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`

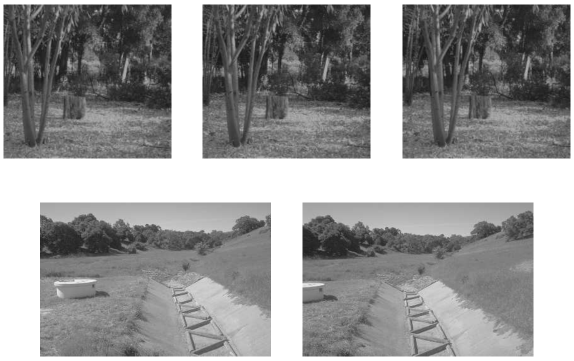

Figure 1: **Natural outdoor scenes used for our experimental investigation.**

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`


original
images