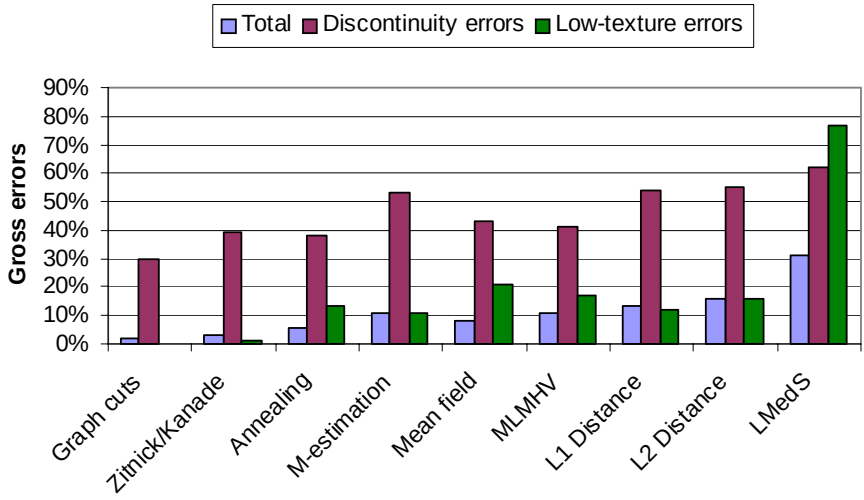


Fig. 7. Performance comparison, using the best results for each algorithm, on the imagery of figure 1.

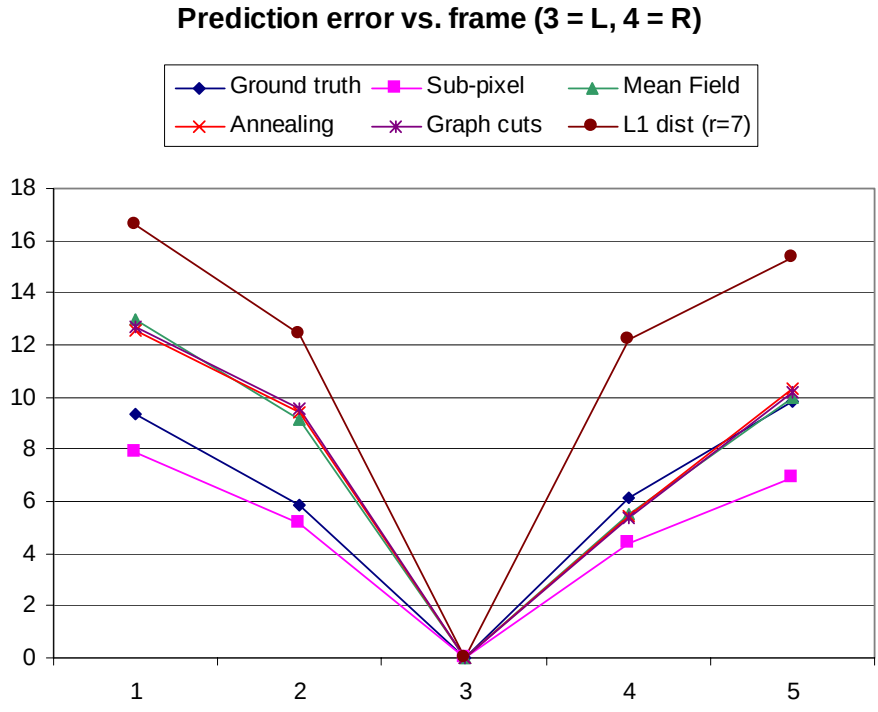


Fig. 8. RMS prediction error (in gray levels) as a function of frame number using the imagery of figure 1.

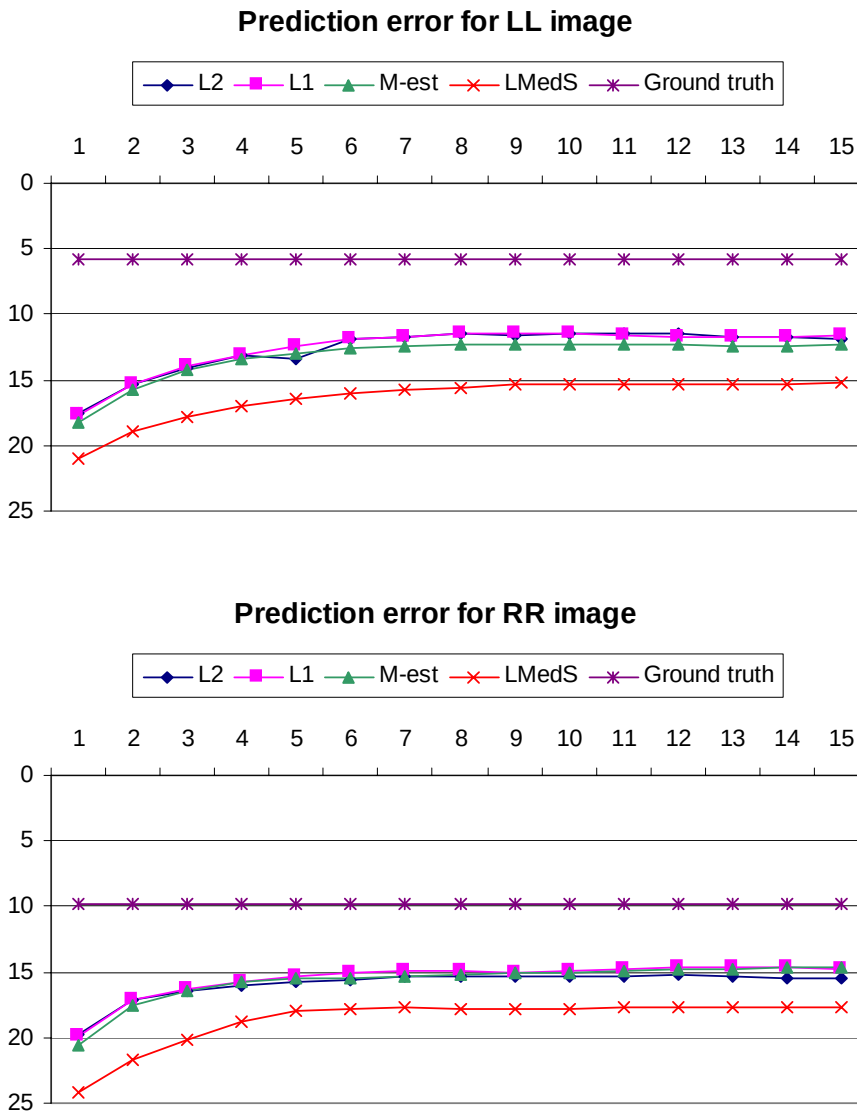


Fig. 9. Performance of correlation methods as a function of window radius, using prediction error.

mistakes are the low-texture regions, where it appears that the least median squares algorithm treats useful textured pixels (e.g., the bolts on the front of the workbench) as outliers.

4.3 Analysis of Prediction Error Data

The prediction error metrics for various algorithms are shown in figures 8 and 9. Figure 8 shows the RMS (uncorrected) prediction error as a function of frame number for four different algorithms and two versions of the ground truth data. The frames being tested are the ones in the middle row of the 5×5 University of Tsukuba data set (we will call these images LLL, LL, L, R, and RR in subsequent discussions). The ground truth data was modified to give sub-pixel estimates using a sub-pixel accurate correlation-based technique whose search range was constrained to stay within a $\frac{1}{2}$ pixel disparity of the original ground truth. As we can see in figure 8, this reduces the RMS prediction error by about 2 gray levels.

We can also see from this figure that error increases monotonically away from the reference frame 3 (the left image), and that prediction errors are worse when moving leftward. This is because errors in the depth map due to occlusions in the right image are more visible when moving leftward (these areas are being exposed, rather than occluded). It is also interesting that the graph cut, mean field, and annealing approaches have very similar prediction errors, even though their ground truth errors differ. Our current conjecture is that this is because graph cuts do a better job of estimating disparity in textureless regions, which is not as important for the prediction task.

Figure 9 shows the prediction error as a function of window size for the four correlation-based algorithms we studied. These figures also suggest that a window size of 7 is sufficient if prediction error is being used as a metric. The shape of these curves is very similar to the ground truth error (figure 5), which suggests that the two metrics are producing consistent results.

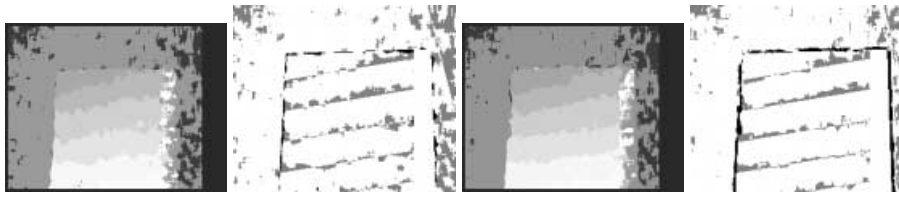
4.4 Results on the Microsoft Research Imagery

The results from running different variants of correlation on the imagery of figure 2 are shown in figure 11. Selected output images are given in figure 10. The overall curves are quite consistent with the results from the Tsukuba imagery. Here, the least median squares algorithm does slightly bettern than the other techniques. This is probably because there are no low-texture regions in this dataset.

5 Discussion

In this paper, we have compared two methodologies for evaluating stereo matching algorithms, and also compared the performance of several widely used stereo algorithms. The two methodologies produce different, but somewhat consistent results, while emphasizing (or de-emphasizing) certain kinds of errors.

The ground truth methodology gives the best possible evaluation of a stereo matcher's quality, since it supposedly knows what the perfect result ("gold standard") should be. However, it is possible for the ground truth to be inaccurate, and it typically is so near discontinuities where pixels are mixtures of values



M-estimation output and errors L_2 output and errors

Fig. 10. Results for correlation algorithms ($r = 4$), on the imagery of figure 2. Errors $> \pm 1$ are shown in black, while errors of ± 1 are shown in gray.

from different surfaces. Quantization to the nearest integer disparity is a further source of error, which we compensate for by only counting errors $> \pm 1$ disparity. Ground truth also weights regions such as textureless or occluded regions (where it is very difficult, if not impossible, to get a reliable result) equally with regions where all algorithms should perform well. In our experiments, we have deliberately excluded occluded regions from our analysis. It may be desirable to treat these regions on the same footing as other potential problem areas (textureless regions and discontinuities). Breaking down the error statistics by their location in the image, is a step towards trying to rationalize this situation.

Intensity prediction error is a different metric, which de-emphasizes errors in low-texture areas, but emphasizes small (one pixel or sub-pixel) errors in highly textured areas. The former is a reasonable idea if the stereo maps are going to be used in an image-based rendering application. Those regions where the depth estimates are unreliable due to low texture are also regions where the errors are less visible. The problem with sub-pixel errors should be fixable by modifying or extending the algorithms being evaluated to return sub-pixel accurate disparity estimates.

A third methodology, which we have not yet evaluated, is the self-consistency metric of Leclerc *et al.* [16]. In this methodology, the consistency in 3D location (or re-projected pixel coordinates) of reconstructed 3D points from different pairs of images is calculated. This shares some characteristics with the intensity prediction metric error used in this paper, in that more than two images are used to perform the evaluation. However, this metric is more stringent than intensity prediction. In low texture areas where the results tend to be error-prone, it is unlikely that the self-consistency will be good (whereas intensity prediction may be good). There is a possibility that independently run stereo matchers may accidentally produce consistent results, but this seems unlikely in practice. In the future, we hope to collaborate with the authors of [16] to apply our different methodologies to the same sets of data.

5.1 Extensions

In work to date, we have already obtained interesting results about the relative performance of various algorithms and their sensitivity to certain parameters

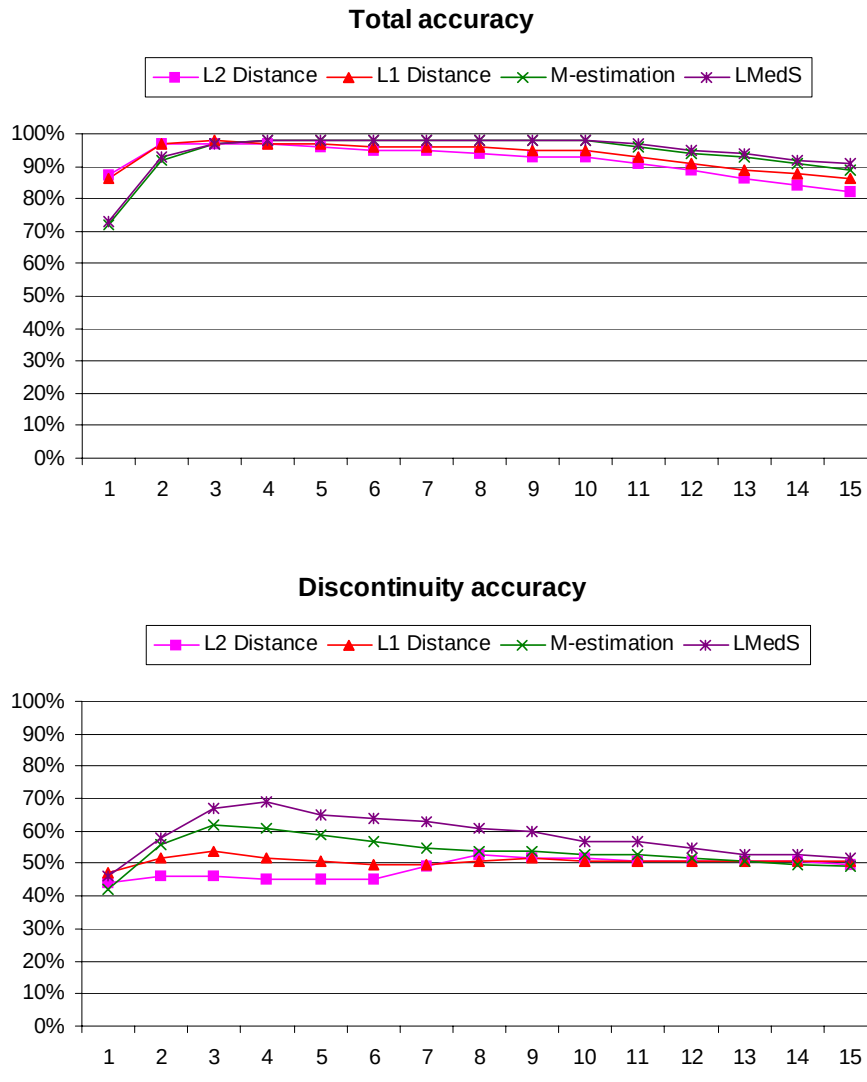


Fig. 11. Performance of standard correlation-based methods as a function of window radius, using the ground truth of figure 2. The graph at top shows errors for all pixels (discarding those that are occlusions according to the ground truth). The graph at bottom only considers pixels that are discontinuities according to the ground truth.

(such as window size). However, there are many additional issues and questions that we are planning to examine in ongoing work. These issues include:

- Sub-pixel issues and sampling error: investigate the effects of using a finer set of (sub-pixel) disparities on both the ground truth and prediction error metrics.
- Study more algorithms, including minimization with non-Potts energy and the use of weighted windows for correlation.
- Evaluate more data sets.
- Determine whether it is more important to come up with the correct energy to minimize, or whether it is more important to find a good minimum.
- Investigate the sensitivity of algorithms to various parameter values.
- Study whether cross-validation (using prediction error in a multi-image stereo dataset) can be used to fine-tune algorithm parameters or to adapt them locally across an image.

We hope that our results on stereo matching will motivate others to perform careful quantitative evaluation of their stereo algorithm, and that our inquiries will lead to a deeper understanding of the behavior (and failure modes) of stereo correspondence algorithms.

Acknowledgements

We are grateful to Y. Ohta and Y. Nakamura for supplying the ground truth imagery from the University of Tsukuba, to various colleagues for furnishing us with their algorithms and/or results, and for the helpful suggestions from the reviewers and program committee. The second author has been supported by the National Science Foundation under contracts IIS-9900115 and CDA-9703470.

References

1. H.H. Baker and T.O. Binford. Depth from edge and intensity based stereo. In *IJCAI81*, pages 631–636, 1981.
2. Stephen Barnard. Stochastic stereo matching over scale. *International Journal of Computer Vision*, 3(1):17–32, 1989.
3. J.L. Barron, D.J. Fleet, and S.S. Beauchemin. Performance of optical flow techniques. *International Journal of Computer Vision*, 12(1):43–77, February 1994.
4. P. N. Belhumeur and D. Mumford. A Bayesian treatment of the stereo correspondence problem using half-occluded regions. In *Computer Vision and Pattern Recognition*, pages 506–512, Champaign-Urbana, Illinois, 1992.
5. Michael Black and P. Anandan. A framework for the robust estimation of optical flow. In *4th International Conference on Computer Vision*, pages 231–236, 1993.
6. Andrew Blake. Comparison of the efficiency of deterministic and stochastic algorithms for visual reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(1):2–12, January 1989.
7. Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast approximate energy minimization via graph cuts. In *Seventh International Conference on Computer Vision (ICCV'99)*, pages 377–384, Kerkyra, Greece, September 1999.

8. Lisa Brown. A survey of image registration techniques. *ACM Computing Surveys*, 24(4):325–376, December 1992.
9. I. Cox, S. Hingorani, S. Rao, and B. Maggs. A maximum likelihood stereo algorithm. *Computer Vision, Graphics and Image Processing*, 63(3):542–567, 1996.
10. U. Dhond and J. Aggarwal. Structure from stereo — a review. *IEEE Transactions on Systems, Man and Cybernetics*, 19(6), 1989.
11. Davi Geiger and Federico Girosi. Parallel and deterministic algorithms from MRF's: Surface reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(5):401–412, May 1991.
12. S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.
13. Y. C. Hsieh, D. McKeown, and F. P. Perlant. Performance evaluation of scene registration and stereo matching for cartographic feature extraction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):214–238, February 1992.
14. S. S. Intille and A. F. Bobick. Disparity-space images and large occlusion stereo. In *Proc. Third European Conference on Computer Vision (ECCV'94)*, volume 1, Stockholm, Sweden, May 1994. Springer-Verlag.
15. H. Ishikawa and D. Geiger. Occlusions, discontinuities, and epipolar lines in stereo. In *Fifth European Conference on Computer Vision (ECCV'98)*, pages 332–248, Freiburg, Germany, June 1998. Springer-Verlag.
16. Y. G. Leclerc, Q.-T. Luong, and P. Fua. Self-consistency: A novel approach to characterizing the accuracy and reliability of point correspondence algorithms. In *DARPA Image Understanding Workshop*, Monterey, California, November 1998.
17. Peter Rousseeuw and Annick Leroy. *Robust Regression and Outlier Detection*. New York: Wiley, 1987.
18. S. Roy and I. J. Cox. A maximum-flow formulation of the n-camera stereo correspondence problem. In *Sixth International Conference on Computer Vision (ICCV'98)*, pages 492–499, Bombay, January 1998.
19. D. Scharstein and R. Szeliski. Stereo matching with nonlinear diffusion. *International Journal of Computer Vision*, 28(2):155–174, July 1998.
20. R. Szeliski. Prediction error as a quality metric for motion and stereo. In *Seventh International Conference on Computer Vision (ICCV'99)*, pages 781–788, Kerkyra, Greece, September 1999.
21. Olga Veksler. *Efficient Graph-based Energy Minimization Methods in Computer Vision*. PhD thesis, Cornell University, July 1999.
22. G. Wolberg and T. Pavlidis. Restoration of binary images using stochastic relaxation with annealing. *Pattern Recognition Letters*, 3:375–388, 1985.
23. Charles Zitnick and Takeo Kanade. A cooperative algorithm for stereo matching and occlusion detection. Technical Report CMU-RI-TR-99-35, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, October 1999.

Discussion

Jean Ponce: You only showed one data set, how much can you say from one data set?

Ramin Zabih: Yes, that's clearly the major weakness so far — there's only one data set. The two different methodologies point in the same direction. It might still be that there's something strange about the way the ground truth was presented, but I think that the agreement is encouraging. So far the state of the art has been just to show off different pictures. We're trying to move beyond that, but getting ground truth for the ground-truth methodology is very difficult. At the least the prediction error stuff allows you to work on lots of different data sets, which we're in the process of doing.

Jean Ponce: Yes, but prediction error is so application dependent. It works very well for image-based rendering, but if you want to do navigation or whatever, then it's not really appropriate.

Ramin Zabih: Yes, I agree.

Yvan Leclerc: Yesterday I talked about our self-consistency methodology for comparing stereo algorithms. I was wondering if you could comment on the relationship between your approach and ours.

Rick Szeliski: For those who missed Yvan's talk, his technique is similar in that you start with a multi-image data set. But instead of what we call prediction error, where you take a depth map computed with one image pair and predict the appearance in all the other images, Yvan's methodology computes depth maps between all possible pairs, and then sees whether they're consistently predicting the same 3D point. I think that it's a very valid methodology. What we hope to do is to test a wider range of data sets with more algorithms, and eventually publish a survey paper, along the lines of the kind of comparative work that we see already in motion estimation. I think it will be essential to include Yvan's methodology as well. Hopefully we'll be able to work out some sort of a joint evaluation. The two metrics won't necessarily give the same results — appearance prediction is oriented towards image-based rendering and is tolerant of errors in low texture regions, whereas Yvan's method is oriented towards structure and might heavily penalize those. Jean's comment is very well taken — this is application-dependent. But you know, in computer vision we've worked on robotics, robotics, robotics. Even when we stopped working on that we still kept the same mind set. But if you look for example at what happens with stereo algorithms when you try to do z-keying — you try to extract the foreground person from a textured background and put something synthetic behind him — the result is horrible, it's just not acceptable, you get these spiky halos full of the wrong pixels. As Luc Robert commented yesterday, we can't use computer vision yet in Hollywood. The reason is basically that we're not focusing on the right problems. That's why I like prediction error — it penalizes you heavily for those visible little single pixels errors.

Yvan Leclerc: Combining our methodologies would be a great idea. Let's do it.

Ramin Zabih: One comment about Jean's point is that in many situations there do seem to be consistent differences between the algorithms. We don't have enough ground truth to do convincing statistics yet, but it looks like the optimization-based approaches are doing better, certainly at discontinuities, and often in low texture areas as well.

Rick Szeliski: One final comment. We have made these data sets available on `http://www.research.microsoft.com/~szeliski/stereo/`, so that people can run their stereo algorithms on them. We are interested in hearing about the results, as we intend to publish a comparative survey of the performance of the different methods we have access to.