Visual Similarity, Judgmental Certainty and Stereo Correspondence [1]

James Z. Wang[2]
School of Information Sciences and Technology
The Pennsylvania State University
504 Rider Building, University Park, PA 16801, USA
FAX (814) 865-5604
jwang@ist.psu.edu

Martin A. Fischler
Artificial Intelligence Center, SRI International
333 Ravenswood Ave., Menlo Park, CA 94025, USA
fischler@ai.sri.com

[1]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. Original figures in this paper are located at the URL: http://www-db.stanford.edu/~wangz/project/stereo/J00/

[2]Work performed when the author was with SRI International and Stanford University.

Visual Similarity, Judgmental Certainty and Stereo Correspondence

James Z. Wang
The Pennsylvania State University

Martin A. Fischler
SRI International
(Summary)

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 feature correspondences 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 98% correctness on *asserted* matches. Finally, we provide a principled basis for relatively dense depth recovery.

Abstract

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 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. When tested on different sets of stereo images, our system has demonstrated above 98% correctness on *asserted* matches.

Keywords: Stereo matching; Statistical methods; Classification tree; Wavelet transform; Semantic overlay.

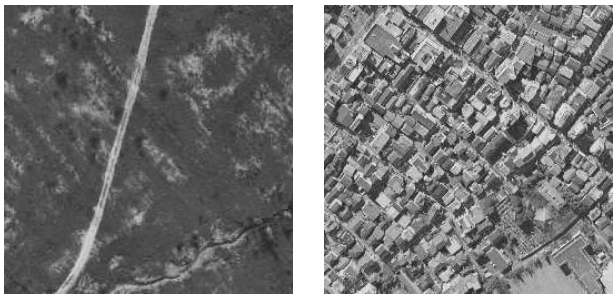
1 Introduction

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 [15, 2, 27, 9, 6, 10, 24, 28, 17, 21, 25, 29] or to select correspondences from a predetermined set that are consistent with a "rigid" spatial configuration.

Stanford Main Quad | Stanford Hill | Stanford Trees | Yosemite Lake

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



Ft. Benning | Kawasaki

Figure 2: **Aerial imagery used for our experimental investigation.**

In this paper, we explore the question of the extent to which judgments of similarity/identity (believed to be the basis of human stereopsis) can be made *essentially* error-free in the context of stereo matching in the natural outdoor world. And further, how such a (possibly sparse) set of correspondences could provide a dense depth model.

The paper is organized as follows. In Section 2, we discuss the key problems to be addressed and our approach to their solution. Section 3 describes our experimental environment. The details of our matching algorithm are given in Section 4. Section 5 presents experimental results on real-world image data. Section 6 discusses the experimental results. Finally, Section 7 presents conclusions and future directions.

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 [19, 7, 22, 1, 20, 16, 13], 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 included in our system 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 pre-attentive, and is based purely on some form of *local matching*. Julesz [11] 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 10) 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. Fischler has developed techniques that could be used to compute a suitable semantic overlay [8]. For the experiments described in this paper, we employed a method based on statistical classification and regresssion trees (CART) [3] developed at Stanford University.

3 The 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 are 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 was intended for ground-level images of natural outdoor scenes; it was not intended to model scenes with man-made objects or aerial views. However, we did apply it to aerial imagery and outdoor scenes with man-made objects. Figures 1 and 2 show some of the images we have used in our experiments.

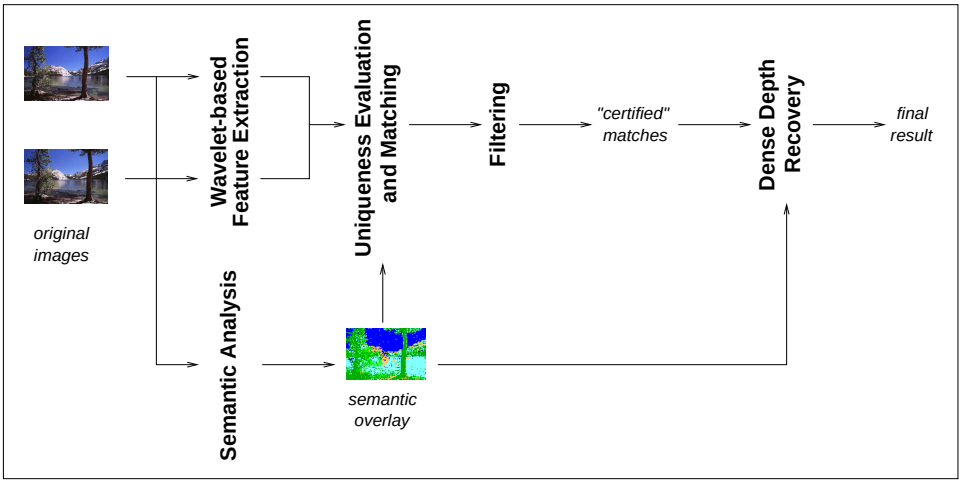


Figure 3: **Basic structure of the current experimental system.**

Our goal in this experimental study was 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 constraint. 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.

A pervasive problem in stereo/matching research is the evaluation of results obtained from experiments using real images, and especially when the data is ground-level imagery of natural outdoor scenes. Some of our earlier evaluations in this effort were based on a "manual" assessment of each asserted match-pair. Our more recent experiments (most involving aerial imagery) exploited some new evaluation ideas and techniques [14] which do not require the availability of "ground truth." The basic idea is that if we can acquire three or more calibrated images that cover the same area, and we can find a common point appearing in two or more asserted matches (across three or more images), then a necessary condition that this set of matches are all correct is that they all "recover" the same ground point. While in a real application we could use the evaluation software to eliminate errors when three or more views were available, for the purposes of this paper, we always restricted our input data to the algorithm to be a single stereo image-pair.

4 The Matching Algorithm

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 3 shows the basic structure of the system. Figure 4 shows the computational flow of our matching algorithm.

In this section, we provide the details of the matching algorithm. The algorithm takes two images as input. It can also take advantage of a precomputed fundamental matrix if available.

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

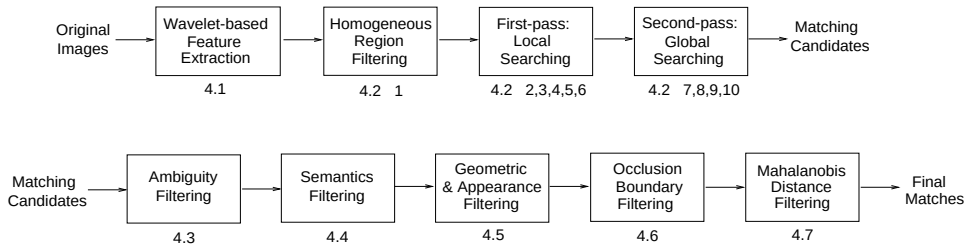


Figure 4: **The main steps in the matching algorithm.**

Assume an image is specified by a set of pixels $\mathfrak{I} = \{(i, j),\ i = 0, ..., m-1,\ j = 0, ..., n-1\}$. We denote the pair of images to be matched as $\mathfrak{I}_1$ and $\mathfrak{I}_2$. We crop and substract the images by their averages so that they are of the same size, $m \times n$, and of roughly the same brightness. If the 3×3 fundamental matrix for the pair of images is given, we denote it to be F.

4.1 Wavelet-based Feature Extraction

Various experiments [30, 26] have shown that Daubechies' wavelets [4, 5, 18, 12] 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 coefficient and six high frequency coefficients, obtained from framed wavelet transforms.

Figure 5: **A normal 3-level wavelet transform with down-sampling.**

1. Apply the Daubechies-4 wavelet filter to each row of the image. We obtain a low-pass vector of length n and a high-pass vector of length n for each row. For each original image, we obtain two matrices of coefficients, each having the same dimensions as the original image. We denote these matrices $L_{\mathfrak{I}_1}$ and $H_{\mathfrak{I}_1}$ for the first image, and $L_{\mathfrak{I}_2}$ and $H_{\mathfrak{I}_2}$ for the second image.

2. *Without* down-sampling, transpose the four matrices, $L_{\mathfrak{I}_1}$, $H_{\mathfrak{I}_1}$, $L_{\mathfrak{I}_2}$ and $H_{\mathfrak{I}_2}$. This step is different from the traditional wavelet transform where a down-sampling is performed.

3. Apply the Daubechies-4 wavelet filter to each row of the four transposed matrices. We obtain a low-pass vector of length m and a high-pass vector of length m for each row. For each of these matrices, we obtain two matrices of coefficients, each having the size $n \times m$. We denote these matrices $LL_{\mathfrak{I}_1}$, $LH_{\mathfrak{I}_1}$, $HL_{\mathfrak{I}_1}$, and $HH_{\mathfrak{I}_1}$, for the first image, and $LL_{\mathfrak{I}_2}$, $LH_{\mathfrak{I}_2}$, $HL_{\mathfrak{I}_2}$, and $HH_{\mathfrak{I}_2}$, for the second image.

4. *Without* down-sampling, transpose the eight matrices. Now we get eight matrices of the same size, $m \times n$. Again, this step is different from the traditional wavelet transform. Figure 5 shows a normal 3-level wavelet transform. Figure 6 shows the notations.

5. Apply Step 1 to Step 4 on the matrices $LL_{\mathfrak{I}_1}$ and $LL_{\mathfrak{I}_2}$ to obtain an additional four matrices for each one of them. Now we have decomposed each original image into seven matrices of distinct frequency bands, three from previous steps and four from this step.

6. Shift the matrices in both dimensions for 4 pixels so that the coefficients in the matrices correspond to the actual location of the pixels in the original image. During the image matching process, we avoid matching in the border area due to the boundary problem with the wavelet filtering.

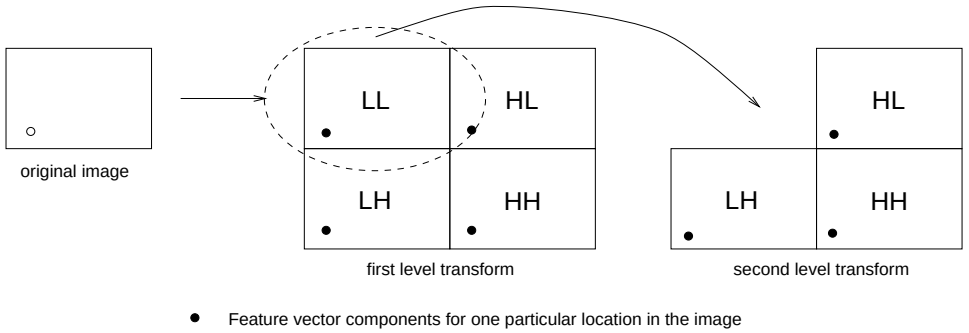


Figure 6: **Forming the feature vector.**

7. For each pixel in the original images, we collect the corresponding coefficients in the seven matrices to form a feature vector of seven dimensions. We denote these 7-dimensional vectors as $\mathfrak{V}_{\mathfrak{I}_1(i,j)}$ and $\mathfrak{V}_{\mathfrak{I}_2(i,j)}$. Figure 6 illustrates the process.

4.2 Uniqueness Evaluation

Correct stereo matching requires an evaluation of both uniqueness and similarity. We first evaluate the uniqueness of the wavelet descriptor at each pixel location in the image. Denote A as the matrix containing the *ambiguity scores* of the *pixels* in the image. The size of A is $m \times n$.

In this step, we do not restrict the search to a single epi-polar line. (Stevenson [23] has shown that human stereo matching also is not restricted to epi-polar lines.)

If the fundamental matrix F is given, we perform the following computation.

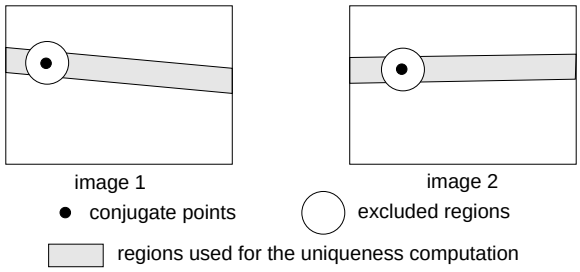


Figure 7: **Uniqueness evaluation.**

1. We exclude homogeneous regions by checking the high frequency wavelet coefficients in the feature vectors. If the coefficients for a particular point are too small, we do not consider that point as a match candidate.

2. Initialize the matrix A to a zero matrix.

3. Set two constants s_1 and s_2 to *small* values (e.g. 3-5).

4. For each pixel $\mathfrak{I}_1(i_1, j_1)$ in the left image, use the fundamental matrix F to find the corresponding band of width $2s_1$ adjacent epi-polar lines of pixels in the right image. Here, i_1 and j_1 are indexes.

5. For each pixel within the band of $2s_1$ epi-polar lines, denoted as $\mathfrak{I}_2(i_2, j_2)$, compute the Euclidean distance between the feature vectors $\mathfrak{V}_{\mathfrak{I}_1(i_1,j_1)}$ and $\mathfrak{V}_{\mathfrak{I}_2(i_2,j_2)}$. Denote the distance as $d(i_1, j_1; i_2, j_2)$. Here, i_2 and j_2 are indexes.

6. Sort the distances $d(i_1, j_1; i'_2, j'_2)$, where (i'_2, j'_2) run over all the pixels in the band of $2s_1$ adjacent epi-polar lines. Find the closest match to $\mathfrak{I}_1(i_1, j_1)$ and denote it as $\hat{\mathfrak{I}}_2(i_2, j_2)$. That is, $d(i_1, j_1; i_2, j_2)$ is the minimum over all d's. We call the pair of points as a conjugate pair.

7. For all pixels $\mathfrak{I}_2(i'_2, j'_2)$ in the second image such that $(i'_2 - i_2)^2 + (j'_2 - j_2)^2 < {s_2}^2$, we compute the maximum $d(i_2, j_2; i'_2, j'_2)$. Denote the maximum as t_2.

8. For all pixels $\mathfrak{I}_2(i'_2, j'_2)$ in the pixel band in the second image such that $(i'_2 - i_2)^2 + (j'_2 - j_2)^2 > {s_2}^2$, if $d(i_1, j_1; i'_2, j'_2) < t_2$ holds, we discard the match. Otherwise,

$$A(i_1, j_1) = A(i_1, j_1) + \frac{1}{d(i_1, j_1; i'_2, j'_2)} \quad .$$

9. For all pixels $\mathfrak{I}_1(i'_1, j'_1)$ in the first image so that $(i'_1 - i_1)^2 + (j'_1 - j_1)^2 < {s_2}^2$, we compute the maximum $d(i_1, j_1; i'_1, j'_1)$. Denote the maximum as t_1.

10. For all pixels $\mathfrak{I}_2(i'_1, j'_1)$ in the pixel band in the first image such that $(i'_1 - i_1)^2 + (j'_1 - j_1)^2 > {s_2}^2$, if $d(i'_1, j'_1; i_2, j_2) < t_1$ holds, we discard the match. Otherwise,

$$A(i_1, j_1) = A(i_1, j_1) + \frac{1}{d(i'_1, j'_1; i_2, j_2)} \quad .$$

We note that t_1 and t_2 are computed thresholds on acceptable uniqueness. Figure 7 shows that the regions around the conjugate pair are excluded for the uniqueness evaluation. If the fundamental matrix F is not given as an input, we use a band of $2s_1$ rows of pixel around the point in the first image to determine the uniqueness of the point.

We now have an ambiguity score matrix A for the image pair. During the matching phase, we require that the components of a match pair is not only similar but also unique.

4.3 Initial Matching

In this part of the process, we try to find a list of about N conjugate pairs that satisfy criteria for both uniqueness and similarity. For our applications, we set N to a (nominal) value of 300. This value is not critical because the system is designed to filter mismatches obtained from this initial matching step.

If the fundamental matrix F is given, we perform the following procedure.

1. Without considering the homogeneous regions, sort values in the ambiguity matrix A.

2. Set $1 \rightarrow c$.

3. If $c > N$, terminate.

4. For each pixel in the left image with the next smallest ambiguity score in A, denoted as $\hat{\mathfrak{I}}_1(i_1, j_1)$, compute the global ambiguity score over the entire right image.

5. If the global ambiguity score is smaller than a threshold determined similar to t_2, we use the Euclidean distance between the feature vectors $\mathfrak{V}_{\mathfrak{I}_1(i_1,j_1)}$ and $\mathfrak{V}_{\mathfrak{I}_2(i'_2,j'_2)}$ to find a best match within the adjacent $2s_1$ epi-polar lines in the right image. Denote the match as $\hat{\mathfrak{I}}_2(i_2, j_2)$.

6. Use the Euclidean distance between $\mathfrak{V}_{\mathfrak{I}_1(i'_1,j'_1)}$ and $\mathfrak{V}_{\mathfrak{I}_2(i_2,j_2)}$ to find a best match within the adjacent $2s_1$ epi-polar lines in the left image. Denote it $\hat{\mathfrak{I}}_1(i_3, j_3)$.

7. If $|i_3 - i_1| > 1$ or $|j_3 - j_1| > 1$, we discard the match.

8. Find the best match of $\mathfrak{I}_1(i_1, j_1)$, denoted $\hat{\mathfrak{I}}_k(i_4, j_4)$, within the entire two images except the neighborhoods of the points $\mathfrak{I}_1(i_1, j_1)$ and $\mathfrak{I}_2(i_2, j_2)$. Denote the distance between $\mathfrak{I}_1(i_1, j_1)$ and $\mathfrak{I}_k(i_4, j_4)$ as d_1.

9. Find the best match of $\mathfrak{I}_2(i_2, j_2)$, denoted $\hat{\mathfrak{I}}_l(i_5, j_5)$, within the entire two images except the neighborhoods of the points $\mathfrak{I}_1(i_1, j_1)$ and $\mathfrak{I}_2(i_2, j_2)$. Denote the distance between $\mathfrak{I}_1(i_1, j_1)$ and $\mathfrak{I}_l(i_5, j_5)$ as d_2.

10. If $d(i_1, j_1; i_2, j_2) < d_1$ and $d(i_1, j_1; i_2, j_2) < d_2$, accept the match $\mathfrak{I}_1(i_1, j_1)$ and $\mathfrak{I}_2(i_2, j_2)$ as a valid conjugate pair for the initial matching stage. Otherwise, discard the match.

11. Set $c + 1 \rightarrow c$. Go to Step 3.

We have now obtained on the order of 300 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.

If the fundamental matrix F is not given as an input, we use a band of $2s_1$ rows of pixel around the point in the images for the computation.

4.4 Semantic Overlay Filtering

If the scene is a ground-level natural out-door scene, we create a rough semantics overlay to eliminate matches in regions (e.g. sky and water) where we have little confidence of finding correct conjugate pairs. If the scene is not a natural out-door scene, we skip this filter step.

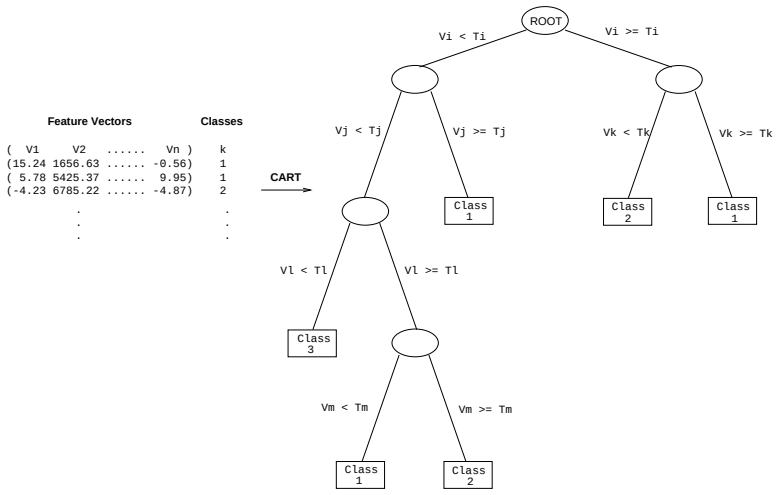


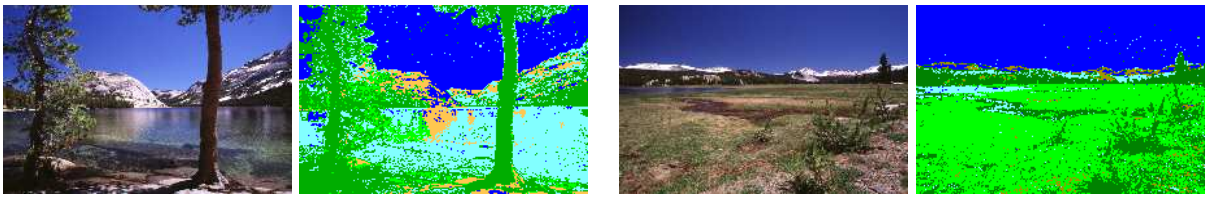
Figure 8: **Generating a classification tree using the CART algorithm.**

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 [3]. CART, developed by Breiman et al., has been widely used in computer-aided clinical diagnosis research. Figure 8 shows the structure of a classification tree generated by the CART algorithm.

In our experiments, we used a sequence of seven training images representing *sky, stone, river/lake, grass* and *tree/forest*. Figure 9 shows five of the seven training images. We use the mean colors and variances of 4×4 blocks in RGB color space as the components of the training feature vector. These features are simple

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

Figure 9: **Training color images used for creating the semantic overlay.**

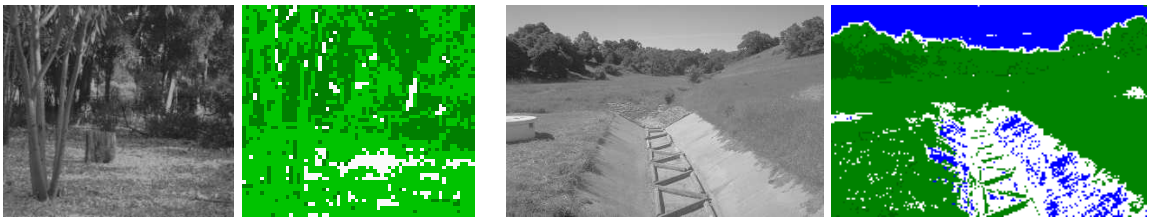


color image *semantic overlay* *color image* *semantic overlay*

Figure 10: **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.

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 components of 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 10 and 11 show the classification results on color and gray-scale images[2]. Each of the five different classes is given a unique "pseudo" color in the final result. The classification results are satisfactory for our application.



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

Figure 11: **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.

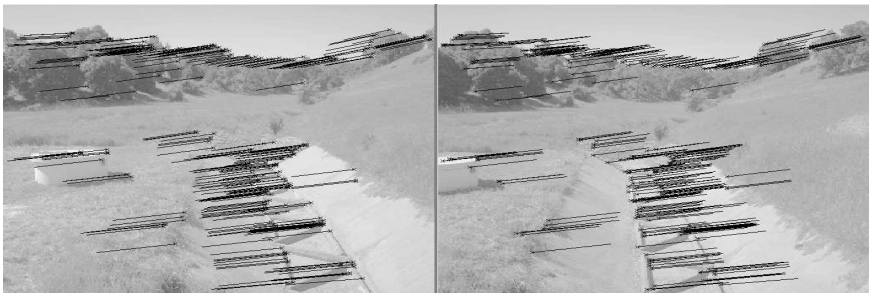
For stereo matching purposes, we exclude regions classified as sky and water because feature-based matching in these regions is not reliable. We can obtain dense stereo matching in the water region by interpolation based on more reliable stereo matches bounding such regions.

[2]The original color figures can be accessed through the WWW at:
`http://www-db.stanford.edu/~wangz/project/stereo/J00/`

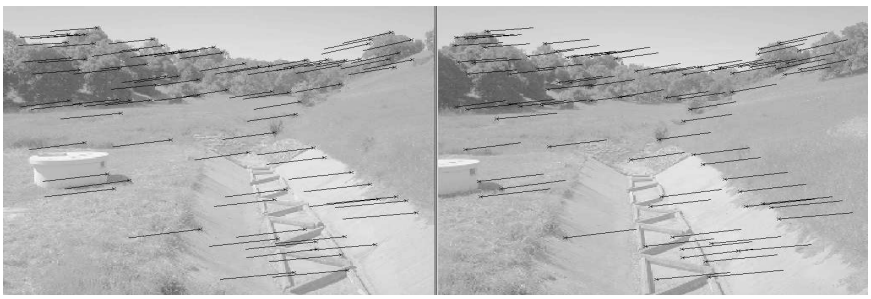
SRI program : 279 matches, 1 error

INRIA's algorithm : 35 matches, 2 errors

Figure 12: **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 errors



INRIA's algorithm : 80 matches, 0 errors

Figure 13: **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 errors

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

Figure 14: **Matching result using our program vs. INRIA's `image-matching` algorithm.** Dark points are the matches found. Lines shown are the disparity vectors.

4.5 Geometric and Appearance Based Filtering

We next compute the fundamental matrix [16] 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 [17] to perform 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.

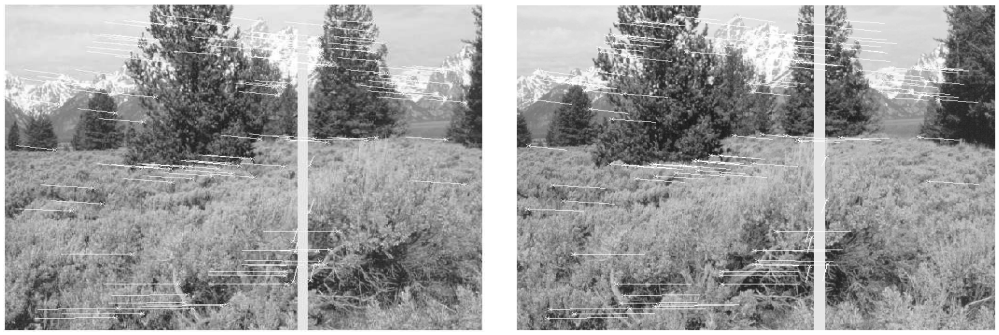
4.6 Occlusion Boundary Filtering

In this step, we eliminate matches on possible occlusion boundaries to avoid matching "psuedo features" composed of both foreground and background components. The procedure is based on the assumption that the intensity differences between corresponding points surrounding a given point on an occlusion boundary is less random than the differences surrounding a given point on a continuous surface.

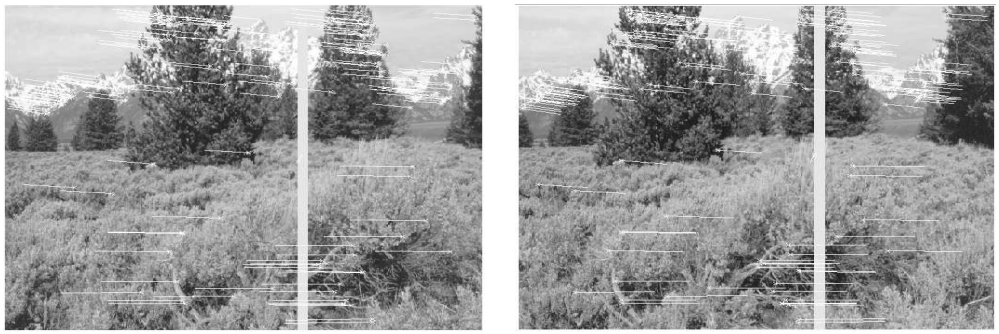
Assume the match $\mathfrak{I}_1(i_1, j_1)$ and $\mathfrak{I}_2(i_2, j_2)$ is to be checked. The procedure is as follows:

1. As indicated in Figure 16, we partition the surrounding 16×16 pixel block of a match point into 4×4 sub-blocks.

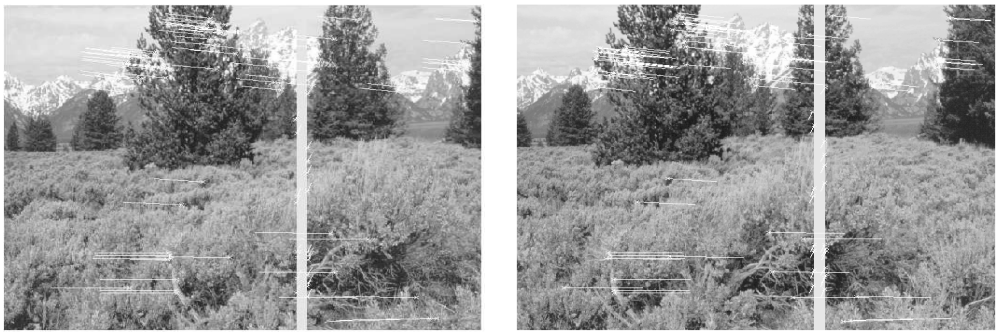
2. Threshold the 16×16 block of intensity value differences between the first image and the second image to obtain a binary "difference block", denoted as B_0. The threshold is determined adaptively for each difference block to maintain about fifty percent ones in the 16×16 block. The threshold is typically around 12 for an 8-bit image.



INRIA's algorithm : 90 matches, 10 errors (*)



SRI's algorithm : 194 matches, 0 errors



106 matches eliminated by the occlusion boundary filter (**)

Figure 15: **Performance of the occlusion boundary filter.** (*) A match near but outside the boundary of the inserted artificial foreground object is considered an error. (**) Of the 106 matches eliminated, 10 matches are near but outside the boundary of the inserted artificial foreground object.

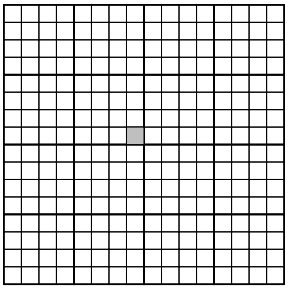


Figure 16: **Partition the** 16×16 **pixel block surrounding a match point into** 4×4 **sub-blocks for the occlusion boundary filtering.**

3. Denote the 4×4 sub-blocks as $B_{i,j}$, where $i = 1, 2, 3, 4$ and $j = 1, 2, 3, 4$. Let

$$p_{i,j} = \frac{summation\ within\ B_{i,j}}{summation\ within\ B_0} \ .$$

4. If $\chi^2 = \sum_{i,j} \frac{(p_{i,j} - \frac{1}{16})^2}{\frac{1}{16}}$ is larger than a threshold, we discard the match pair from the list of initial matches.

Here we use the equivalent of a Chi-square test to evaluate the hypothesis that the intensity differences in the 16×16 block are uniformly distributed. The two thresholds currently used were determined based on experiments on real data. If the first threshold was chosen to produce 50% ones, then the Chi-square test with 15 degrees of freedom and a rejection threshold of $0.15 \times 128 = 18.2$ would result in a 25% probability of rejecting a valid match.

Figure 15 shows the performance of the occlusion boundary filter. We inserted an artificial occlusion boundary in each member of a pair of ground-level images. INRIA's program asserted some incorrect matches on the occlusion boundary. Our program eliminated all potential matches on the inserted occlusion boundary.

4.7 Mahalanobis Distance Filtering

For each semantic label (separately) 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 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 less than 2 percent – while still returning on the order of 1-2 "certified correct" points per scan-line.

4.8 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 18 illustrates the pixels to be examined in this step.

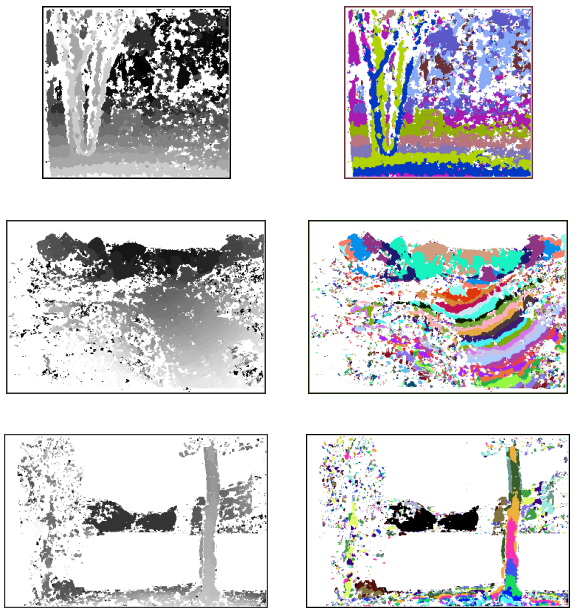


Figure 17: **Results from the recursive dense depth recovering process using our program.** The disparity image is shown. White regions are the no-match regions. No interpolation has been performed.

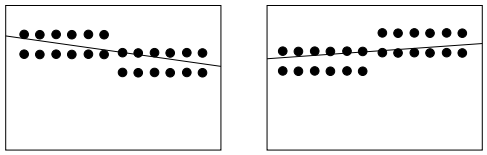


Figure 18: **Pixels surrounding the corresponding epi-polar line in each of the images are searched for final dense matching.**

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

5 Experimental Results

SRI program : 296 matches, 0 errors

INRIA's algorithm : 50 matches, 10 errors

Figure 19: **Matching result using our program vs. INRIA's `image-matching` algorithm.** Dark points are the matches found. Lines shown are the disparity vectors.

The system has been implemented using C on a Pentium III LINIX PC. Figures 12, 13, 17, 14, 19 and 21 show sample matching results obtained using our system compared with using INRIA's `image-matching` algorithm.

We have performed a series of more than 100 experiments. For the ground-level natural outdoor scenes, we visualize the disparity maps to determine obvious errors (i.e., matches at least a few pixels away from the true match). For the aerial imagery, we use the SCT (self-consistency test) system [14] developed by SRI. Table 1 summarizes the data sets we have used. Figure 20 shows the cumulative distribution functions of

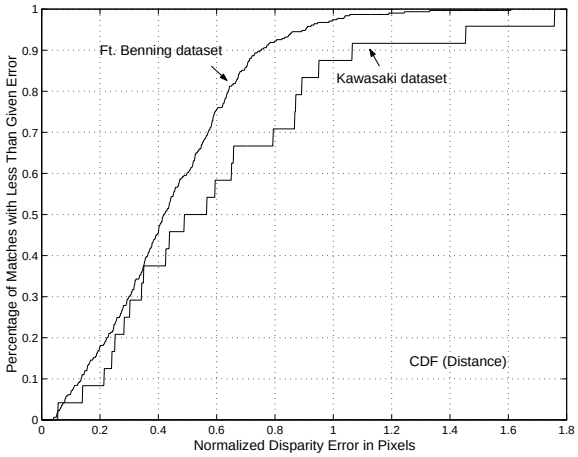
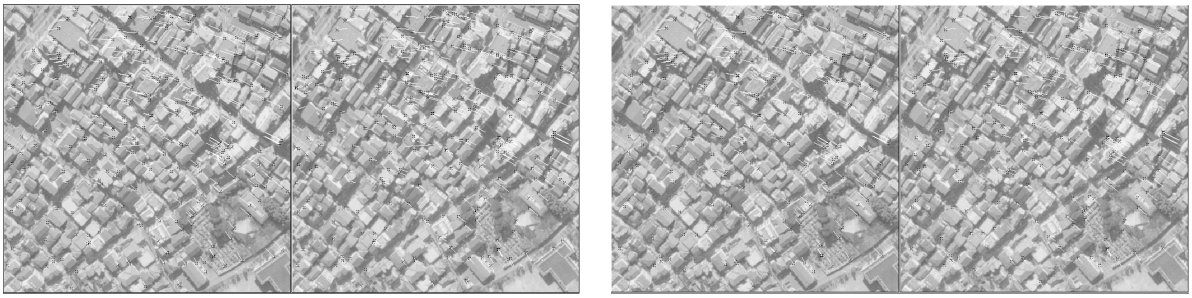


Figure 20: **Self-consistency test (SCT) results using the SRI program.**

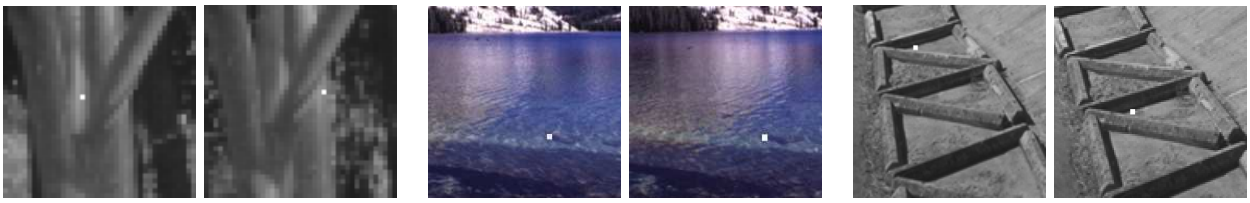


SRI program : 273 matches, 2 errors *INRIA's algorithm : 90 matches, 1 error*

Figure 21: **Matching result using our program vs. INRIA's `image-matching` algorithm.** Dark points are the matches found. Lines shown are the disparity vectors.

errors, obtained from the SCT system. Our system was not intended to provide sub-pixel accuracy matching data.

6 Discussion



Occlusion *Specular Surface (lake)* *Repeated Structure*

Figure 22: **Typical mismatches that we are trying to eliminate.** Bright points are the mismatches eliminated by our filters.

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 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 22).

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.

7 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

Location	Stanford Main Quad	Stanford Hill	Stanford Trees	Yosemite Lake	Ft. Benning	Kawasaki
# of Images	2	2	2	2	18×6 (**)	6×6 (**)
Image Dimensions ($rows \times cols$)	294×445	512×768	233×256	294×447	400×400	450×450
Externally Supplied Data	F	F	F	F, S	F	F
Acquisition Geometry						
Type	ground	ground	ground	ground	air	air
Nominal Camera-baseline	2m	2m	1m	3m	N/A	N/A
Nominal Range in Disparity	15-60 pixels	60-100 pixels	0-20 pixels	20-50 pixels	0-30 pixels	0-30 pixels
Ground Surface Resolution (meters/pixel)	N/A	N/A	N/A	N/A	0.30	0.50
Performance of the **SRI** System						
# of Conjugate Pairs Returned	295	300	285	289	240-300 (*)	240-280 (*)
Valid Matches	$> 99\%$	$> 99\%$	$> 99\%$	$> 99\%$	$> 99\%$	$> 98\%$
Gross Errors	$< 1\%$	$< 1\%$	$< 1\%$	$< 1\%$	$< 1\%$	$< 2\%$
Self Consistency	N/A	N/A	N/A	N/A	100% within 2 pixels	100% within 2 pixels
Performance of the **INRIA** System						
# of Conjugate Pairs Returned	50	80	50	50	80	70
Valid Matches	$> 80\%$	$> 99\%$	$> 90\%$	$> 80\%$	$> 99\%$	$> 98\%$
Gross Errors	$< 20\%$	$< 1\%$	$< 10\%$	$< 20\%$	$< 1\%$	$< 2\%$

Table 1: **Performance comparison of the SRI system and the INRIA system.** F: fundamental matrices. S: semantic overlay computed. (*) The SRI system did not find more than 100 matches in some cases. (**) number of sets $\times$ number of images in a set. Lenses with 50mm focal length and 35mm film format were used for ground level scenes.

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.

Since the matching problem is "open-ended," 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 for 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

[1] E. B. Barrett, P. Payton, M. H. Brill, N. N. Haag, Linear resection, intersection, and perspective-independent model matching in photogrammetry: theory, Appl. Digital Image Processing XIV, A. Tescher (ed.), Proc. SPIE, vol. 1567, 1991, pp. 142-169.

[2] R.C. Bolles, H.H. Baker, D.H. Marimont, Epipolar-plane image analysis: an approach to determining structure from motion, International Journal of Computer Vision, 1:7-55, 1987.

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

[4] I. Daubechies, Orthonormal bases of compactly supported wavelets, Communications on Pure and Applied Mathematics, vol. 41, no. 7, Oct. 1988, pp. 909-996.

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

[6] O. Faugeras, What can be seen in three dimensions with an uncalibrated stereo rig?, ECCV'92, Lecture notes in Computer Science, vol 588, G. Sandini (ed.), Springer-Verlag, pp. 563-578, 1992.

[7] M. A. Fischler, R. C. Bolles, Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, Comm. ACM, vol. 24, no. 6, Jun. 1981, pp. 381-95.

[8] M. A. Fischler, Robotic vision: sketching natural scenes, ARPA Image Understanding Workshop, Feb. 1996.

[9] P. Fua, Y. G. Leclerc, Registration Without Correspondences, Proc. CVPR, Seattle, Jun. 1994.

[10] R.I. Hartley, R. Gupta, T. Chang, Stereo from uncalibrated cameras, Proc. CVPR, 1992.

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

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

[13] Y. G. Leclerc, Q.-T. Luong, P. Fua, Self-consistency: a novel approach to characterizing the accuracy and reliability of point correspondence algorithms, Proc. DARPA Image Understanding Workshop, 1998.

[14] Y. G. Leclerc, Q.-T. Luong, P. Fua, Characterizing the performance of multiple-image point-correspondence algorithms using self-consistency, Proc. ICCV, Greece, September 1999.

[15] B.D. Lucas, T. Kanade, An iterative image registration technique with an application to stereo vision, Proc. International Joint Conference on Artificial Intelligence, Vancouver, pp. 674-9, 1981.

[16] Q.-T. Luong, O. D. Faugeras, The fundamental matrix: theory, algorithms, and stability analysis, Int J of Computer Vision, vol. 17, no. 1, 1996, pp. 43-76.

[17] D. P. McReynolds, D. G. Lowe, Rigidity checking of 3D point correspondences under perspective projection, IEEE Tran. PAMI, vol. 18, no. 12, Dec. 1996, pp. 1174-85.

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

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

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

[21] 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, vol. 25, no. 3, Kluwer, 1997, pp. 203-29.

[22] P. J. Rousseeuw, Robust regression and outlier detection, Wiley, New York, 1987.

[23] S. B. Stevenson, C. M. Schor, Human stereo matching is not restricted to epipolar lines, Vision Research, Elsevier, vol. 37, no. 19, Oct. 1997, pp. 2717-23.

[24] P. H. S. Torr, Motion segmentation and outlier detection, University of Oxford Thesis, 1995.

[25] P. H. S. Torr, D. W. Murray, The development and comparison of robust methods for estimating the fundamental matrix, Int J. Computer Vision, vol. 24, no. 3, Kluwer, Sep/Oct 1997, pp. 271-300.

[26] J. Z. Wang, J. Li, R. M. Gray, G. Wiederhold, Unsupervised multiresolution segmentation for images with low depth of field, IEEE Tran. PAMI, vol. 23, no. 1, 2001.

[27] J. Weng, T. S. Huang, N. Ahuja, Motion and structure from two prospective views: algorithms, error analysis, and error estimation, IEEE Tran. PAMI, vol. 11, 1989, pp. 451-76.

[28] 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, vol. 78, no. 1-2, Elsevier, Oct 1995, pp. 87-119.

[29] Z. Zhang, Determining the epipolar geometry and its uncertainty: a review, Int J. Computer Vision, vol. 27, no. 2, Kluwer, 1998, pp. 161-95.

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

About the Author – JAMES Z. WANG is the PNC Technologies Career Development Professor at the School of Information Sciences and Technology and Department of Computer Science and Engineering at the Pennsylvania State University. He received a Summa Cum Laude Bachelor's degree in Mathematics and Computer Science from University of Minnesota (1994), an M.S. in Mathematics and an M.S. in Computer Science, both from Stanford University (1997), and a Ph.D. degree from Stanford University Biomedical Informatics and Computer Science Departments (2000). He has been a visiting scholar at Uppsala University in Sweden, SRI International, IBM Almaden Research Center, and NEC Computer and Communications Research Lab. His research interests include semantics-sensitive image retrieval, large-scale genomic database retrieval, image classification and processing, and computer vision.

About the Author – MARTIN A. FISCHLER is a principal scientist of SRI International and a fellow of the IEEE. He directs all vision and perception research in the SRI Artificial Intelligence Center. He received a B.S. in Electrical Engineering from the College of the City of New York, an M.S. and a Ph.D. in Electrical Engineering, both from Stanford University.