Spacetime Stereo: A Unifying Framework for Depth from Triangulation

James Davis
Honda Research Institute
jedavis@ieee.org

Diego Nehab
Princeton University
diego@cs.princeton.edu

Ravi Ramamoorthi
Columbia University
ravir@cs.columbia.edu

Szymon Rusinkiewicz
Princeton University
smr@cs.princeton.edu

Abstract

Depth from triangulation has traditionally been investigated in a number of independent threads of research, with methods such as stereo, laser scanning, and coded structured light considered separately. In this paper, we propose a common framework called spacetime stereo *that unifies and generalizes many of these previous methods. Viewing specific techniques as special cases of this general framework leads to insights regarding solutions to many of the traditional problems of individual techniques. Specifically, we discuss a number of possible applications such as improved recovery of static scenes under variable illumination, spacetime stereo for moving objects, structured light and laser scanning with multiple simultaneous stripes or patterns, and laser scanning of shiny objects. To suggest the practical utility of the framework, we use it to analyze two of these applications—recovery of static scenes under variable, but uncontrolled and unstructured illumination, and depth estimation in dynamic scenes. Based on our analysis, we show that methods derived from the spacetime stereo framework can be used to recover depth in situations in which existing methods perform poorly.*

Keywords: Depth from Triangulation, Stereo, Spacetime Stereo.

1 Introduction

A representation of three dimensional scene geometry is required for many tasks in computer vision, robotic navigation, computer graphics, and rapid prototyping, and a variety of techniques have been proposed for acquiring the geometry of real-world objects. This paper considers methods that obtain depth via triangulation. Within this general family, a number of methods have been proposed including stereo [16, 30], laser stripe scanning [5, 13, 14, 20], and time- or color-coded structured light [3, 9, 17, 18, 31]. Although a deep relationship exists between these methods, as illustrated in the classification of figure 1, they have been developed primarily in independent threads of the academic literature, and are usually discussed as if they were separate techniques. This paper presents a general framework called spacetime stereo for understanding and classifying methods of depth from triangulation. By viewing each technique as an instance of a more general framework, solutions to many of the traditional limitations within each sub-space become apparent.

Depth from triangulation makes use of at least two known scene viewpoints. Corresponding features from the different viewpoints are identified, and rays are intersected to find the 3D position of each feature. Determining the correct correspondence between viewpoints is the fundamental challenge, and it is in this area that the various methods can be distinguished.

Most previous surveys classify triangulation techniques into *active* and *passive* methods [5, 12, 25, 34]. Active techniques, such as laser scanning and structured light, intentionally project illumination into the scene in order to construct easily identifiable features in order to minimize the difficulty involved in determining correspondence. In contrast, passive stereo algorithms attempt to find matching image features between a pair of general images about which nothing is known a priori. This classification has become so pervasive that we believe it is artificially constraining the range of techniques proposed by the research community.

This paper proposes a different classification of algorithms for depth from triangulation. We characterize methods by the domain in which corresponding features are located. Techniques such as traditional laser scanning and passive stereo typically identify features purely in the *spatial domain*; i.e., correspondence is found by determining similarity of pixels in the image plane. Methods such as time-coded structured light and temporal laser scanning make use of features

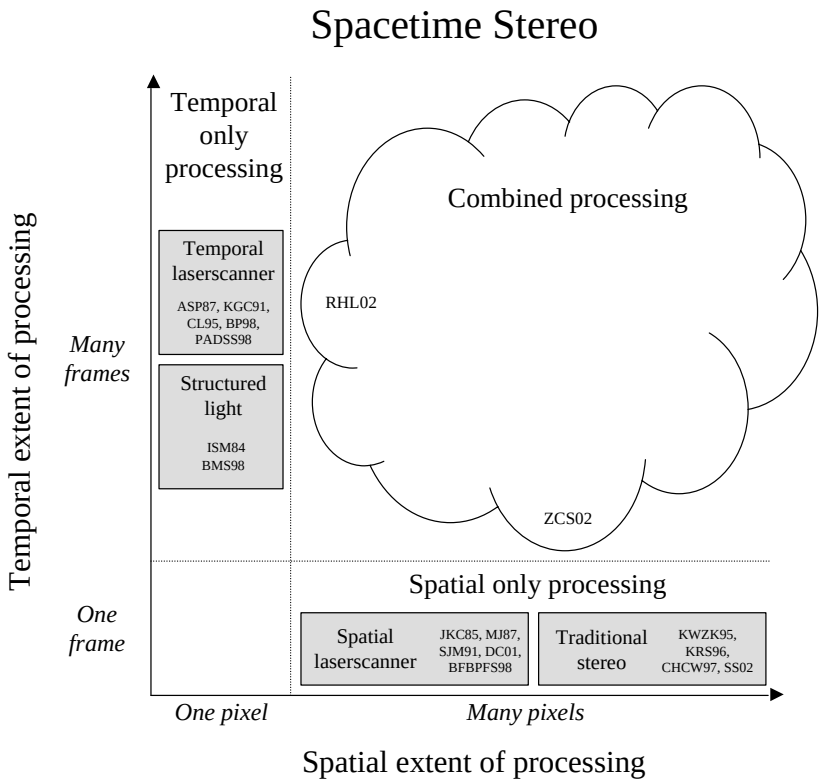

Figure 1: *Most existing depth from triangulation techniques are specific instances of the more general class of spacetime stereo reconstruction. Because these methods have been developed largely independently, they have often been artificially constrained to a small range of variation. Understanding that all these techniques lie in a continuum of possible methods can lead to previously unexplored modifications and hybrids.*

which lie predominantly in the *temporal domain*. That is, pixels with similar appearance over time are considered to be corresponding. Most existing methods locate features wholly within either the spatial or temporal domains. However it is possible, and this paper will argue desirable, to locate features within both the space and time domains using the general framework of *spacetime stereo*.

The remainder of this paper proposes a new framework for triangulation methods built around identifying corresponding features in both the space and time domains. This framework of space-time stereo allows a deeper understanding of the relationship between existing techniques that were previously considered separately. In particular, it unifies certain active and passive techniques by recognizing that they perform very similar computations to establish correspondences between two viewpoints. In addition, the new framework suggests extensions to existing techniques to permit

greater flexibility, accuracy, or robustness. We propose a number of these extensions, and describe new systems to which they may be applied. This unified framework and discussion are the primary contributions of this work.

In order to evaluate the practical utility of this framework we analyze the accuracy of depth recovery for two particular classes of scenes. The first is those in which geometry is static but illumination undergoes uncontrolled variation. We call this condition *unstructured light*, to distinguish it both from structured light methods in which lighting variation is strictly calibrated, and from passive stereo in which lighting variation is typically ignored. In our experiments, this variation is produced by the light and shadows from a hand held flashlight, or using a hand-held laser pointer. The tradeoffs between space and time are investigated by evaluating the possible combinations of spatial and temporal processing. The second class of scenes are those in which the object moves. Again we investigate the tradeoffs between spatial and temporal processing. In both cases, we demonstrate results indicating that spacetime stereo can recover depth maps with greater accuracy and robustness than traditional spatial-only stereo.

This paper is a considerably expanded version of a previous conference paper [15], and includes new results on shape recovery for dynamic scenes, as well as a discussion of optimal spacetime windows in that context. We are not alone in proposing that spatio-temporal information may be useful. Zhang et al. have simultaneously developed methods similar to ours, focusing on recovery of dynamic scenes rather than on constructing an organizing framework [37]. Other applications have been explored as well. For example, Shechtman et al. suggest that a spatio-temporal framework will be useful for increasing the resolution of video sequences [33].

The rest of this paper is organized as follows. In section 2, we describe the Spacetime Stereo framework. In section 3, we discuss previous work, classifying previous approaches as special cases of this new framework. In section 4, we discuss a number of possible extensions and improvements to previous methods enabled by the spacetime stereo framework. In section 5, we show reconstruction results and analysis of optimal spacetime stereo windows for both static and dynamic scenes. Finally, in section 6 we present conclusions and discuss future work.

2 Spacetime Stereo

In this section, we introduce our spacetime stereo framework for characterizing depth-from-triangulation algorithms. We discuss traditional spatial stereo, temporal stereo, and how they may be combined into a common spacetime stereo framework. Finally, we categorize errors in the spatial and temporal domains, showing the relationships between the two. In the next section, we will discuss previous work, classifying these methods in the spacetime stereo framework based on whether they identify features in the spatial or temporal domains.

2.1 Traditional (spatial) stereo

The spacetime stereo framework can most naturally be understood as a generalization of traditional passive stereo methods that operate entirely within the spatial (image) domain. Traditional stereo depth reconstruction proceeds by considering two viewpoints in known positions, and attempting to find corresponding pixels in the two images. This search for correspondence can proceed either by searching for specific features such as corners in each of the images, or more typically via matching of arbitrary spatial windows in the first image to corresponding regions along the epipolar line in the second image. More specifically, stereo finds correspondences by minimizing a matching function, which in its simplest form is

$$\left\| I_1(V_s(x_1)) - I_2(V_s(x_2)) \right\|^2. \tag{1}$$

Here I_1 is the intensity in image 1, I_2 is the intensity in image 2, and V_s is a vector of pixels in a *spatial* neighborhood close to x_1 (or x_2). This is the standard minimization of sum of squared differences to find the best matching pixel x_2^*.

There is a natural tradeoff in choosing the size of the neighborhood to be used. If the neighborhood is too small (an extreme case is a single pixel), there may be many pixels along the epipolar line that match the pixel in the first image equally well. If the neighborhood is larger, we often obtain more disambiguating information (since we are matching larger vectors). However, more information does not always become available. For example, regions of constant color introduce no new information. Further, we increase the chance that the spatial window will include depth discontinuities. In this situation, it becomes difficult or impossible to find correct matches. Because

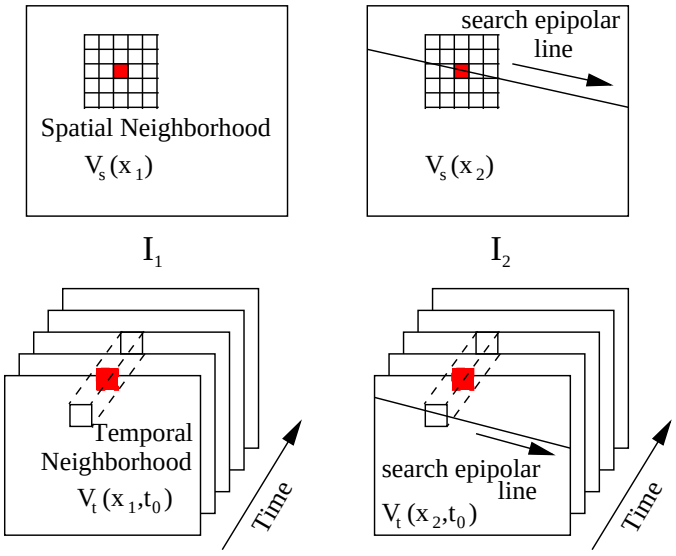

Figure 2: *Comparison of spatial (top) and temporal (bottom) stereo. In spatial stereo, the epipolar line is searched for similar spatial neighborhoods. In temporal stereo, the search is for similar temporal variation.*

of this tradeoff between finding a unique match (which may not be possible if the neighborhood is too small) and avoiding discontinuities and distortions (which can occur if the neighborhood is too large), traditional stereo methods sometimes lack robustness and often can not return dense depth estimates.

2.2 Temporal stereo

In order to show how spacetime stereo reconstruction relates to traditional spatial stereo, let us first consider a scene with static geometry that is viewed for multiple frames across time. In this new *temporal* stereo setting, we match a single pixel from the first image against the second image. As previously discussed, a unique match is unlikely and the size of the matching vector must be increased. Rather than increasing this vector by considering a neighborhood in the spatial direction, it is possible to increase the vector in the temporal direction, as shown in figure 2.

More specifically, we minimize a matching function,

$$\left\| I_1(V_t(x_1,t_0)) - I_2(V_t(x_2,t_0)) \right\|^2. \tag{2}$$

This is analogous to equation 1, except that now instead of a spatial neighborhood we consider a *temporal* neighborhood V_t around some central time t_0.

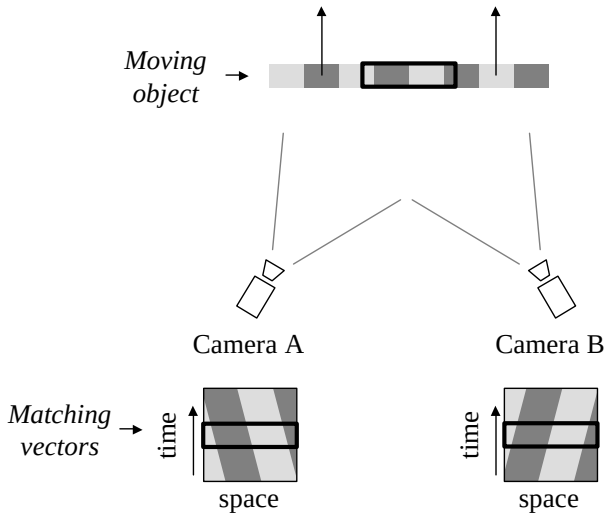

Figure 3: *Distortions in temporal stereo caused by a moving object. The matching vector is drawn in 2D with one spatial dimension and one time dimension. Rows in the matching vectors represent moments in time. Although the highlighted time instant is in direct correspondence, the temporal neighborhood has been subjected to distortion.*

Under some conditions, a temporal matching vector is preferable to the traditional spatial vector, such as if the lighting in a static scene is changing over time. In this case a long temporal sequence can be used to construct a matching vector. This vector may contain significantly more disambiguating information than a spatial matching vector, since, unlike with spatial windows, there are no disadvantages to increasing the window size. On the other hand, temporal matching can fail for dynamic scenes, since the same image pixel may no longer correspond in different frames to the same object. We will discuss this issue at the end of the section, showing that scene motion in temporal stereo is exactly analogous to depth variation in spatial stereo, leading to similar difficulties with increased neighborhood size.

2.3 Spacetime stereo

In general, there is no reason to restrict the matching vector to lie entirely along either the spatial or temporal axes. The matching vector can be constructed from an arbitrary spatio-temporal region around the pixel in question. In the case of rectangular regions, a window of size $N \times M \times T$ can be chosen, where N and M are the spatial sizes of the window, and T is the dimension along the

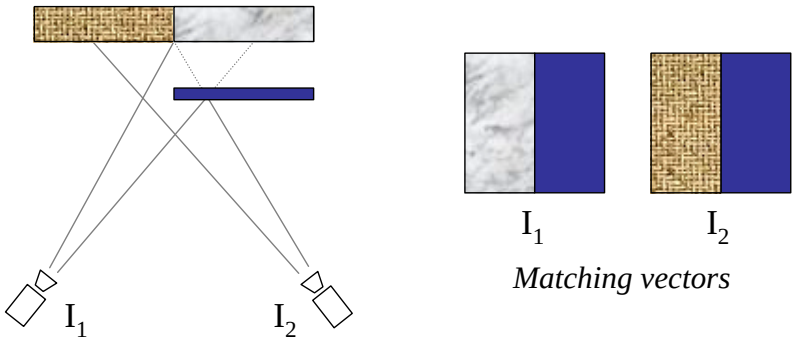

Figure 4: *Near a depth discontinuity, the spatial windows used by traditional stereo can contain multiple objects, as in this example. This often results in incorrect reconstructed depths.*

time axis. In our general framework, we would seek to optimize the matching function,

$$\left\| I_1(V_{st}(x_1,t_0)) - I_2(V_{st}(x_2,t_0)) \right\|^2. \tag{3}$$

It is clear that there is no mathematical distinction between the spatial and temporal axes.

This framework of generalized spacetime stereo is the main contribution of this paper. In section 3 we will classify many previous methods as particular instances of this more general framework. Seeing these techniques as part of a continuum rather than as isolated techniques can lead to previously unexplored modifications and hybrids.

2.4 Spatial and temporal domain errors

Matching errors can arise in both the spatial and temporal domains, and there is a natural tradeoff in determining the size of the neighborhood to use. In this subsection, we will discuss the relationship between errors in the spatial and temporal domains.

In spatial stereo matching, regions of constant texture (e.g. solid white objects) create difficulties, since increasing the size of the matching vector does not introduce new information to disambiguate likely matches. Similarly, in temporal stereo, regions with constant illumination over time do not introduce any new information. Spatial matching will perform best on objects textured with high spatial frequency, and temporal matching will perform best when the scene illumination has high temporal frequency.

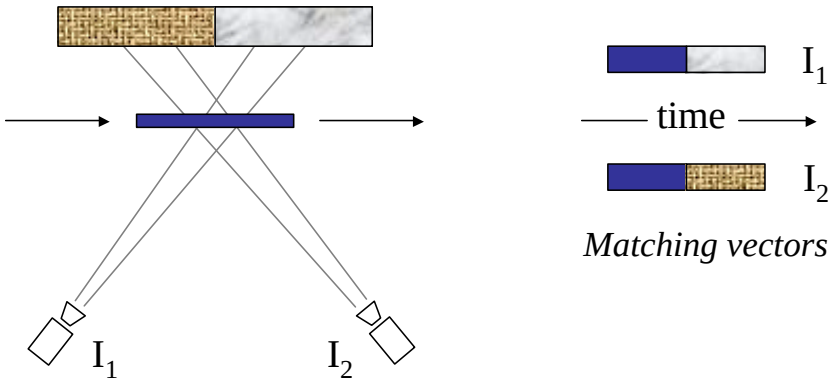

Figure 5: *Temporal stereo errors because of temporal depth discontinuities from motion. Initially, the two views do match, but as the box moves, they no longer do so. This situation is analogous to that in spatial stereo matching.*

In spatial stereo, a patch in the scene will create identical, undistorted images on both cameras only if it is correctly oriented at constant depth with respect to the two spatial viewpoints. A planar patch with other orientations will in general be subject to an arbitrary projective warp (homography) between the two views. Analogously, in the temporal domain, moving objects will not have constant depth over time and will produce a similar distortion in the time direction, as shown in figure 3.

Lastly, in spatial stereo matching, depth discontinuities between objects create matching neighborhoods with two separate regions that cannot be simultaneously matched. Typically, one region is on the foreground, and one on the background, as shown in figure 4. When a temporal matching vector is used, moving objects cause the same sort of discontinuity. If an object moves, leaving a view of the background, a discontinuity will exist in the temporal vector, as shown in figure 5. At first, the vectors do indeed match; at some point, a discontinuity creates a new region and the vectors no longer match, a situation analogous to the spatial case.

3 Previous methods

Several well-investigated categories of research are in fact special cases of the general spacetime stereo framework discussed above. These include traditional stereo, time-coded structured light,

and laser stripe scanning. While this paper does not attempt an exhaustive survey of existing methods, a classification of the algorithms discussed is given in figure 1. Note that the well-defined categories of existing research determine feature correspondence purely within either the spatial or temporal domains, and the realm of spatio-temporal processing remains largely unexplored.

3.1 Stereo

Traditional stereo matching is a well studied problem in computer vision. A number of good surveys exist [16, 30]. As discussed in section 2.1, traditional stereo matches vectors in the spatial or image domain to determine correspondence. In passive stereo methods, no attempt is made to create easy features for correspondence, and the vectors or spatial neighborhoods matched are arbitrary. Active stereo methods project a high-frequency static pattern onto the object to aid in determining correspondences, improving performance in areas of little texture [11, 22, 23].

Another common depth estimation technique is photometric stereo [35]. In this approach, multiple light source positions are used with a fixed camera. Variations in shading allow surface normals, and thus surfaces to be estimated. Although this method seems similar in that it makes use of lighting variations, it is a fundamentally different method since it obtains shape from shading, rather than using triangulation. Hybrid technologies that combine both methods have been proposed [4], as have techniques that use Helmholz reciprocity to provide robustness to lighting and BRDF variations [38].

Some researchers have encoded camera motion as a temporal sequence, and applied volumetric processing [6]. Although the method of epipolar analysis is well known, and uses similar terminology to this work, it is not directly related. The "spatio-temporal volumes" of that work encode camera position, rather than time, making it more closely related to multibaseline stereo and structure from motion. Similarly, our framework is not directly related to recent research which provides unifying theories for multiperspective [32] and multiocular [2] stereo.

3.2 Time-coded structured light

Time-coded structured light methods determine depth by triangulating between projected light patterns and an observing camera viewpoint. A recent survey of these methods is by Batlle et al. [3].

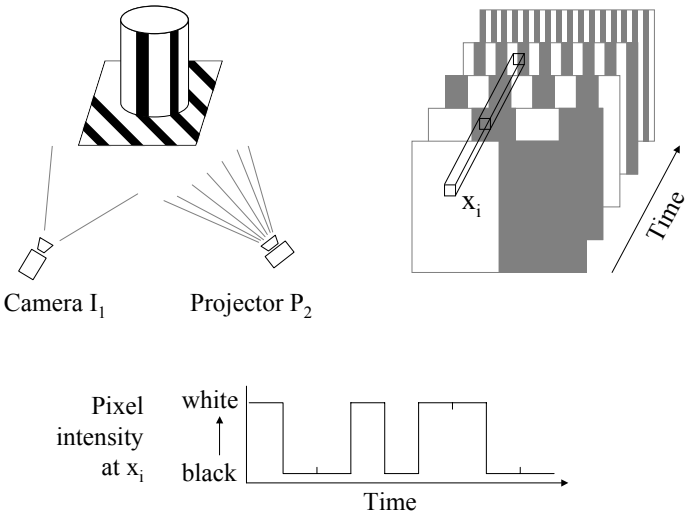

Figure 6: *Structured light scanning. A set of known temporal patterns is projected onto the object. These patterns induce a temporal matching vector, shown below.*

The projector illuminates a static scene with a temporally varying pattern of light stripes. The patterns are arranged such that every projected column of pixels can be uniquely identified. Thus, the depth at each camera pixel is determined based on the particular pattern observed. That is, the matching vector is temporal and is matched against a known database of projected patterns and their associated depths, as shown in figure 6. Although the example in this figure is simplistic, a wide variety of projected patterns are possible, and much of the work in this area has focused on designing optimum patterns in terms of either minimum sequence length or robustness, such as the gray coding used by Inokuchi et al. [18].

From the above description, we can see that structured light is a special case of spacetime stereo with matching in the temporal domain. The matching error metric can be written as

$$\left\| I_1(V_t(x_1,t_0)) - P_2(V_t(x_2,t_0)) \right\|^2, \tag{4}$$

which is similar to equation 2 except that we have replaced the second image I_2 with known projected patterns P_2. This is functionally equivalent to having a *virtual* second camera collocated with the projector. The virtual camera has the same viewpoint as the lightsource, so the virtual image it captures can be assumed identical to the projected light. By making conceptual use of a second camera, depth recovery in structured light systems can be described in terms of correspondence between images, similar to traditional stereo.

It should be noted that the second camera need not be virtual. Using an additional real camera has a number of benefits, including improving the robustness of correspondence determination to variations in object reflectance [11], and generating high quality ground truth stereo test images [31].

3.3 Laser stripe scanning

A typical laser scanner has a single camera and a laser that sweeps across a scene. Many geometries have been proposed, but for the purposes of this discussion all behave similarly. A plane of laser light is generated from a single point of projection and is moved across the scene. At any given time, the camera can see the intersection of this plane with the object. Depths can be determined using either spatial or temporal processing, and both types of scanners have been built. Informative surveys have been provided by Besl [5] and Jarvis [20].

Most commercial laser scanners function in the spatial domain. The laser sheet has an assumed Gaussian cross section, and the location of this Gaussian feature is known in the laser frame of reference. Given a known laser position, the epipolar line in the camera image is searched for a matching Gaussian feature [28]. This match determines corresponding rays, and thus a depth value. Since the feature set lies only on one line in image space, rather than densely covering the image plane, only a single stripe of depth values is recovered. This process is repeated many times with the laser positioned such that the stripe of features is in a new location.

The search for a laser stripe is conceptually similar to sparse feature matching in that we are looking for features with a known signature in the spatial domain, and matching these features between views. Spatial laser scanning is subject to some of the same difficulties that complicate traditional stereo matching. In particular, no good match is possible in the neighborhood of a depth discontinuity.

Laser scanners that function in the temporal domain have also been built [1, 21]. As the laser sweeps past each pixel, the time at which the peak intensity is observed is recorded and used to establish correspondence, as shown in figure 7. Curless and Levoy [13] provide an analysis of the benefits that temporal correlation provides over the traditional spatial approach in the context of

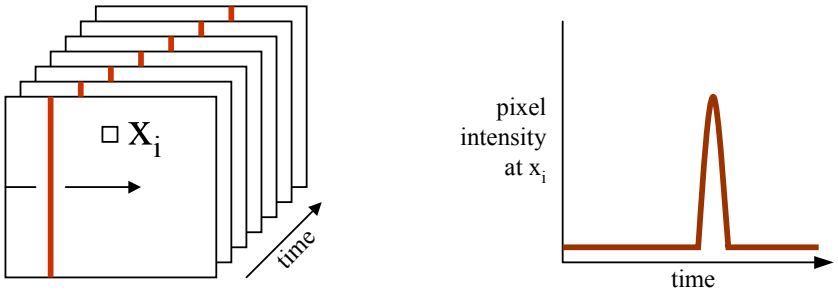

Figure 7: *Temporal processing of laser scan data. The peak intensity in the temporal matching vector indicates the time at which the laser stripe crosses a pixel, in turn indicating the pixel's depth.*

laser scanning. Moreover, they show that the optimal matching uses feature vectors that are not strictly aligned with the time axis, but are "tilted" in spacetime.

It should be noted that systems qualitatively similar to laser scanners can be built by replacing the laser stripe with any well-defined and uniquely identifiable light pattern. For instance, Bouguet and Perona [8] have demonstrated a scanner that uses a planar shadow generated using a calibrated lamp and hand-held stick.

As with coded structured light, laser scanning can be framed as standard stereo matching by replacing the calibrated laser optics with a second calibrated camera. With this modification, the laser stripe functions as the high frequency texture desirable for stereo matching, though since the variation only occurs in a small region, only a small amount (one stripe's worth) of valid data is returned at each frame. Multi-camera implementations have been built that find correspondence in both the spatial [7, 14, 24] and temporal [26] domains.

3.4 Previous methods in the spacetime stereo framework

As we have seen, most previous triangulation systems can be thought of as operating either in the purely-spatial or purely-temporal domains. Recently, however, researchers have begun to investigate structured light systems that make use of both space and time, though typically with many restrictions. One such system uses primarily temporal coding, adding a small spatial window to consider stripe *boundaries* (i.e., adjacent pairs of stripes) [17, 27]. Another approach uses a primarily spatial coding, adding a small temporal window to better locate stripes [36]. Still another approach considers "tilted" space-time windows that have extent in both space and time, but are only a single pixel thick [13].

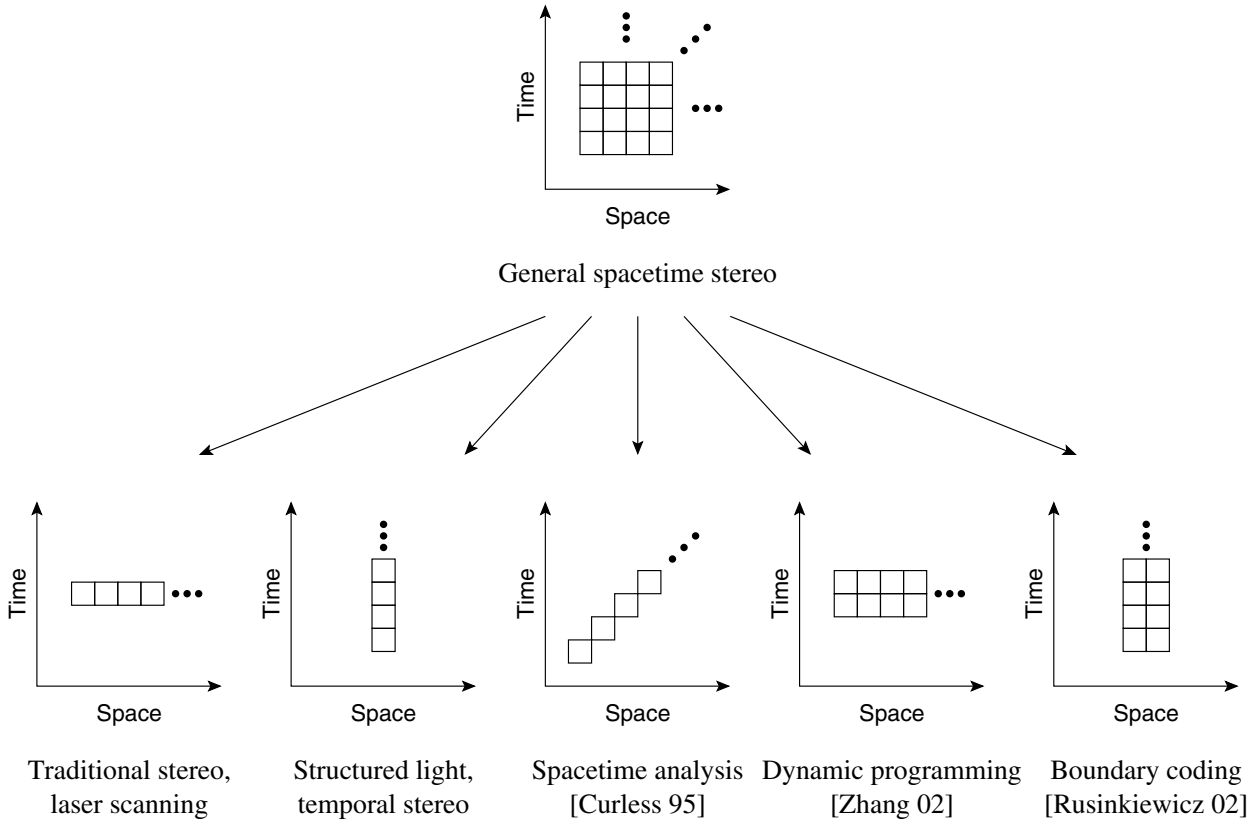

Figure 8: *Previous triangulation methods can be considered to be special cases of the framework of space-time stereo. Most methods use purely-spatial or purely-temporal windows, while a few others use other, restricted, classes of window shapes.*

Thus, as shown in Figure 8, some previous methods have begun to explore the benefits of windows that are not purely spatial or temporal. In the remainder of this paper, however, we argue that even these methods were limited in the class of matching windows they considered, and that expanding the domain of methods to encompass arbitrary space-time windows leads to improvements in robustness and flexibility.

4 Applications

The chief benefit of the spacetime stereo framework is its ability to suggest new depth acquisition methods that (a) incorporate ideas from many different systems that have traditionally been considered separately, and (b) optimize the size and shape of matching windows for the continuity,

texture, and motion characteristics of the scene. In this section, we first propose elements of previous systems that could be applied to wider classes of scanners, then consider specific new designs that are expected to provide better performance than existing systems.

4.1 Broader application of ideas from previous systems

Several of the research threads considered above, such as stereo vision or laser-stripe scanning, have developed particular methods for obtaining greater accuracy or robustness. Although these methods have often been applied only within the community that invented them, they could have a large impact on broader classes of spacetime stereo systems.

Regularization: Because of the difficulty of obtaining correspondences in traditional passive stereo, sophisticated methods have been developed that employ regularization to trade off resolution for robustness. The methods range from simple smoothing and outlier detection to complex techniques that filter data while preserving discontinuities [10]. Other spacetime stereo applications can benefit from these types of methods, particularly those that have traditionally been sensitive to ambiguity during matching (e.g. active methods for moving objects).

Ordering constraints: One frequently-encountered special case of the regularization paradigm relates to ordering constraints. That is, for the most part one expects two objects to appear in the same left-to-right order as seen from multiple viewpoints. Unfortunately, because of depth variation, the ordering constraint is not absolute. Therefore, algorithms must be used that permit but discourage ordering constraint violations. Though most common in stereo systems, ordering constraints could be incorporated into any triangulation-based scanner.

Data-sensitive windows: Many systems use correlation windows that are not axis aligned and perfectly rectangular. Rather, they follow the data to attempt to incorporate as much information as possible while not crossing discontinuities. For example, the non-linear diffusion approach of Scharstein and Szeliski effectively grows the matching windows independently at each pixel, as long as doing so reduces uncertainty [29]. The stripe boundary coding approach described earlier actually tracks boundaries as they move from frame to frame, leading to substantially better

performance for moving scenes than fixed windows [17]. These and other types of data-adaptive matching can improve almost any triangulation algorithm. In fact, data-sensitive algorithms can be thought of as a generalization, performed at run time, of the kind of window size analysis described in Section 5.

4.2 New scanner designs

We now discuss a number of specific possible extensions and improvements to existing methods. While this is intended primarily as a thought exercise to illustrate the utility of the spacetime stereo framework, section 5 will present results from implementations of the first two applications, as proof of the practical utility of the spacetime stereo framework.

Static scenes under variable illumination: Consider a static scene in the presence of uncontrolled but variable illumination, i.e., unstructured light. Existing methods do not make use of all available information to recover scene geometry in this case. Traditional stereo makes good use of any naturally occurring spatial features, but ignores the temporal domain. Active methods such as laser scanning and structured light use temporal information, but require very specific and highly calibrated features in order to determine correspondence. It is possible to design a hybrid of these methods that makes use of both naturally occurring features *and* the temporal domain.

In section 5, we analyze scenes of this class to discover the optimal spatio-temporal matching window. Based on this analysis, we show results on a few sample scenes. One potential application of spatiotemporal stereo is to large objects like buildings, cities, and mountains that are beyond the reach of existing active lighting methods, but often have naturally occurring variable illumination in the form of sunlight (with additional variation added by moving clouds).

Spacetime stereo for moving objects: The depth of moving objects has usually been recovered using spatial stereo. The primary reason for this is the simplicity of treating each time instant individually. However, as discussed previously, it is meaningful and potentially beneficial to apply temporal matching, even for scenes with moving objects.

The optimal spacetime matching window depends on the speeds with which objects in the scene move. For static scenes, a long temporal window will give optimal results. For scenes with

quickly moving objects, a short temporal window is desirable to avoid the distortions shown in figure 3. When objects move at intermediate speed, a spacetime matching window with extent in both space and time is optimal. Section 5 both analyzes scenes of this class and shows examples of recovered motion.

Structured light with no precision projectors: Structured light systems typically make use of precise time-varying lighting patterns. A relatively expensive projector, synchronized with the camera, is required to produce calibrated time varying patterns. In contrast, "active" stereo seeks to enhance spatial stereo matching using a static pattern projected by an inexpensive slide projector. For these systems, however, the ambiguities and limitations of exclusively-spatial matching still apply.

Using the framework of spacetime stereo, the strengths of these methods can be combined. For example, an imprecise but time varying illumination pattern can be created using an inexpensive motor to rotate a pattern in front of a light source. Since the projected light is no longer known, at least two real cameras are required. High-quality depth is recovered not by establishing correspondence in the spatial domain as in active stereo, but rather by correlating the temporal variation seen by the two cameras.

Laser scanning with multiple stripes: Laser scanning systems have traditionally provided the highest quality models; however, they have relatively slow acquisition times. Researchers have attempted to increase the rate at which models are acquired by sweeping the laser quickly and using a high speed camera [21], but achieving high speeds requires expensive customized hardware. Another approach is to add additional laser stripes [19]. Unfortunately, additional stripes introduce potential ambiguities in determining correspondence. This has typically been addressed by making surface continuity assumptions. Spacetime stereo allows these continuity assumptions to be relaxed by introducing disambiguating information from the time domain, so that a wider range of objects can be recovered by fast scanning systems.

Laser scanning of somewhat specular objects: One difficulty in traditional laser scanning is with regard to specular objects. The laser stripe tends to reflect and create additional illumination on other parts of the surface, essentially creating multiple laser stripes. These interreflections

make stripe peak detection ambiguous and interfere with proper reconstruction. As before, this situation can be improved by using a second real camera. In the case of a temporal matching vector, the spurious laser stripes will then simply create additional information in the time domain, and reconstruction will not be seriously compromised. Of course, if the object is sufficiently specular, then view dependent effects will become predominant, and performance will degrade; however, it should be possible to obtain additional robustness for many objects that exhibit some specularity but are primarily Lambertian.

5 Results

The spacetime stereo framework naturally gives rise to the question of optimal spatial-temporal window size. The best spacetime window will be scene and lighting dependent; however specific data sets and classes of scenes can be analyzed in terms of relative error.

We have chosen to investigate two classes of scenes corresponding to the first two potential applications in the previous section. First, we will consider scenes in which geometry is static but illumination varies in an unstructured manner. Second, we will look at scenes with moving objects.

5.1 Static Scenes

The first class of scenes we investigated includes static objects illuminated by unstructured but variable lighting. We choose this class because it includes scenes for which existing methods usually perform poorly. Consider the case of textureless geometry lit by uncontrolled natural illumination, such as sunlight. Traditional stereo methods will often not be able to recover any depth information in the textureless areas. On the other hand, active methods are not usually applicable since the illumination does not include the carefully controlled lighting on which they depend.

By analyzing error across the full range of possible spacetime window sizes, we can select the best parameters for reconstructing scenes in this class, which in this case turns out to be purely temporal processing or *temporal stereo*. Based on our analysis, we present visual results showing that spacetime stereo is capable of recovering depth with greater accuracy than traditional spatial-only analysis.

Figure 9: *Sample stereo pairs for the two scenes used in our experiments. Note the specularities on the cat sculpture and the regions of uniform texture on the wood blocks, both of which make traditional spatial stereo matching difficult.*

Experimental setup: We used two scenes to evaluate our method, pictured in figure 9. One consists of blocks of wood, while the other contains a sculpture of a cat and a teapot. Stereo pairs were acquired using a single camcorder and mirrors to produce two viewpoints. The working volume is approximately $50cm^3$, and the viewpoints have a baseline separation of approximately 60 degrees. Each viewpoint was manually calibrated using a target.

In addition, we have experimented with a variety of different lighting configurations. Figure 10 shows the cat scene under three unstructured lighting conditions—moving a flashlight manually across the objects, moving a hand in front of a light source to cast a sequence of shadows, and using a hand-held laser pointer to illuminate the scene with a moving line. On the left, we show one frame from each temporal sequence. On the right, we show the temporal matching vectors for a single pixel.

Note that the three lighting scenarios give rise to very different patterns of intensity variation. On top, the flashlight provides broad smooth lighting, and the pixel intensity varies relatively smoothly over time. In the middle, the pixel intensity diagram is binary or bi-level, with sharp transitions corresponding to when shadows appear or disappear. In the bottom row, the laser is narrow and much brighter than the ambient lighting, so the pixel intensity is mostly dark, with a few peaks corresponding to laser illumination. It remains a subject of future work to investigate the specific advantages and disadvantages of various illumination variations, and how these may

Figure 10: *Stereo pairs for the cat-and-teapot scene under different kinds of lighting variations. On the left is one frame in the image sequence. On the right is the variation in intensity over time plotted for one pixel. Top: manually moving a flashlight around the scene. Middle: moving a hand in front of a light source to cast a sequence of shadows. Bottom: using a hand-held laser pointer to illuminate the objects. Note the different characteristics of the temporal matching vector in each case.*

be combined optimally. In this paper, we merely demonstrate that we are able to produce good reconstructions using spacetime stereo, under a variety of illumination conditions.

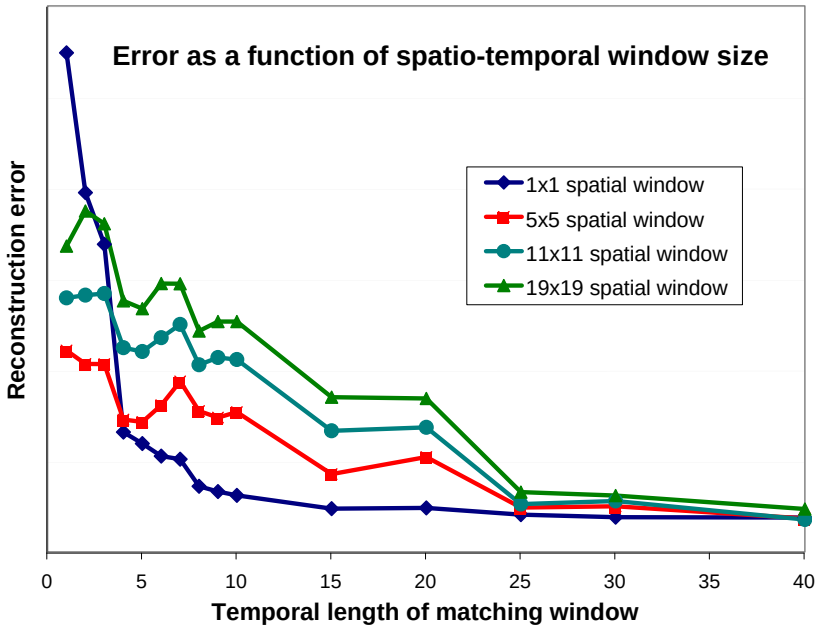

Figure 11: *Error as a function of spatio-temporal window size for the wood-block scene illuminated with a flashlight.*

Spatiotemporal matching: In order to characterize the performance of spacetime stereo, we choose a single data set and investigate all possible spatio-temporal window sizes. In this section we present results of our analysis of the sequence in which wooden blocks are illuminated by a flashlight.

For each spacetime window we computed the average depth error. Since ground truth is unavailable, we approximate "truth" as the visually estimated best result obtained from processing our other data sets of the same scene. Error is computed as the mean absolute Euclidean distance between a given test reconstruction and "ground truth." The temporal order of frames in the video sequence was randomly shuffled to negate any effects caused by the specific path of flashlight motion. This also has the effect of increasing the temporal information available in short temporal windows, since it removes correlation between neighboring frames.

In figure 11, we show the accuracy of reconstruction as a function of both spatial and temporal window size. For all spatial window sizes, we can see that increasing temporal window length is beneficial. Since the examined dataset is of a static scene, this result confirms our expectations. There are no adverse effects from increasing the temporal length, and new information becomes available that increases the probability of finding the correct match. Another insight, confirmed by

the graph, is that after only a few frames of temporal information become available, it is no longer desirable to use any spatial extent at all: the lowest error was obtained using a spatial window of only a single pixel. This corresponds to the fact that spatial windows behave poorly near depth discontinuities.

For clarity, only four data sets were shown. Similar results were obtained in additional tests of six other spatial window sizes. Furthermore, since a 1x1 spatial window produced the best results, we verified that error continues to decrease as the temporal window grows to span the entire sequence.

Although an analysis of only one sequence is shown, we believe that the conclusions generalize to similar scenes. In particular, with static scene geometry and variable illumination it is desirable to use a purely temporal matching vector.

Comparison of Spatial and Temporal matching: To show the utility of the spacetime stereo framework, we use our conclusions from the preceding analysis and compare purely spatial matching, as in standard stereo, with purely temporal matching. Spatial matching is computed using a 13×13 window; results were visually similar for other spatial window sizes. Temporal matching uses a single pixel, with a time neighborhood including the entire temporal sequence, as per equation 2. A hand-drawn mask is used to limit comparison to regions that are visible from both viewpoints.

We first consider the same sequence, in which wood blocks are illuminated with a flashlight. The top of figure 12 compares spatial matching (left), with temporal matching (right). Spatial stereo matching is unreliable because the wooden blocks have large regions of almost uniform texture. Hence, the results are uneven and noisy. On the other hand, lighting variation creates texture in the time domain, making temporal matching robust. To show that our results generalize to a variety of conditions, we repeated the experiment using different geometry and lighting. The bottom of figure 12 contains a comparison of spatial and temporal processing on the sequence in which a sculpted cat is subjected to shadowing. The results are similar: temporal matching produces better results than spatial matching.

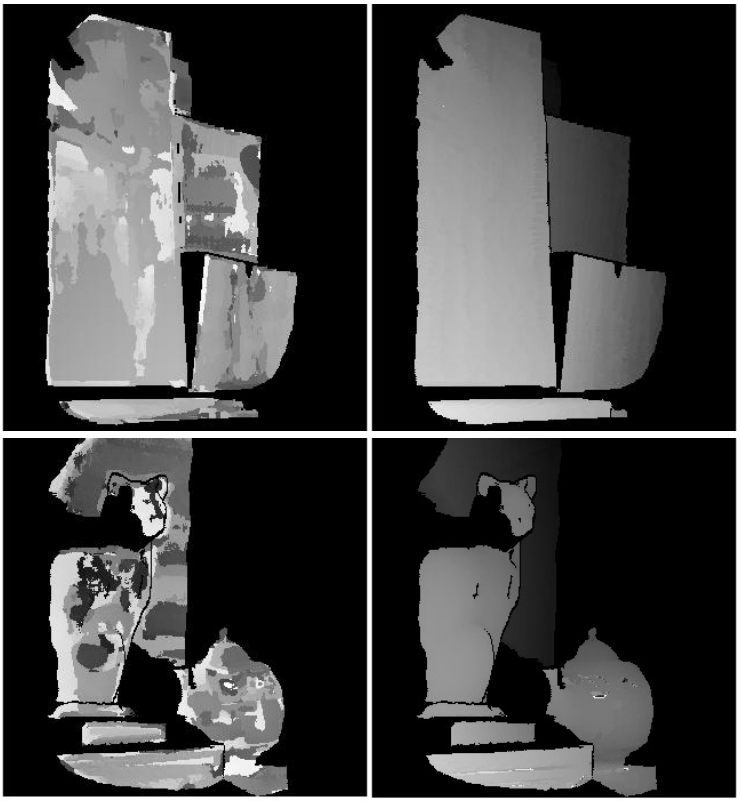

Figure 12: *Depth reconstruction (shading corresponds to estimated depth) using spatial stereo matching with* 13×13 *neighborhoods (left), and temporal stereo (right). On top are the wooden blocks with lighting variation by manually moving a flashlight. Below is the cat and teapot scene with lighting variation from shadows. Note that traditional spatial stereo depth estimates are uneven and noisy while temporal stereo is relatively robust and accurate.*

5.2 Moving Scenes

For scenes with motion, a different selection of spacetime window size is likely optimal. Under these conditions there is a tradeoff in the temporal domain between obtaining additional information and introducing confounding distortions. We expect U-shaped error curves, in which accuracy first improves and then decays as the temporal window size increases.

Experimental Setup: Moving objects require significantly higher-frequency (but still uncontrolled) lighting variation than do static objects. In order to accommodate this need we revised our experimental arrangement. A pair of cameras with a triangulation angle of approximately 15

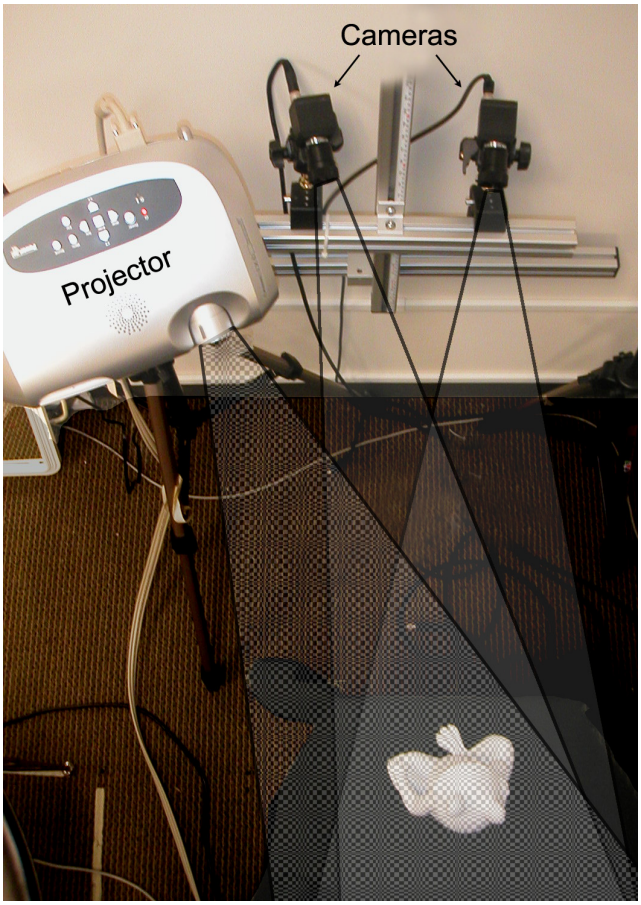

Figure 13: *Experimental setup. Two synchronized cameras capture stereo views at 40Hz, while the projector displays random high frequency patterns at 60Hz.*

degrees are arranged to observe a working volume of $30cm^3$. Instead of using a hand-held light source, an LCD projector is placed outside the camera baseline, but as nearby as is feasible, as shown in figure 13. As before, the cameras are calibrated and synchronized with respect to one another, but the light source is completely uncalibrated. Since the projected image can be varied at 60Hz, arbitrary high frequency lighting variation is possible. We simply project random patterns of stripes onto the scene. Our cameras are capable of capturing at approximately 40Hz. Figure 14 shows a captured stereo pair.

In order to evaluate the optimal window size when objects are moving, it is necessary to obtain ground truth data. Since this is not possible while an object actually is moving, we created "moving" data sets using stop motion photography. The frog statue was moved by hand under both linear and rotational motion, and a single image was taken at each position. When combined

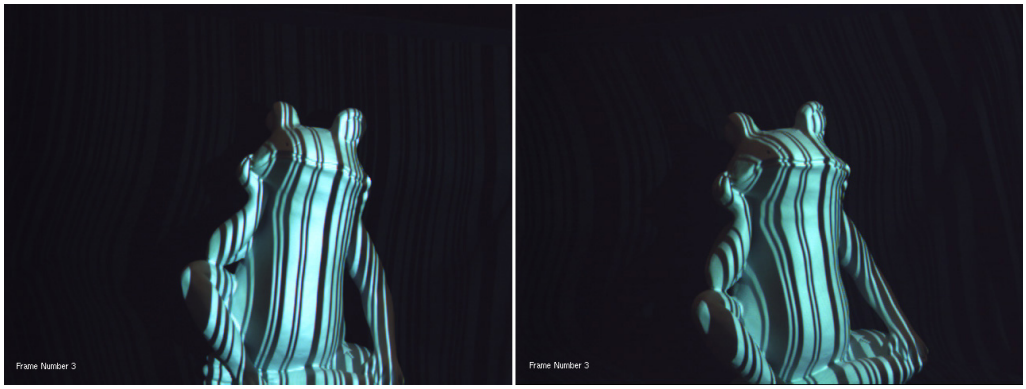

Figure 14: *An example of a stereo pair captured using the moving-object rig shown in Figure 13.*

these images simulate actual object motion. In order to obtain ground truth for a given frame, the frog was left stationary while additional lighting variation was projected and recorded. Figure 15 shows the high quality reconstruction that is possible using temporal stereo when the object is not moving.

Spatiotemporal matching: For each moving sequence, depth was computed using all possible combinations of spatiotemporal window sizes and compared to ground truth. Since depth recovery of moving scenes is more error prone than is that of static scenes, we use a measure of robustness rather than L2 norm to evaluate error. For each window size, robustness is computed as the number of pixels for which the computed and ground truth disparity differ by at most 1 pixel. Computation is limited to those pixels for which ground truth disparity exists.

In the first condition, the frog was moved along a linear path at the rate of 1mm per frame. Figure 16 shows the robustness of various window sizes. As expected, since the object is in motion, it is no longer preferable to use a very large temporal window. Disambiguating information must come from somewhere, and since the temporal window is smaller, a single-pixel spatial window no longer provides good results. In this case, we found a 9x9x3 spatiotemporal window to be optimal. We also computed the optimum window size when the frog was subjected to rotation. When we used a rotation speed of 0.3 degrees per frame, the optimal temporal window size was 8 frames, and spatial window size 3x3. Figure 17 shows the robustness under this condition.

If we increase the rotation speed by an order of magnitude to 3.0 degrees per frame (a relatively very high rate of rotation), the optimal temporal window size becomes very short, reducing to 2 frames. In this extreme case, object motion is so large that it is essentially best to treat each frame

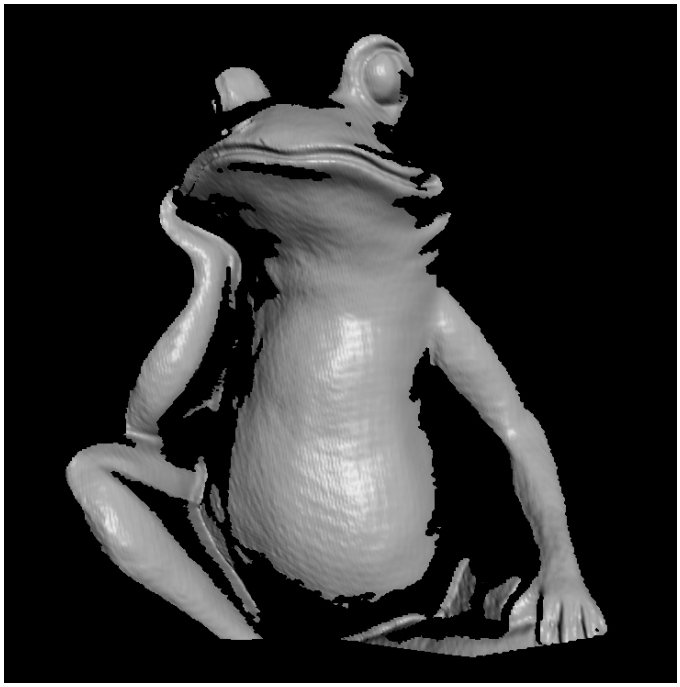

Figure 15: *Estimated depth using temporal stereo when the object is static*

separately with spatial stereo. This is analogous to the opposite extreme of static objects, where purely temporal stereo is the optimal configuration. However, as seen earlier, in many cases, a spatiotemporal window provides the best depth information.

The optimal window size is a function of the speed of object motion, the camera frame rate, the spatial texture available and the rate of temporal lighting variation. While it would be difficult to quantify the exact window size that should be used under every condition, it is possible to make a few qualitative statements. Either slow moving objects or fast cameras allow a longer temporal window, while fast objects or slow cameras require a short temporal window. Both spatial and temporal texture is desirable, and it should have a frequency roughly equivalent to the sampling frequency along that dimension.

Capturing motion: In order to demonstrate the capability of spacetime stereo on dynamic scenes, we captured the motion of a deforming face. Rather than use stop motion photography as in the previous experiments, the cameras captured video at 40Hz, while the projector displayed stripe patterns at 60Hz. Depth was recovered at each frame of the sequence using a window size of 7x1x7. This window size was chosen because both the horizontal and temporal dimensions have

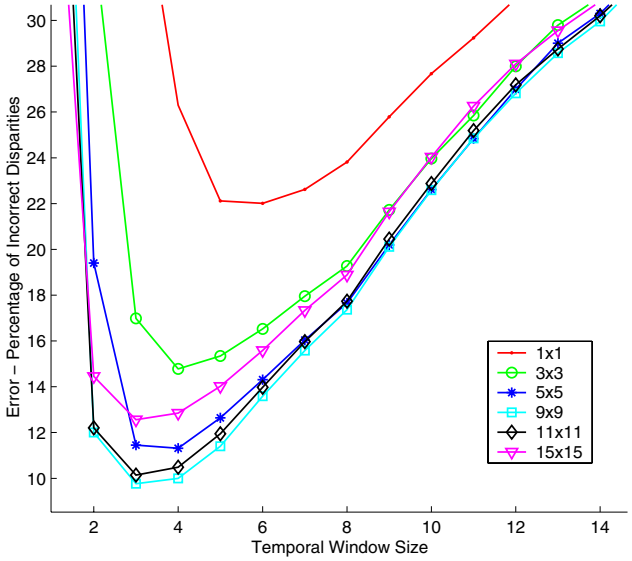

Figure 16: *Matching error for a linearly moving scene as a function of temporal window size, for a variety of spatial window sizes. The the result is a U-shaped curve for which the error first decreases with more disambiguating information, but then increases as motion makes matching difficult. Hence, a finite temporal window is desirable, and a 9x9x3 spacetime window is seen to provide best results in this case.*

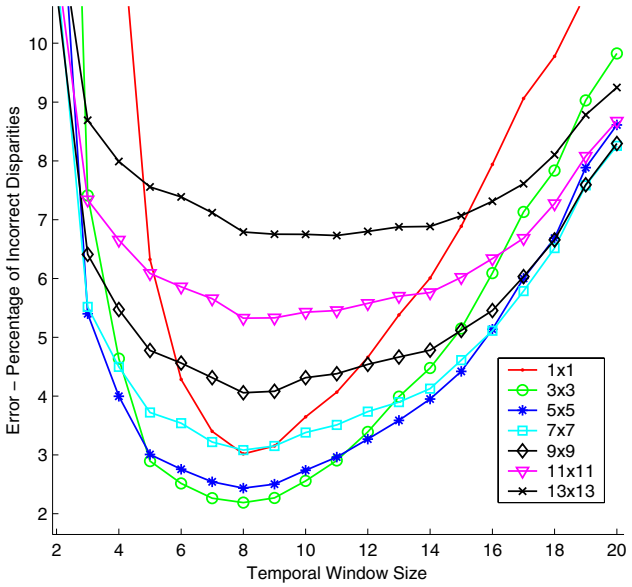

Figure 17: *Matching error for a rotating scene, as a function of temporal window size for several spatial window sizes. The result is a U-shaped curve similar to the linear motion case. In this case, a 3x3x8 spatiotemporal window is optimal, and is better than either spatial or temporal matching alone.*

high frequency texture that is useful for matching. The vertical dimension (which is aligned with our stripe pattern) has relatively little texture, so does not contribute substantially to matching.

Figure 18: *Depth estimates on a dynamic scene (one of the authors smiling), captured at 40 Hz. We show two views (frontal on top and side on bottom), for three instants in time. Note recovery of subtle features like the cheek deformation. Spacetime stereo provides a new way of recovering depth in dynamic scenes, which has been difficult for previous algorithms.*

The recovered depth was triangulated and is shown rendered with lighting in figure 18. Note in particular that the subtle motion of the cheek deformation while smiling is faithfully captured.

6 Conclusions and Future Work

This paper has introduced a new classification framework, spacetime stereo, for depth from triangulation. Rather than distinguish algorithms as active or passive, we classify algorithms based on the spatial or temporal domain in which they locate corresponding features. This classification unifies a number of existing techniques, such as stereo, structured light, and laser scanning into a continuum of possible solutions, rather than segmenting them into disjoint methods. From this unified view a number of possible extensions and hybrid methods emerge, potentially allowing for improved stereo recovery of moving scenes, structured light scanning with multiple simultaneous systems, faster and cheaper variants, and laser scanning of shiny objects.

As a demonstration of the utility of the spacetime stereo framework, we have analyzed the performance of various spatio-temporal matching windows on two classes of scenes—those with unstructured but variable illumination, and those with moving objects. Based on this analysis we have demonstrated depth recovery results that are superior to those obtainable using traditional spatial-only stereo. In future work, we wish to extend our analysis to determine optimal methods and patterns for generating variable lighting.

In summary, we believe the framework proposed in this paper provides a useful way of thinking about many triangulation-based depth extraction methods, and the insights from it will lead to new applications.

References

[1] K. Araki, Y. Sato, and S. Parthasarathy. High speed rangefinder. In *SPIE vol 850: Optics, Illumination, and Image Sensing for Machine Vision*, pages II–184–II–188, 1987.

[2] S. Baker, T. Sim, and T. Kanade. When is the shape of a scene unique given its light-field: A fundamental theorem of 3D vision? *PAMI*, 25(1), 2003.

[3] J. Batlle, E. Mouaddib, and J. Salvi. Recent progress in coded structured light as a technique to solve the correspondence problem: A survey. *Pattern Recognition*, 31(7):963–982, 1998.

[4] F. Bernardini, I. Martin, and H. Rushmeier. High-quality texture reconstruction from multiple scans. *IEEE TVCG*, 7(4), 2001.

[5] P. Besl. *Active Optical Range Imaging Sensors, in Advances in Machine Vision, chapter 1*, pages 1–63. 1989.

[6] R. Bolles, H. Baker, and D. Marimont. Epipolar-plane image analysis: An approach to determining structure from motion. *IJCV*, pages 7–56, 1987.

[7] N. Borghese, G. Ferrigno, G. Baroni, A. Pedotti, S. Ferrari, and R. Savare. Autoscan: A flexible and portable 3D scanner. *IEEE Computer Graphics and Applications*, 18(3):38–41, 1998.

[8] J. Bouguet and P. Perona. 3D photography on your desk. In *ICCV*, pages 43–50, 1998.

[9] K. L. Boyer and A. C. Kak. Color-encoded structured light for rapid active ranging. *Trans. PAMI*, 9(1), 1987.

[10] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *Trans. PAMI*, 23(11), November 2001.

[11] C. Chen, Y. Hung, C. Chiang, and J. Wu. Range data acquisition using color structured lighting and stereo vision. *Image and Vision Computing*, 15(6):445–456, June 1997.

[12] B. Curless. Overview of active vision techniques. In *SIGGRAPH 99 Course on 3D Photography*, 1999.

[13] B. Curless and M. Levoy. Better optical triangulation through spacetime analysis. In *ICCV*, pages 987–994, 1995.

[14] J. Davis and X. Chen. A laser range scanner designed for minimum calibration complexity. In *Third International Conference on 3D Digital Imaging and Modeling*, 2001.

[15] J. Davis, R. Ramamoorthi, and S. Rusinkiewicz. Spacetime stereo: A unifying framework for depth from triangulation. In *CVPR*, pages II–359–II–366, 2003.

[16] U. Dhond and J. Aggarwal. Structure from stereo—a review. *IEEE Transactions on Systems, Man and Cybernetics*, 19(6), 1989.

[17] O. Hall-Holt and S. Rusinkiewicz. Stripe boundary codes for real-time structured-light range scanning of moving objects. In *ICCV*, pages 359–366, 2001.

[18] S. Inokuchi, K. Sato, and F. Matsuda. Range-imaging for 3D object recognition. In *ICPR*, pages 806–808, 1984.

[19] J. Jalkio, R. Kim, and S. Case. Three dimensional inspection using multistripe structured light. *Optical Engineering*, 24(6):966–974, 1985.

[20] R. Jarvis. A perspective on range finding techniques for computer vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(2):122–139, 1983.

[21] T. Kanade, A. Gruss, and L. Carley. A very fast VLSI rangefinder. In *IEEE International Conference on Robotics and Automation*, pages 1322–1329, 1991.

[22] S. Kang, J. Webb, C. Zitnick, and T. Kanade. A multibaseline stereo system with active illumination and real-time image acquisition. In *ICCV*, pages 88–93, 1995.

[23] A. Koschan, V. Rodehorst, and K. Spiller. Color stereo vision using hierarchical block matching and active color illumination. In *ICPR*, pages I 835–839, 1996.

[24] G. Medioni and J. Jezouin. An implementation of an active stereo range finder. In *Optical Society of America Technical Digest Series vol. 12, Tropical Meeting on Machine Vision*, pages 34–51, 1987.

[25] D. Poussart and D. Laurendeau. *3-D Sensing for Industrial Computer Vision, in Advances in Machine Vision, chapter 3*, pages 122–159. 1989.

[26] K. Pulli, H. Abi-Rached, T. Duchamp, L. Shapiro, and W. Stuetzle. Acquisition and visualization of colored 3D objects. In *ICPR*, pages 11–15, 1998.

[27] S. Rusinkiewicz, O. Hall-Holt, and M. Levoy. Real-time 3D model acquisition. *ACM Trans. on Graphics (SIGGRAPH 2002 proceedings)*, 21(3):438–446, 2002.

[28] P. Saint-Marc, J. Jezouin, and G. Medioni. A versatile PC-based range finding system. *IEEE Transactions on Robotics and Automation*, 7(2):250–256, 1991.

[29] D. Scharstein and R. Szeliski. Stereo matching with non-linear diffusion. In *CVPR*, 1996.

[30] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 47(1):7–42, 2002.

[31] D. Scharstein and R. Szeliski. High-accuracy stereo depth maps using structured light. In *CVPR*, 2003.

[32] S. Seitz. The space of all stereo images. In *ICCV*, pages 26–33, 2001.

[33] E. Shechtman, Y. Caspi, and M. Irani. Increasing space-time resolution in video. In *ECCV*, 2002.

[34] T. C. Strand. Optical three-dimensional sensing for machine vision. *Optical Engineering*, 24(1):33–40, 1985.

[35] R. Woodham. Photometric method for determining surface orientation from multiple images. *Optical Engineering*, 19(1):139–144, 1980.

[36] L. Zhang, B. Curless, and S. Seitz. Rapid shape acquisition using color structured light and multi-pass dynamic programming. In *IEEE 3D Data Processing Visualization and Transmission*, 2002.

[37] L. Zhang, B. Curless, and S. Seitz. Spacetime stereo: Shape recovery for dynamic scenes. In *CVPR*, 2003.

[38] T. Zickler, P. Belhumeur, and D. Kriegman. Helmholtz stereopsis: Exploiting reciprocity for surface reconstruction. In *ECCV*, pages III 869–884, 2002.