Visual Similarity, Judgmental Certainty and Stereo Correspondence *

James Ze Wang† Martin A. Fischler‡
Artificial Intelligence Center, SRI International, Menlo Park, CA 94025

Abstract

Normal human vision is nearly infallible in modeling the visually sensed physical environment in which it evolved. In contrast, most currently available computer vision systems fall far short of human performance in this task, and further, they are generally not capable of being able to assert the correctness of their judgments. In computerized stereo matching systems, correctness of the similarity/identity-matching is almost never *guaranteed*. In this paper, we explore the question of the extent to which judgments of similarity/identity can be made essentially error-free in support of obtaining a relatively dense depth model of a natural outdoor scene. We argue for the necessity of simultaneously producing a crude scene-specific semantic "overlay". For our experiments, we designed a wavelet-based stereo matching algorithm and use "classification-trees" to create a primitive semantic overlay of the scene. A series of mutually independent filters has been designed and implemented based on the study of different error sources. Photometric appearance, camera imaging geometry and scene constraints are utilized in these filters. When tested on different sets of stereo images, our system has demonstrated above 97% correctness on *asserted* matches. Finally, we provide a principled basis for relatively dense depth recovery.

1 Introduction

Vision, by animals or machines, is an inductive process which results in the construction of models, or theories, about the sensed environment. Unlike mathematical assertions, with respect to which one can make absolute judgments about correctness (actually, only about consistency with some assumed set of axioms), any assertion about the physical world can only be disconfirmed – never established with certainty. Never the less, our introspection and experience assures us that normal human vision is almost infallible in modeling the visually sensed physical environment in which we evolved and with which we directly interact. It is almost never the case that there is a *hole* in our visual field where our visual system can't produce an instantiated model, and it is very rare that our visually produced models cause us to fail in some task because they were *incorrect*. Even in the case of illusions, it is not obvious that our visually guided behavior would suffer from the same errors our conscious introspection is subject to. (Obviously, geometric modeling becomes less reliable as the distance from the sensor increases.)

In contrast, most currently available computer vision systems fall far short of human performance in this task, and additionally, they make no attempt, or are generally not capable of being able to assert the correctness of their judgments in proposing correspondences required for dense stereo depth modeling. In computerized stereo matching systems correctness of the similarity/identity matching is almost never *guaranteed*. There are some important exceptions, especially in regard to "structure-from-motion" problems where efforts are made to either statistically predict and verify the accuracy of the 3-D registration methods [Weng et al., 1989, Fua and Leclerc, 1994, Torr, 1995, Zhang et al., 1995, McReynolds and Lowe, 1996, Pennec and Thirion, 1997, Torr and Murray, 1997, Zhang et al., 1998] or to select correspondences from a predetermined set that are consistent with a "rigid" spatial configuration.

In this paper, we explore the question of the extent to which judgments of similarity/identity (believed to be the bias of human stereopsis) can be made *essentially* error-free in the context of stereo matching in

*This work was sponsored by the Defense Advanced Research Projects Agency under contract DACA76-92-C-0008 monitored by the U.S. Army Topographic Engineering Center, Alexandria, VA. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Advanced Research Projects Agency, the United States Government, or SRI International.

†Also of Department of Computer Science and Department of Medical Informatics, Stanford University, Stanford, CA 94305. Email: wangz@cs.stanford.edu

‡Email: fischler@ai.sri.com

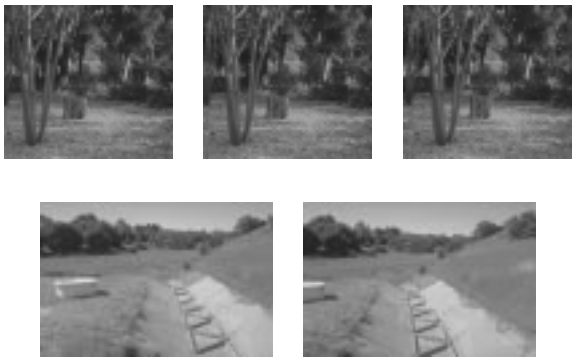


Figure 1: **Natural outdoor scenes used for our experimental investigation.**

the natural outdoor world. And further, how such a (possibly sparse) set of correspondences could provide a dense depth model.

2 The Central Problems

Human intelligence would be relatively worthless in a non-causal world. To exploit causality, it is necessary to be able categorize and recognize objects and events, in order to predict what will happen next or to take appropriate action based on past experience.

In machine vision, the categorization problem is central and pervasive. In this paper we examine one of the simplest instances of this problem – the problem of establishing stereo correspondence – and address the key question: *How can one be "certain" that a stereo match is correct.*

In order to answer this question, and exploit the answer, we address the following issues: what is visual similarity/uniqueness and how can we measure it; what is judgmental certainty and how can it be established; what is the role of semantic scene understanding in judgments about stereo correspondence.

2.1 Visual Similarity

A similarity metric for assigning distinct objects membership in a classification scheme can be completely arbitrary and is almost certain to be problem dependent. For example, we would not expect the metric used for classifying/recognizing flaws in a printed circuit board to be the preferred metric for correctly classifying images of trees according to species. Even when we restrict similarity judgments to the *identity* classes of real 3-D world objects (the distinct objects themselves, as opposed to class membership(s) of these distinct objects), there is a large set of alternative metrics that depend on how we define (or can acquire) our available observations, what we mean by an *object*, and how we intend to use the answer. For example, if we recognize the front and rear views of the same person in two different images, this could be useful for some purposes but relatively worthless for geometric recovery via stereo correspondence. Thus, any meaningful discussion of matching and the corresponding quantification of "degree of similarity" must be *grounded* in a specific problem. We use stereo vision as the grounded reference for evaluating our contribution. In this regard, we wish to understand and duplicate human stereo competence, but not necessarily the explicit mechanisms employed by the HVS.

We note that the human visual system operates in real-time, *below* the conscious level, to produce a 3-D representation of the environment. It is reasonable to assume that stereopsis is pre-attentive. This would normally imply that it uses little or no scene-specific contextual knowledge in arriving at its instantaneous judgments, but follows a preselected procedure (or algorithm). We will argue that effective stereo in the outdoor world must involve scene-specific context. Thus, a *solution* to the problem of designing an *infallible* stereo machine cannot be based solely on comparing the intensity variations in two (or more) images.

2.2 Judgmental Certainty

The problem that the HVS appears to have solved, the ability to make uniformly correct judgments in an uncertain world, is a core problem we address in this paper. There are, essentially, only two ways of judging when a fallible process has produced a correct answer:

1. Apply some known criterion/condition or test for correctness (that may not be competent in itself to obtain the desired answer). In mathematics we might not know how to prove a given theorem, but we know how to check a proof when offered one (regardless of the reliability of the source of the proof).

2. Get the *opinions* of a suitably sized collection of *informed independent* sources, and accept the proposed solution only when there is both a sufficient *consensus* of agreement, *and* when additional criterion for a valid model are satisfied: the additional criterion include stability (e.g., the derived model does not change in a significant way under *small* perturbations of the data

or the viewing conditions) and limited model *complexity* (given too many free variables in a model, it can be made consistent with any collection of data).

In this paper we focus on method (2) for establishing judgmental certainty. The application of this approach to problems in vision requires a careful examination of what is meant by the terms *informed* and *independent* in the vision context.

In its most fundamental sense, by **independent** opinions we mean that the errors made by the sources of these opinions, with regard to some given problem, are uncorrelated.

By **informed**, we mean (at least) that a process is more likely than pure chance to produce a correct answer. We will show later that an opinion can only be informed relative to some specific collection of error types/conditions. In particular, we must ultimately be concerned with:

- incorrect assumptions
- an incomplete model (e.g., some key variables are omitted – such as lens distortion in the context of a perspective imaging model)
- incomplete set of observations/information
- incorrect observations/information
- approximations (e.g., due to the finite resolution of measuring devices, and also, the representation of continuous numerical quantities in a machine)
- incorrect implementation (e.g., nerve damage, programming errors)
- probabilistic algorithms or a guessing strategy (errors are expected)
- an inappropriate utility function

Some of the corresponding visual phenomena include: (1) occlusion (2) ambiguity (3) distortion (4) incorrect assumptions about (or approximations with respect to) reflectance, surface continuity, camera geometry, illumination (5) computational errors or numerical instability in computing optical or geometric transforms.

2.3 Three Primary Information Sources for Image-Based Scene Modeling

We consider three primary information sources for image-based geometric scene modeling: (1) the image(s): photometric appearance and shape (2) the camera(s): imaging geometry constraints (3) the scene: scene domain and scene-specific constraints such as physical, semantic, geometric, photometric relationships and regularities.

2.3.1 Photometric Appearance-Based Similarity

From a statistical/signal-processing point of view, the objects of interest can be characterized using an attribute-vector of measurements made on the objects, and we then quantify the similarity relationship between two objects by the "normalized" distance between their attribute vectors. We note that even correlation-based matching can be viewed in this way – here the attribute vector is the ordered set of intensity values in the *correlation* patch. Never the less, it is difficult to deal with certain types of similarity problems using this formalism. In particular, line drawings cannot be well described this way, and more important, the local image appearance of (say) grass or other types of *nearby* vegetation is highly unstable to small shifts in viewing position. While we question the adequacy of vector-space characterization as the sole basis for natural outdoor scene stereo matching/modeling there are very few other practical alternatives available at present.

2.3.2 Imaging-Geometry Based Constraints on Feature Matching

Advances made over the past two decades in projective geometry and robust statistical estimation [Mikhail, 1976, Fischler and Bolles, 1981, Rousseeuw, 1987, Barrett et al., 1991, Mundy and Zisserman, 1992, Luong and Faugeras, 1996, Leclerc et al., 1998], appear to provide a relatively complete basis for exploiting imaging geometry in both depth recovery and in rejecting point correspondences that are inconsistent with the derived camera model. In this paper, we have little to add in this area. However, we do employ projective constraints beyond those directly associated with camera modeling. For example, We have implemented a filter that uses a plane-to-plane linear transform to reject errors

associated with semantically identified planar scene features.

And, of course, we do not wish to imply that additional advances are not needed in this area. We note that the HVS is not completely dependent on a projective model of the imaging process – it can recover a "qualitative" geometric model of a scene from highly distorted images.

2.3.3 Scene Based Constraints on Feature Matching

It is almost universally the case that stereo-based depth-recovery systems are designed to operate without reference to scene semantics. In the case of the HVS, it is commonly assumed that stereopsis occurs very early in the visual processing chain, is preattentive, and is based purely on some form of *local matching*. Julesz [Julesz, 1971] has shown that stereopsis can occur in the absence of any meaningful information in the individual images of a stereo pair. Never the less, we argue in this paper that stereo depth recovery in the natural outdoor world must invoke scene-specific semantic knowledge to create a relatively dense meaningful depth-model. For example, in the Tenaya Lake picture (Figure 4) most of the scene is composed of either sky or lake. The lake is especially interesting in that it can appear as a large *mirror-like* surface. Reflected objects can be matched to produce a depth map which is consistent over a large number of views, but which is incorrect. Under circumstances where the the water surface is refractive rather than reflective, we can again form valid matches which produce incorrect depth measurements (if we make the usual assumption that light travels in straight lines). On the other hand, if we know we are looking at a large flat body of water, we could profit from its known planar geometry to constrain matching of objects on its immediate boundary, and to obtain a correct depth model (via interpolation) for its surface. We can even correct for its refractive properties if we need to estimate its depth. In the case of the sky, we might determine that it is homogeneous, and thus not suitable for matching, without knowing what it is. However, where the sky is visible through the tree foliage, a purely geometric system might well try to interpolate depth from the surrounding valid matches – this works for the lake (which might also appear photometrically homogeneous) but is obviously incorrect for sky patches. There are a large number of similar considerations (e.g., fire, smoke, snow, insubstantial surfaces – such as grass or foliage, ...) that force one to conclude that some form of crude semantic overlay must be available to support a depth recovery system whose results must be reasonably complete and correct. We have previously developed techniques that could be used to compute a suitable semantic overlay [Fischler, 1996], but for most of the experiments described in this paper, we employed a method based on recent work using classification-trees [Breiman et al., 1984].

3 The Current Experimental Environment

We assume that the images to be processed were obtained with a stereo camera configuration similar to the HVS. Two essentially identical cameras (or a single camera) that is used to view the scene at approximately the same time from two closely spaced locations. The cameras have vertically oriented image-planes (approximately) parallel to each other. The two images of a stereo pair should be quite similar to each other modulo some projective and lens distortion, a horizontal shift in scene content, some differences in occluded regions, and some intensity variation due to film processing and non-Lambertian reflective surfaces in the scene. The currently implemented experimental configuration is limited to both black-and-white and color images of natural outdoor scenes; it is not intended to model scenes with man-made objects or aerial views. Figures 1 shows some stereo images we have used in our experiments.

Our goal in this experimental study is not the implementation and testing of the complete stereo system we envision, but rather to demonstrate that we can can extract a set of correct matches with a specified maximum percentage of errors in each uncontrived stereo pair we process and then show that based on such a sampling of "known correct" matches, and a "semantic overlay" constructed (nominally) in parallel with the matching results, we can obtain a dense depth map that is superior to conventional (2-image) stereo models. Some of the components and processing steps in our experimental work were chosen for convenience and accessibility, rather than reflecting the ideal design.

In order to compare our results to existing state-of-the art stereo/matching systems, and to illustrate the importance of the concepts proposed in our paper, we took advantage of an excellent publicly accessible `image-matching` algorithm[1] which implements a robust technique for binocular image matching by exploiting the epi-polar con-

[1] Available from INRIA at:
`http://www.inria.fr/robotvis/personnel/zzhang`

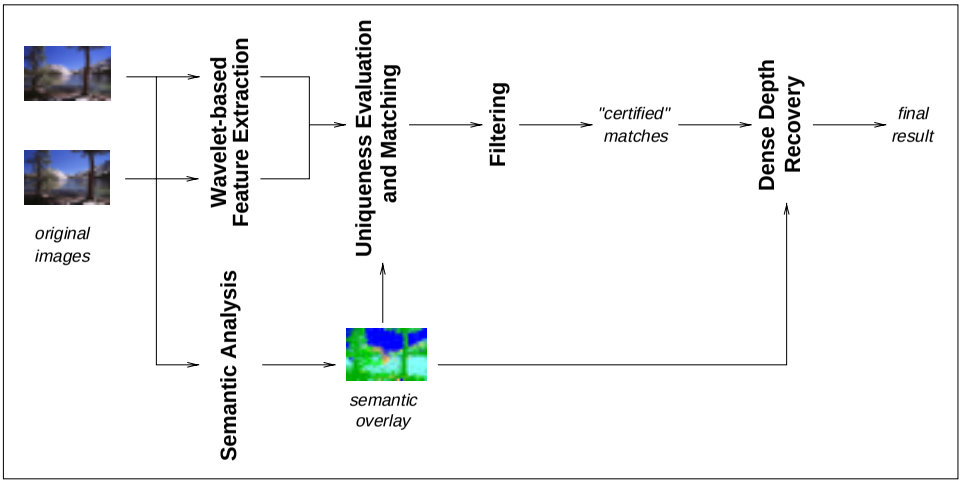


Figure 2: **Basic structure of the current experimental system.**

straint. It uses correlation and relaxation methods to find an initial set of matches, and then use the Least Median of Squares technique to discard false matches. (We realize that INRIA's latest developments (e.g. [Pennec and Thirion, 1997]) on stereo matching are not necessarily included in the available software.)

3.1 The System Architecture

The current experimental stereo configuration consists of several modules: a wavelet-based feature extraction module, a semantic analysis module, a uniqueness evaluation and matching module, a filtering module and an interpolation module for dense depth recovery. Figure 2 shows the basic structure of the system.

3.1.1 Wavelet-based Feature Extraction

For a given stereo pair, we first perform 2-level wavelet frame transforms on the images. Daubechies-4 wavelet is used to separate high frequency and low frequency information stored in the images. Unlike traditional wavelet transforms, we do not perform down sampling. Various experiments [IEEE, 1993, Wang et al., 1998a, Wang et al., 1998b] have shown that Daubechies' wavelets [Daubechies, 1988, Daubechies, 1992, Meyer, 1993, Kaiser, 1994] are well suited for characterizing localized information in natural signals such as sounds and images. We characterize the local intensity information at each pixel location in each image with a vector of seven wavelet coefficients, i.e. one low frequency coefficients and six high frequency coefficients in the six high level frequency bands. For color images, we use 21 wavelet coefficients obtained from three wavelet transforms in RGB color space.

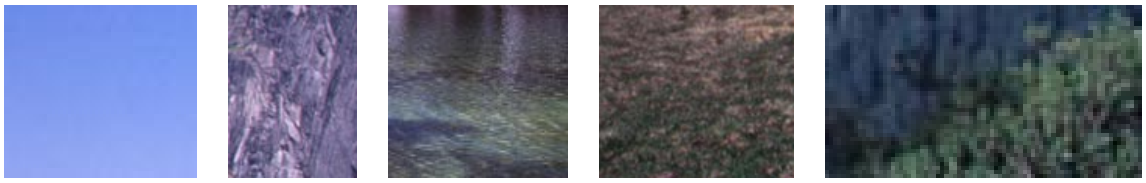
3.1.2 Semantic Analysis Using CART

We derive a rough semantic overlay of each image of the stereo pair. For most of the experiments discussed in this paper, we use training samples from a few scenes similar (but distinct) from the experimental scenes to create a decision tree structure using the classification and regression trees (CART) algorithms [Breiman et al., 1984]. CART, developed by Breiman et al., has been widely used in computer-aided clinical diagnosis research.

In our experiments, we used a sequence of seven training images representing *sky, stone, river/lake, grass* and *tree/forest*. Figure 3 shows five of the seven training images. We use the mean colors and variances of 4×4 blocks in RGB color space as the training feature vector. These features are simple but appear capable of distinguishing the above five classes. For gray scale images, we use only the mean intensity and variance of 4×4 blocks as the training feature vector.

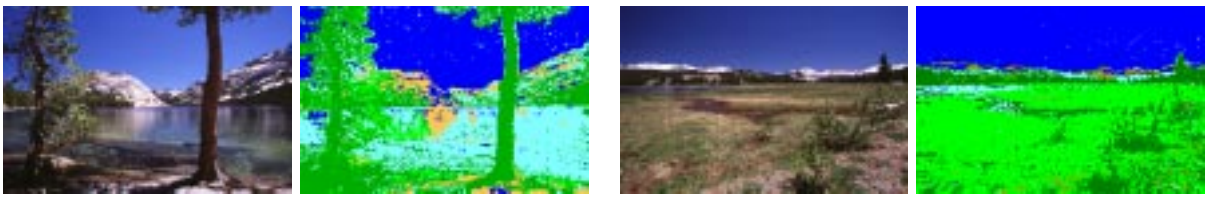
It takes about one minute on a Pentium PC to create the classification tree structure. After the classification tree is created, it takes only a few seconds to classify a given image to create the semantic overlay for a color image of 768×512 pixels. Figure 4 and 5 show the classification results on color and gray-scale images[2]. Each of the five different classes is given a

[2]The original color figures can be accessed through the WWW at:
`http://www-db.stanford.edu/~wangz/project/stereo/IUW98/`



sky *stone* *river/lake* *grass* *tree/forest*

Figure 3: **Training color images used for creating the semantic overlay.**



color image *semantic overlay* *color image* *semantic overlay*

Figure 4: **Semantic analysis of outdoor scenes using the classification and regression trees (CART) algorithm.** No post-processing is performed. Color scheme: Deep blue for sky, yellow for stone, light blue for river/lake, light green for grass, deep green for tree/forest.

unique "pseudo" color in the final result. The classification results are satisfactory for our application.

For stereo matching purposes, we exclude regions classified as sky and water because feature-based matching in these regions is not reliable. We can obtain dense stereo matching in the water region by interpolation based on more reliable stereo matches bounding such regions (or possibly, say a rock, situated within the lake boundary).

3.1.3 Uniqueness Ranking and Filtering

We first *filter out* points in large ambiguity sets (e.g. the sky and the lake) for computational efficiency, and then compute the uniqueness score/ranking of each point in both of the images.

The uniqueness score/ranking of a given point is the sum of the reciprocal Euclidean distances (in wavelet feature space) to every other (non-eliminated) point in both images of the pair. We eliminate from the computation the closest (in feature-space) point in the conjugate image and its adjacent (in image space) neighbors as the nominally correct match. The most unique points have the smallest uniqueness scores.

To reduce computational complexity, we approximate the above "ideal" computation. For example, points with high "partially-computed" uniqueness scores are discarded without completing their evaluation. In this step, we do not restrict the search to the same epi-polar line. (Stevenson [Stevenson and Schor, 1997] has shown that human stereo matching also is not restricted to epi-polar lines.)

We now filter the ranked points to obtain on the order of 300 to 500 conjugate pairs, where each pair satisfies the following condition: each member of a pair has only one other potential match in the set of unique points, and this single match is its conjugate in the other image.

3.1.4 Geometric and Appearance Based Filtering

We next compute the fundamental matrix [Luong and Faugeras, 1996] that models the imaging geometry between the two images of the stereo pair and eliminate all conjugate pairs that fail to satisfy the epi-polar "rigidity" condition. Since we nominally assume that we know the internal camera parameters (as needed to fully exploit the semantic overlay), in an ideal system we would replace the epi-polar constraint with the more comprehensive collinearity constraint [McReynolds and Lowe, 1996] to do the rigidity checking. We then further filter the surviving pairs on the basis of additional constraints derived from assumptions about scene geometry affecting two or more conjugate pairs.

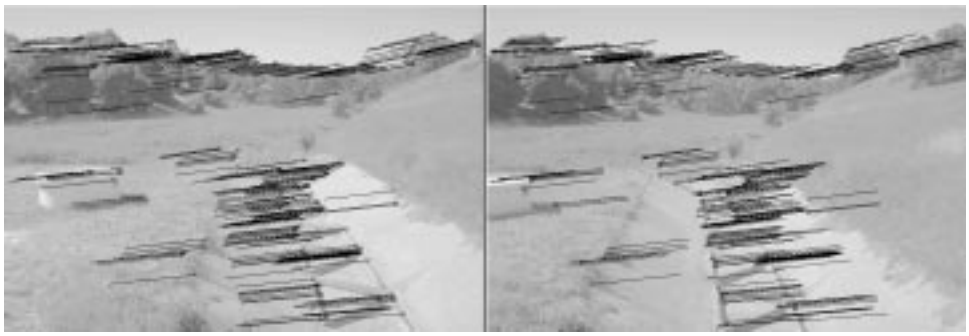
We use the surviving pairs and their wavelet-based feature vectors to compute a 7×7 covariance matrix and then rank the remaining pairs on the basis of the Mahalanobis distance between members of the each conjugate pair. Based on the assumption that the differences between the wavelet characterizations of

gray-scale image *semantic overlay* *gray-scale image* *semantic overlay*

Figure 5: **Semantic analysis of outdoor scenes using the classification and regression trees (CART) algorithm.** No post-processing is performed. Color scheme: Deep blue for sky, light blue for river/lake, light green for grass, deep green for tree/forest, white for non-classified regions.

SRI program : 279 matches, 1 error *INRIA's algorithm : 35 matches, 2 errors*

Figure 6: **Matching result using our program vs. INRIA's `image-matching` algorithm.** Dark points are the matches found. Lines shown are the disparity vectors. Our system found 279 matches including 1 mismatch (marked with white lines). INRIA's system found 35 matches including 2 mismatches.



SRI program : 300 matches, 0 error

INRIA's algorithm : 80 matches, 0 error

Figure 7: **Matching result using our program vs. INRIA's `image-matching` algorithm.** Dark points are the matches found. Lines shown are the disparity vectors.

SRI program : 289 matches, 4 error

INRIA's algorithm : 50 matches, 10 errors within the lake

Figure 8: **Matching result using our program vs. INRIA's `image-matching` algorithm.** Dark points are the matches found. Lines shown are the disparity vectors.

the members of a correctly associated conjugate pair can be approximated by a Gaussian process, we could set a threshold based on the Chi-squared distribution that allows us to eliminate any matches that have a probability of greater than (approximately) 2% of being in error. (The squared Mahalanobis distance has a Chi-squared distribution under the Gaussian assumption.) What if the Gaussian assumption does not hold?? We have found that the Mahalanobis distance consistently produces an acceptable ordering of the image points with respect to uniqueness for the class of natural scenes we are concerned with and it is possible to select a fixed threshold that virtually eliminates all but a very small percentage of errors – experimentally found to be on the order of 2-3 percent – while still returning on the order of 1-2 "certified correct" points per scan-line.

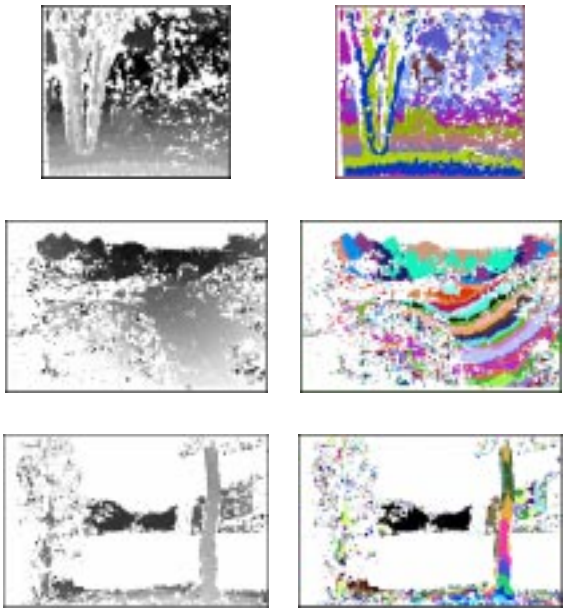
3.1.5 Dense-Modeling/Interpolation

The semantic overlay, certified correct matches, and computed epi-polar geometry, allow us to partition the images into *subregions* which are recursively processed by the above strategy.

The recursive search in this step is limited to the set of pixels surrounding the corresponding epi-polar lines in each of the two images. Figure 10 illustrates the pixels to be examined in this step.

Since each subregion has fewer points to cause mismatches, we obtain additional (nominally) correct matches and thus the final number of conjugate pairs use to construct the 3-D scene model, while a function of scene content (e.g., the extent of the sky region), is not constrained by the size of our initial set of *certified conjugate points*. Dense modeling of depth is based on the assured correct matches and the semantic overlay to provide an informed basis for interpolation.

At present, we have focused on sky and water constraints in exploiting the semantic overlay. Obviously, the sky regions are not assigned any finite depth value – they serve mainly to prevent the formation of incorrect correspondences or interpolation. It can easily be shown that for the imaging configuration we are assuming (known internal camera parameters and horizontal principal ray, we can estimate the elevation of a horizontal surface (e.g. a lake) relative to the focal point of the camera from a single correct correspondence of a point on or adjacent to the horizontal lake surface; and the distance to any point on the surface



gray-scale *random colormap*

Figure 9: **Results from the recursive dense depth recovering process using our program.** The disparity image is shown. White regions are the no-match regions. No interpolation has been performed.

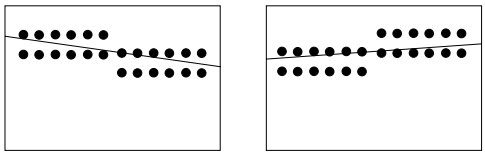


Figure 10: **Pixels surrounding the corresponding epi-polar line in each of the images are searched for final dense matching.**

or surface-level boundary of the lake can then also be directly computed without any additional correspondences.

3.2 Computational Complexity

The process we currently have in place is computationally expensive; computational complexity is $O(n^{1.5})$ for image of n pixels. However, we may apply tree-structured vector quantization (TSVQ) [Gersho and Gray, 1991, Riskin and Gray, 1991] to significantly speed-up the process. As reported in [Wang et al., 1998c], linear indexing time and constant or near constant searching time can be achieved using TSVQ for searching a Euclidean feature space.

3.3 Experimental Results

The system has been implemented using C on a UNIX platform. The discrete fast wavelet transforms were performed on a Pentium Pro 300MHz LINUX workstation. It takes about 40 seconds for the feature space classification module to process all the 200,000 feature vectors for a pair of images of 384×256 pixels. Approximately 40 minutes CPU time is required to perform the dense depth recovery.

Figures 6, 7, 9 and 8 show sample matching results obtained using our system compared with using INRIA's `image-matching` algorithm. Table 1 shows the efficiency of two of the filters measured for the tree scene (Figure 1).

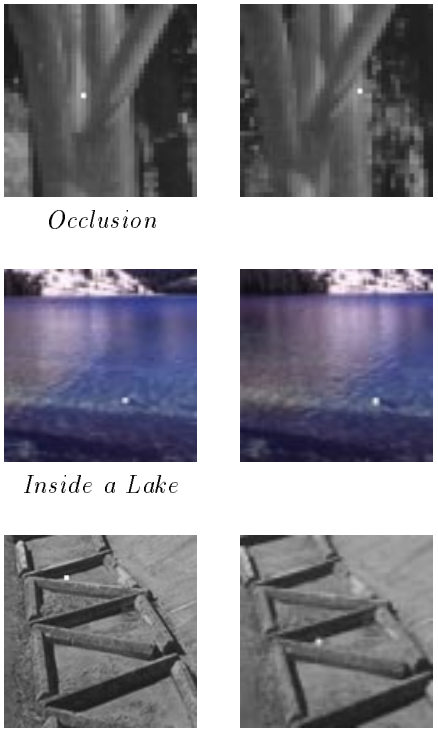
3.4 Discussion

What conclusions can we draw from the experiments? Both the SRI and INRIA algorithms made almost no errors in the "Lambertian regions" of the three test scenes, but the filtering efficiency (retention of correct correspondences for an essentially zero error rate) was much higher for the SRI algorithm. Thus our ability to create a complete and valid depth model, even for the "normal" regions of the natural scenes, was significantly greater. In the case of the sky regions, both algorithms did well, but for the water there were significant differences. Here, as expected, without the semantic overlay, the INRIA algorithm had a high error rate – on the order of 20 percent of the returned matches.

We believe that the key to high efficiency in the filtering step is to have an initial collection of error-free matches to be used to construct the covariance matrix and thus also the rank ordering of the points with respect to expectation of a correct match. To the extent

Filters	Total Matches Given to Filter	Eliminated Errors	Eliminated Valid Matches	Non-Eliminated Errors
Mahalanobis Distance Filter	293	9	0	2
Scene Geometry Filter	284	3	2	1

Table 1: **Filter efficiency (sequential execution) for the tree scene image pair.** Starting with 293 matches returned from the epi-polar filter, the two following filters found 12 additional errors but eliminated two valid matches and failed to find one remaining error.



Occlusion

Inside a Lake

Repeated Structure

Figure 11: **Typical mismatches that we are trying to eliminate.** Bright points are the mismatches eliminated by our filters.

that incorrect correspondences are included, the correctness ordering (Mahalanobis distance) is "noisy" and a threshold chosen to eliminate almost all errors will be forced to also eliminate many correct correspondences. Thus, by preventing the sky and water regions from producing any correspondences, we improve the efficiency of the filters, even for parts of the scene outside of the sky and water regions. This explains why we were willing to pay a high computational price for the uniqueness computation in addition to the construction of the semantic overlay.

The uniqueness ranking that we assign to each conjugate pair is based on "all" the information present in both images. We assume that an ambiguity condition detected far from the original point in the containing image, or far from the associated epi-polar line in the conjugate image, still suggests an increased probability of an undetected mismatch (e.g., due to occlusion so that only one close but incorrect match is found on the correct epi-polar line itself) – we have found many examples where this is indeed the case (see Figure 11).

We assume that the only valid basis for certainty judgments is the consensus of informed independent opinions. Photometric measures based on different characterizations of the image intensity pattern are not likely to be truly independent. Constraints from the nature of the imaging process add additional necessary, but not sufficient criteria for a correct match. Thus, the other available information sources, especially constraints based on semantics, physical laws, and known or assumed scene geometry must be invoked if we are to have any hope of duplicating the performance of human stereopsis.

Stereo modeling of a natural scene requires a parallel (primitive) semantic overlay to provide a basis for informed interpolation. This observation and its implications are central to our approach and a major departure from related work on this subject.

4 Conclusions and Future Work

In this paper, we addressed two related problems. First, we have explored the question of the extent to which judgments of similarity/identity can be made essentially "error-free." Most current approaches to robust matching focus on obtaining a consistent geometric model under a highly simplified set of assumptions about the imaging process and world being modeled. In the natural outdoor world, consistency is not sufficient; even a valid match does not insure correct depth recovery (e.g., the Tenaya-lake example). In the two image case, camera geometry constraints can, at best, restrict matching to epi-polar lines; at this point conventional systems usually rely on some form of local appearance matching and statistical arguments to complete the construction of the depth model. We show examples from non-contrived images where the statistics are valid but the matching is still incorrect. We argue that the HVS does not make these mistakes because it uses scene semantics as an additional, and more powerful, constraint on potential matches.

Second, we have examined the requirements for "human-level" stereo modeling in the natural outdoor world. Avoiding matching errors is only half the job: we can eliminate all the errors by eliminating all the matches. Consistency and statistical decision theory are not a sufficient basis for obtaining a relatively complete model when a significant portion of the scene content is "unmatchable" (i.e., when such matching is based strictly on intensity variations in the imagery). Interpolation into the unmatched regions can only be accomplished in a principled way if a semantic constraints are invoked and if semantic modeling accompanies geometric recovery.

This paper is still *work-in-progress*. We are attempting to better define the requirements of the semantic overlay, to make its automatic construction more robust, and to use it more effectively in the stereo matching process.

Acknowledgments

We would like to thank Yvan Leclerc, Quang-Tuan Luong, Marsha Jo Hannah, and various other researchers of the SRI AI Center and Jia Li of Stanford University Electrical Engineering Department for discussions and help. In particular, Yvan was involved in the early work leading to this paper and made important contributions with respect to our approach to evaluating uniqueness. We would also like to thank INRIA for making publicly available a highly robust image matching program for purposes of comparative evaluation.

References

[Barnard and Fischler, 1982] S T Barnard and M A Fischler, Computational Stereo, *ACM Surveys*, December 1982.

[Barrett et al., 1991] E B Barrett, P Payton, M H Brill and N N Haag, Linear Resection, Intersection, and Perspective-independent Model Matching in Photogrammetry: Theory, *Appl. Digital Image Processing XIV*, editor A. Tescher, Proc SPIE 1567, p. 142-169, 1991.

[Bhat and Nayar, 1998] D N Bhat and S K Nayar, Stereo and Specular Reflection, *Int J Computer Vision*, Kluwer, 26(2):91-106, 1998.

[Breiman et al., 1984] L Breiman, J H Friedman, R A Olshen and C J Stone, *Classification and Regression Trees*, Chapman and Hall, 1984.

[Daubechies, 1988] I Daubechies, Orthonormal bases of compactly supported wavelets, *Communications on Pure and Applied Mathematics*, 41(7):909-996, October 1988.

[Daubechies, 1992] I Daubechies, *Ten Lectures on Wavelets*, CBMS-NSF Regional Conference Series in Applied Mathematics, 1992.

[Fischler and Bolles, 1981] M A Fischler and R C Bolles, Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography, *CACM*, 24(6):381-95, June 1981; also, *Readings in Computer Vision*, (M A Fischler and O Firschein, eds.), Morgan Kaufmann, pp 726-40, 1987.

[Fischler, 1996] M A Fischler, Robotic Vision: Sketching Natural Scenes, *ARPA Image Understanding Workshop*, Feb 1996.

[Fua and Leclerc, 1994] P Fua and Y G Leclerc, Registration Without Correspondences, *CVPR* Seattle, June 1994.

[Gersho and Gray, 1991] A Gersho and R M Gray, *Vector Quantization and Signal Compression*, Kluwer Academic Publishers, 1991.

[Julesz, 1971] *B Julesz, Foundations of Cyclopean Perception*, Univ. of Chicago, Ill, 1971.

[Kaiser, 1994] G Kaiser, *A Friendly Guide to Wavelets*, Birkhauser, Boston, 1994.

[Leclerc et al., 1998] Y G Leclerc, Q.-T. Luong and Pascal Fua, Self-consistency: a Novel Approach to Characterizing the Accuracy and Reliability of Point Correspondence Algorithms, *DARPA Image Understanding Workshop*, 1998.

[Luong and Faugeras, 1996] Q.-T. Luong and O D Faugeras, The Fundamental Matrix: Theory, Algorithms, and Stability Analysis *Int J of Computer Vision*, 17(1):43–76, 1996.

[McReynolds and Lowe, 1996] D P McReynolds and D G Lowe, Rigidity Checking of 3D Point Correspondences Under Perspective Projection, *IEEE PAMI*, 18(12):1174-85, Dec 1996.

[Meyer, 1993] Y Meyer, *Wavelets: Algorithms & Applications*, SIAM, Philadelphia, 1993.

[Mikhail, 1976] E M Mikhail, *Observations and Least Squares*, IEP, New York, 1976. Also publ. by Harper and Row, 1980.

[Mundy and Zisserman, 1992] J Mundy and A Zisserman (eds), *Geometric Invariance in Computer Vision*, MIT Press, Cambridge, Mass., 1992.

[Pennec and Thirion, 1997] X Pennec, J.-P. Thirion, A Framework for Uncertainty and Validation of 3-D Registration Methods Based on Points and Frames, *Int J of Computer Vision*, 25(3):203-229, Kluwer, 1997.

[Riskin and Gray, 1991] E A Riskin and R M Gray, A Greedy Tree Growing Algorithm for the Design of Variable Rate Vector Quantizers. *IEEE Trans. Signal Process.*, November 1991.

[Rousseeuw, 1987] P J Rousseeuw, *Robust Regression and Outlier Detection*, Wiley, New York, 1987.

[Stevenson and Schor, 1997] S B Stevenson and C M Schor, et al., Human Stereo Matching is Not Restricted to Epipolar Lines, *Vision Research*, Elsevier, 37(19):2717-23, Oct 1997.

[Torr, 1995] P H S Torr, Motion Segmentation and Outlier Detection *University of Oxford Thesis*, 1995.

[Torr and Murray, 1997] P H S Torr and D W Murray, The Development and Comparison of Robust Methods for Estimating the Fundamental Matrix, *Int J Computer Vision*, Kluwer, 24(3):271-300, Sept/Oct 1997.

[Wang et al., 1998a] J Z Wang, G Wiederhold, O Firschein and X W Sha, Content-based Image Indexing and Searching Using Daubechies' Wavelets, *Int J Digital Libraries(IJODL)*, 1(4):311-328, Springer-Verlag, 1998.

[Wang et al., 1998b] J Z Wang, J Li, G Wiederhold and O Firschein, System for Classifying Objectionable Images, to appear in *Computer Communications J*, Elsevier Science, 1998.

[Wang et al., 1998c] J Z Wang, J Li and G Wiederhold, WISE: A Wavelet-based Image Search Engine with Efficient Feature Vector Clustering and Classification, submitted for journal publication, 1998.

[Weng et al., 1989] J Weng, T S Huang, and N Ahuja, Motion and Structure from Two Prospective Views: Algorithms, Error Analysis, and Error Estimation, *IEEE PAMI*, 11:451-476, 1989.

[Zhang et al., 1995] Z Zhang, R Deriche, O Faugeras, Q.-T. Luong, A Robust Technique for Matching Two Uncalibrated Images Through the Recovery of the Unknown Epipolar Geometry, *Artificial Intelligence*, 78(1-2):87-119, Elsevier, Oct 1995.

[Zhang et al., 1998] Z Zhang, et al., Determining the Epipolar Geometry and Its Uncertainty: a Review, *Int J Computer Vision*, 27(2):161-95, Kluwer, 1998.

[IEEE, 1993] Special Issue on Wavelets and Signal Processing, *IEEE Trans. Signal Processing*, Vol.41, Dec. 1993.