Spacetime Stereo: A Unifying Framework for Depth from Triangulation

James Davis, *Member*, *IEEE*, Diego Nehab,
Ravi Ramamoorthi, and
Szymon Rusinkiewicz, *Member*,
IEEE Computer Society

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. To show the practical utility of the framework, we develop two new algorithms for depth estimation: depth from unstructured illumination change 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.

Index Terms—Depth from triangulation, stereo, spacetime stereo.

1 Introduction

This paper considers methods that obtain depth via triangulation. Within this general family, a number of methods have been proposed including stereo [15], [28], laser stripe scanning [3], [12], [13], [18], and time or color-coded structured light [2], [8], [16], [17], [29]. Although a deep relationship exists between these methods, as illustrated in the classification of Fig. 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 some of the traditional limitations within each subspace become apparent.

Most previous surveys classify triangulation techniques into *active* and *passive* methods [3], [11], [24], [31]. Active techniques, such as laser scanning and structured light, intentionally project illumination into the scene in order to construct easily identifiable features and 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 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 insight that triangulation methods can be unified into a single framework is the primary contribution of this work. The success of a proposed framework can be measured by its simplicity and its power to bring new insights. We believe that this framework is sufficiently simple that most readers will find it intuitive and almost obvious in retrospect. To illustrate the framework's power to provide insight, we introduce two new methods for recovering depth that have not been previously explored in the literature.

The first new method applies temporal processing to scenes 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 handheld flashlight. The second new method applies spacetime processing to scenes in which the object moves. In addition to demonstrating the method, we analyze the necessity of spacetime processing, and show that optimal reconstruction is possible only by simultaneously using both the space and time domains.

This paper is a considerably expanded version of a previous conference paper [14] 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 [33]. 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 [30].

2 Spacetime Stereo

In this section, we introduce our spacetime stereo framework for characterizing depth-from-triangulation algorithms.

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 no reason to restrict the matching vector to lie entirely in a single spatial image plane. By considering multiple frames across time, we can extend the matching window into the temporal domain, as shown in Fig. 2. In general, 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

- J. Davis is with Honda Research Institute, USA, 800 California Street, Suite 300, Mountain View, CA 94041. E-mail: jedavis@ieee.org.
- D. Nehab and S. Rusinkiewicz are with the Department of Computer Science, Princeton University, 35 Olden Street, Princeton, NJ 08544. E-mail: {diego, smr}@cs.princeton.edu.
- R. Ramamoorthi is with the Department of Computer Science, Columbia University, 450 Computer Science Bldg., 500 W. 120 Street, New York, NY 10027. E-mail: ravir@cs.columbia.edu.

Manuscript received 16 Oct. 2003; revised 9 July 2004; accepted 6 Aug. 2004. Recommended for acceptance by L. Quan.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0321-1003.

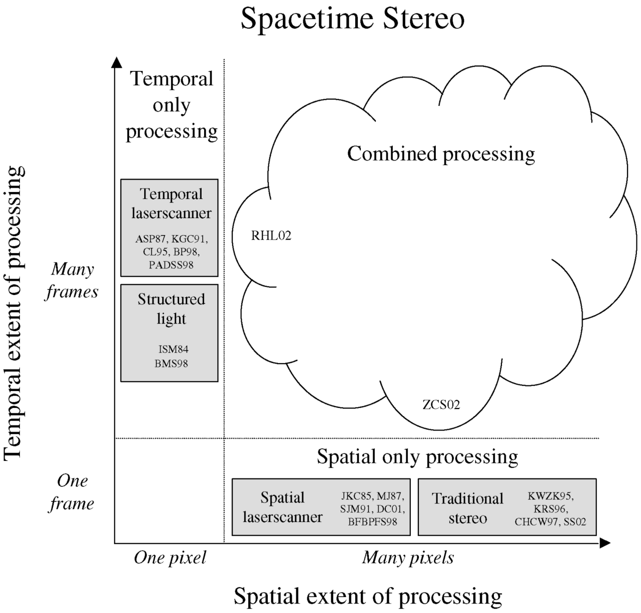

Fig. 1. Most existing depth from triangulation techniques are specific instances of the more general class of spacetime stereo reconstruction. Because these methods have been developed largely independently, they have often been artificially constrained to a small range of variation. Understanding that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

window and T is the dimension along the time axis. In this general case, we would seek to optimize the matching function,

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

It is clear that there is no mathematical distinction between the spatial and temporal axes. By choosing $T = 1$, we reduce to traditional spatial-only stereo matching. By choosing $N = M = 1$, we use a purely temporal matching window. 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 general, the precise lighting and scene characteristics will determine the optimal size for the spacetime matching window.

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.

Stereo. Traditional stereo matching is a well-studied problem in computer vision. A number of good surveys exist [15], [28]. As discussed in Section 2, traditional stereo matches vectors in the spatial or image domain to determine correspondence.

Surprisingly, no existing stereo methods make use of the temporal domain. Presumably this is due to the ubiquitous classification of techniques into passive and active. Passive techniques are assumed to have no lighting variation and, thus, no need for temporal processing. The framework and examples presented in this paper make clear that it is beneficial to extend existing stereo algorithms to use this additional source of information.

It should be noted that although epipolar analysis includes the language of "temporal" imaging, that work encodes camera motion on the temporal axis and is thus more closely related to multibaseline stereo processing [5].

Most stereo methods can be described as a pipeline of local matching followed by global regularization. Since the ambiguities of passive stereo provide poor quality local correspondence, nearly all state of the art stereo research focuses on methods for global regularization such as dynamic programming [22] or graph cuts [9]. In contrast, this paper focuses on improving the local operator used for matching, using absolutely no method of global regularization. Many of the global methods for improved matching in the context of traditional spatial windows could be easily extended to include spatiotemporal windows.

Methods for improving the local matching metric in stereo have also been proposed, such as adaptive windows [23] and robust matching metrics [4]. In this work, we use very simple matching in order to isolate the importance of using the spacetime domain. In particular, we use constant size rectangular windows and accept the disparity that minimizes the SSD as shown in (2). More sophisticated matching metrics will of course improve the results beyond those shown in this paper.

Some stereo implementations do make use of actively projected texture in order to aid the correspondence search. For example, Kang et al. project an uncalibrated sinusoidal pattern and reconstruct depth using a real-time multibaseline solution [20]. We group techniques such as this with traditional stereo, rather than with the coded structured light methods discussed in the next section, because the correspondence search is inherently spatial rather than temporal.

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. [2]. The projector illuminates a static scene with a temporally varying pattern of light. The patterns are arranged such that every projected column of pixels can be uniquely identified. Thus, the depth at each camera pixel is uniquely determined based on the particular pattern observed.

These systems rely on strictly controlled lighting and most existing implementations are very careful to synchronize projectors, use one system at a time, and remove ambient illumination. The work presented in this paper unifies structured light methods with stereo matching and, thus, eliminates the need for precise control over all aspects of scene lighting.

Although depth recovery in these systems is not typically described in terms of stereo matching, they do fall within the

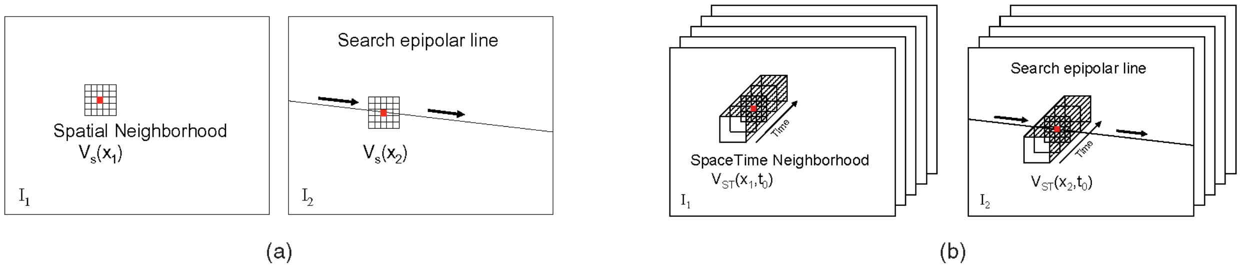

Fig. 2. (a) Comparison of spatial and (b) spacetime stereo. In spatial stereo, the epipolar line is searched for similar spatial neighborhoods. In spacetime stereo, the search is for similar spatio-temporal variation.

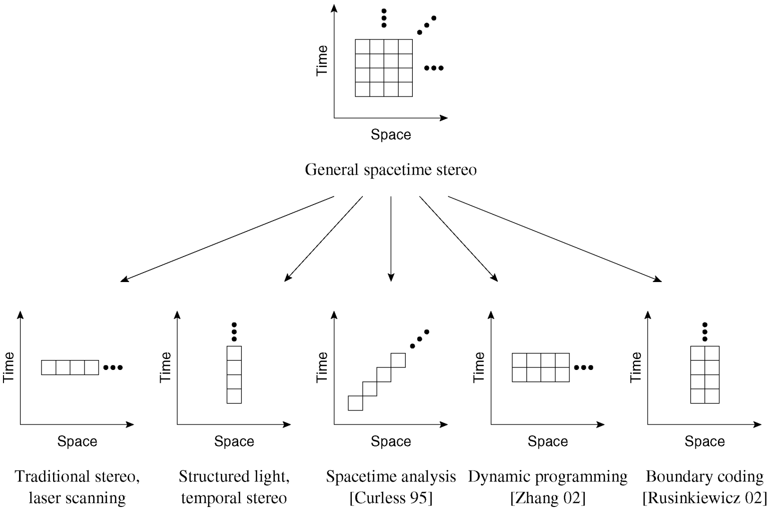
Time
Space
General spacetime stereo
Traditional stereo, laser scanning
Structured light, temporal stereo
Spacetime analysis [Curless 95]
Dynamic programming [Zhang 02]
Boundary coding [Rusinkiewicz 02]

Fig. 3. Previous triangulation methods can be considered to be special cases of the framework of spacetime stereo. Most methods use purely-spatial or purely-temporal matching windows, while a few others use other, restricted, classes of window shapes.

spacetime framework. The camera matching vector is purely temporal and is matched against a known database of projected patterns and their associated depths. The matching error metric can be written as

$$\left\| I_1(V_t(x_1, t_0)) - P_2(V_t(x_2, t_0)) \right\|^2, \tag{3}$$

which is similar to (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 [10], and generating high-quality ground truth stereo test images [29].

Laser stripe scanning. Another alternative is laser scanning. 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. Informative surveys have been provided by Besl [3] and Jarvis [18].

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 [27]. 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.

Laser scanners that function in the temporal domain have also been built [1], [19], [7]. As the laser sweeps past each pixel, the time at which the peak intensity is observed is recorded and used to establish correspondence. Curless and Levoy [12] provide an analysis of the benefits that temporal correlation provides over the traditional spatial approach in the context of 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.

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. Two-camera implementations have been built that find correspondence in both the spatial [6], [13], [21] and temporal [25] domains.

Partial spacetime methods. 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) [16], [26]. Another approach uses a primarily spatial coding, adding a small temporal window to better locate stripes [32]. Still another approach considers "tilted" space-time windows that have extent in both space and time, but are only a single pixel thick [12].

Thus, as shown in Fig. 3, some previous methods have begun to explore the benefits of windows that are not purely spatial or temporal. However these methods were limited in the class of matching windows they considered, and expanding the domain of methods to encompass arbitrary space-time windows leads to improvements in robustness and flexibility.

4 Depth from Unstructured Illumination Change

Consider the class of scenes which includes static objects illuminated by unstructured but variable lighting. This class includes scenes for which existing methods perform poorly, such as textureless geometry lit by uncontrolled natural illumination, such as sunlight. Traditional spatial stereo methods will not be able to recover any depth information in the textureless areas without resorting to global smoothness assumptions. On the other hand, active methods are not applicable since the illumination does not include the carefully controlled lighting on which they depend.

Spacetime stereo is able to recover high quality depth maps for this class of scenes. 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 turns out to be purely temporal processing or *temporal stereo*. To illustrate the gains from this analysis, we present visual results showing that temporal stereo is capable of recovering depth with far greater accuracy than traditional spatial-only analysis. The reader should

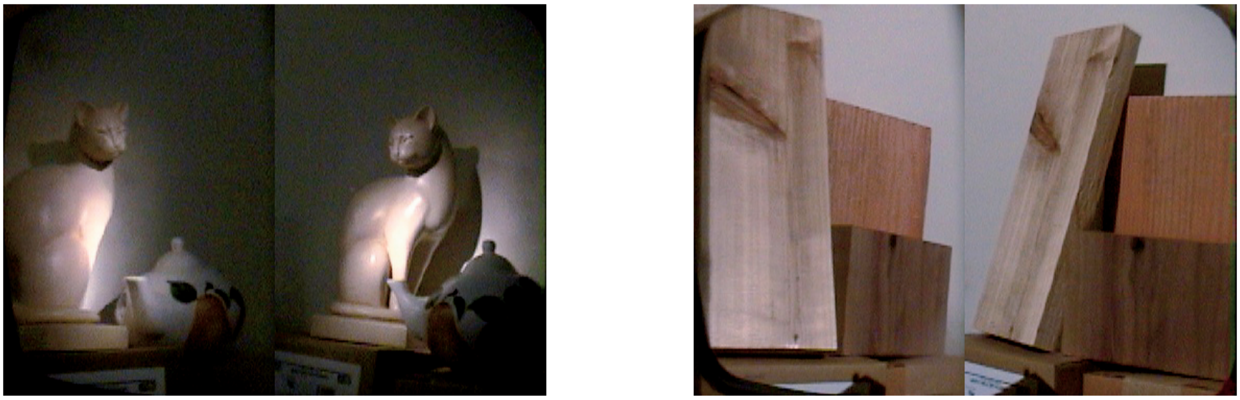

Fig. 4. Sample stereo pairs for the two scenes used in our experiments. The cat is illuminated by a flashlight which was moving slowly over the scene. Note the regions of uniform texture and lighting which make traditional spatial stereo matching difficult.

keep in mind that although temporal stereo is straightforward in light of the spacetime framework, it represents a truly new algorithm which has not been investigated previously.

Experimental setup. We used two scenes to evaluate our method, pictured in Fig. 4. 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.

We have experimented with a variety of different lighting configurations, 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. We have found that we are able to produce good reconstructions using spacetime stereo, under a variety of illumination conditions.

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."

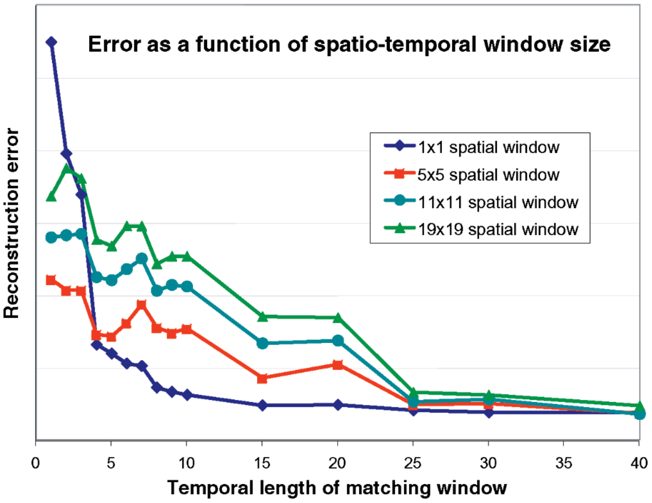

Fig. 5. Error as a function of spatio-temporal window size for the wood-block scene illuminated with a flashlight.

In Fig. 5, 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. 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 spatial window sizes are shown. Similar results were obtained in additional tests of six other spatial window sizes. Furthermore, we verified that error continues to decrease as the temporal window grows to span the entire sequence.

It should be noted that 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 has the effect of increasing the temporal information available in short temporal windows, since it removes correlation between neighboring frames. As a result, using a 1×1 spatial window becomes optimal after only four frames of temporal information are available. If we had not shuffled the frames, the number of frames required to outperform spatial stereo would have been higher, related to the speed at which the flashlight moves. The original sequences in this case had approximately 400 frames, and 50-100 frames would have been required to obtain a good approximation of depth.

Although an analysis of only one sequence is shown, we have recovered depth for hundreds of scenes and believe that the conclusions generalize. 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 practical 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 (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. Fig. 6 compares spatial matching, with temporal matching. 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

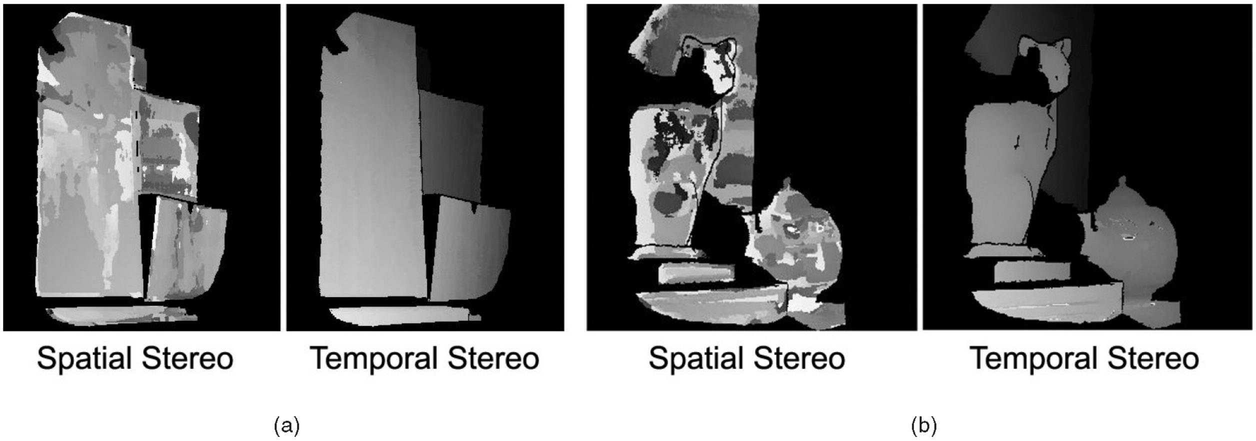

Fig. 6. Comparison of depth reconstruction (shading corresponds to estimated depth) using spatial stereo matching with 13×13 neighborhoods and temporal stereo. On the left (a) are the wooden blocks with lighting variation by manually moving a flashlight. On the right (b) is the cat and teapot scene with lighting variation from shadows. Note that traditional spatial stereo depth estimates are uneven and noisy while temporal stereo is relatively robust and accurate.

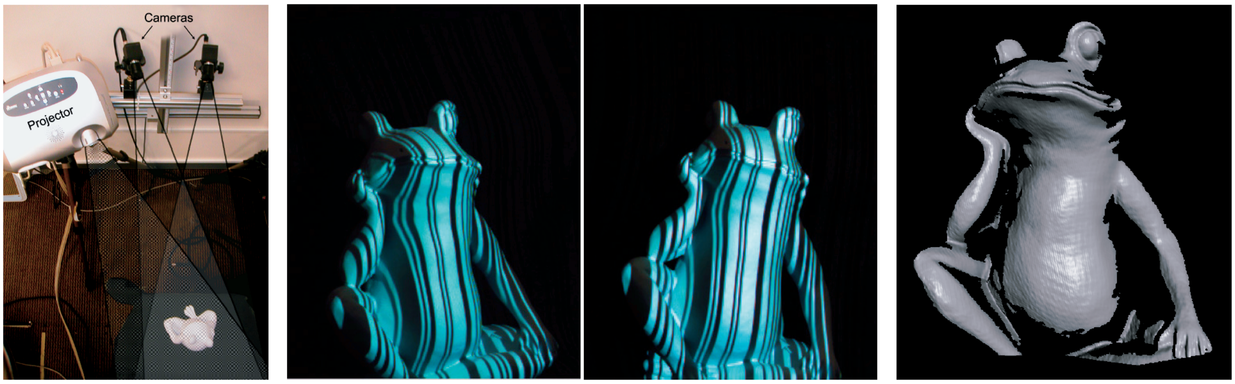

Fig. 7. Experimental setup. Two synchronized cameras capture stereo views at 40Hz, while the projector displays random high-frequency patterns at 60Hz.

variety of conditions, we repeated the experiment using different geometry and lighting; a sculpted cat was subjected to shadowing. The results are similar: temporal matching produces much better results than spatial matching.

The scene of a white cat in front of a white wall was designed to be difficult or impossible for spatial stereo. Nevertheless, some readers may be surprised that spatial stereo produces such poor depth estimates. We would like to reiterate that, in order to compare only the proposed changes to local matching, no global regularization was used in these experiments. The addition of smoothness constraints would presumably improve the recovered depth regardless of whether spatial or temporal matching was used.

5 Depth of Moving Scenes

Scenes with motion represent a new set of challenges. Traditional passive stereo can process each frame of a sequence, but produces relatively low-quality results. Active methods can not in general be applied, since nearly all rely on a static object. The spacetime stereo framework provides a solution. By subjecting the scene to high frequency illumination variations, a spacetime window can be used to recover depth. Although this is a straightforward application of the spacetime framework, it is unlikely that it would have been proposed by either the passive or active triangulation communities. The passive community would not propose active lighting, and the active community strictly controls lighting and does not speak in terms of stereo matching.

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 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 on the left in Fig. 7. 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. The middle of Fig. 7 shows a captured stereo pair. On the right is a reconstructed and rendered view of the object, captured while stationary. Note that although the lighting was unknown the resulting accuracy is equivalent to a laser scanner.

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 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.

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 the percentage of correct disparities as a measure of robustness rather than L2 norm to evaluate error. For each window size, robustness is computed as the percentage of pixels for which the computed and ground truth disparity differ by at most 1.0 pixels. Computation is limited to those pixels for which ground truth disparity exists.

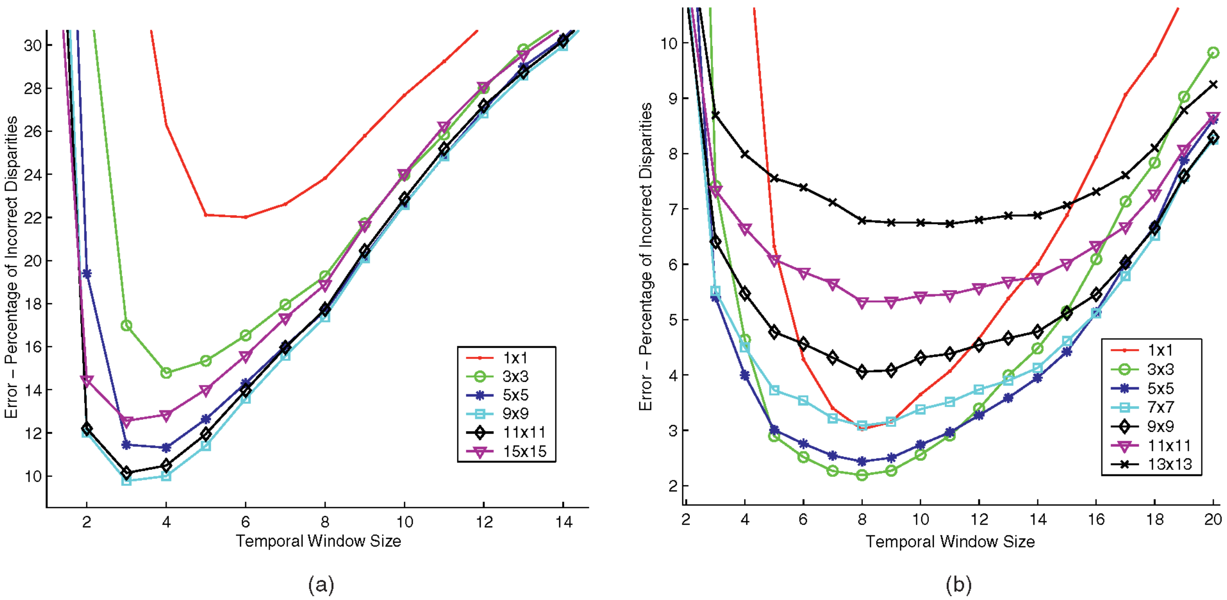

Fig. 8. (a) Matching error for a linearly moving scene as a function of temporal window size, for a variety of spatial window sizes. The result is a U-shaped curve for which the error first decreases with more disambiguating information, but then increases as motion makes matching difficult. Hence, a finite temporal window is desirable, and a $9 \times 9 \times 3$ spacetime window is seen to provide best results in this case. (b) 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 $3 \times 3 \times 8$ spatiotemporal window is optimal and is better than either spatial or temporal matching alone.

For scenes with motion there is a tradeoff in the temporal domain between obtaining additional information and introducing confounding distortions. If we repeat the analysis performed on static scenes, we expect U-shaped error curves, in which accuracy first improves and then decays as the temporal window size increases.

In the first condition, the frog was moved along a linear path at the rate of 1 mm per frame. This is equivalent to roughly three to four pixels of motion in the image. Fig. 8a 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 $9 \times 9 \times 3$ 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 3×3. Fig. 8b shows the robustness under this condition.

It is reasonable to wonder what would happen if the object moves either faster or slower. We increased the rotation speed by an order of magnitude to 3.0 degrees per frame (a relatively very high rate of rotation). Although the plot is not shown, the optimal temporal window size becomes very short, reducing to two frames. In this extreme case, object motion is so large that it is essentially best to treat each frame separately with spatial stereo.

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. Although it is true that projected lighting will improve any stereo algorithm, we have shown that for some scenes *optimal* reconstruction requires the use of a spatiotemporal window.

When the object moves either very quickly or very slowly a degenerate form of spacetime stereo is optimal. For fast scenes spatial-only stereo is desirable, while for static scenes temporal-only stereo is desirable. Both spatial and temporal texture is desirable, and this texture 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 real 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 $7 \times 1 \times 7$. This window size was chosen because both the horizontal and temporal dimensions have 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. This sequence can not be reconstructed reliably using either spatial-only matching or temporal-only matching. The recovered depth was triangulated and several frames are shown rendered with lighting in Fig. 9a. The complete video is available at http://graphics.stanford.edu/papers/SpacetimeStereo/.

Rendered images of polygonal models with lighting are sensitive to the mesh surface normal. Since we show data prior to regularization it will appear noisy even if the error has low magnitude. Fig. 9b visualizes the mesh after filling holes and smoothing the surface normals. These steps are analogous to the global regularization that is ubiquitously used in traditional stereo. Fig. 9c shows a plot of the mesh depth along the line indicated above. Note that the noise level is well below 1 mm.

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.

As a demonstration of the utility of the spacetime stereo framework, we introduce two new hybrid methods: depth from unstructured illumination, and shape recovery for dynamic scenes using spacetime windows. We have demonstrated depth recovery results that are superior to those obtainable using traditional spatial-only stereo in both cases. Further, we have analyzed the optimal spacetime windows and shown that for some classes of scenes space time windows must be used for optimal reconstruction.

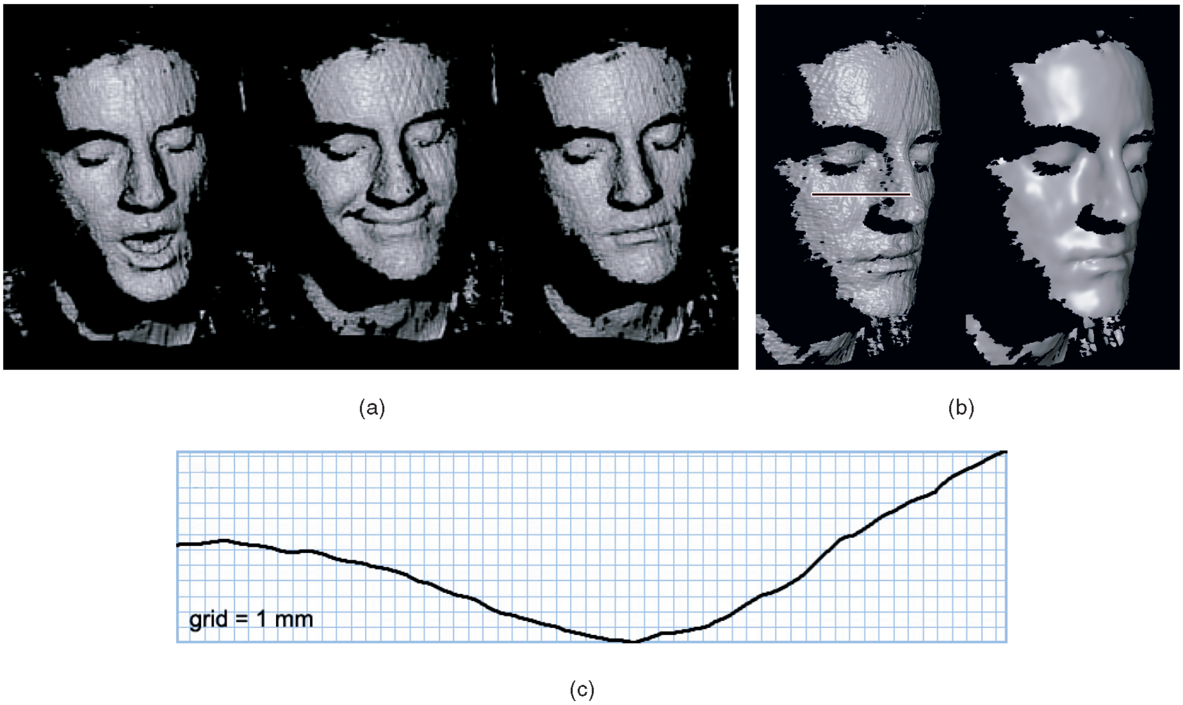

Fig. 9. (a) Depth estimates of three frames in a dynamic scene (one of the authors smiling), captured at 40 Hz. Note recovery of subtle features like the cheek deformation. (b) Recovered geometry before and after filling holes and smoothing mesh normals. (c) Plot of the mesh depth along the line indicated above. Note that although noisy normal estimates are perceptually distracting, the actual mesh geometry is accurate to under a millimeter, evident by visual inspection of the smoothness of this plot.

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,” *Proc. SPIE Optics, Illumination, and Image Sensing for Machine Vision,* vol. 850, pp. II-184-II-188, 1987.

[2] 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,* vol. 31, no. 7, pp. 963-982, 1998.

[3] P. Besl, *Active Optical Range Imaging Sensors, in Advances in Machine Vision,* chapter 1, pp. 1-63, 1989.

[4] M.J. Black and P. Anandan, “A Framework for the Robust Estimation of Optical Flow,” *Proc. Fourth Int'l Conf. Computer Vision,* pp. 231-236, 1993.

[5] R. Bolles, H. Baker, and D. Marimont, “Epipolar-Plane Image Analysis: An Approach to Determining Structure from Motion,” *Int'l J. Computer Vision,* pp. 7-56, 1987.

[6] 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,* vol. 18, no. 3, pp. 38-41, 1998.

[7] J. Bouguet and P. Perona, “3D Photography on Your Desk,” *Proc. Fourth Int'l Conf. Computer Vision,* pp. 43-50 1998.

[8] K.L. Boyer and A.C. Kak, “Color-Encoded Structured Light for Rapid Active Ranging,” *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 9, no. 1, 1987.

[9] Y. Boykov, O. Veksler, and R. Zabih, “Fast Approximate Energy Minimization via Graph Cuts,” *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 23, no. 11, Nov. 2001.

[10] C. Chen, Y. Hung, C. Chiang, and J. Wu, “Range Data Acquisition Using Color Structured Lighting and Stereo Vision,” *Image and Vision Computing,* vol. 15, no. 6, pp. 445-456, June 1997.

[11] B. Curless, “Overview of Active Vision Techniques,” *Proc. SIGGRAPH 99 Course on 3D Photography,* 1999.

[12] B. Curless and M. Levoy, “Better Optical Triangulation through Spacetime Analysis,” *Proc. Int'l Conf. Computer Vision,* pp. 987-994, 1995.

[13] J. Davis and X. Chen, “A Laser Range Scanner Designed for Minimum Calibration Complexity,” *Proc. Third Int'l Conf. 3D Digital Imaging and Modeling,* 2001.

[14] J. Davis, R. Ramamoorthi, and S. Rusinkiewicz, “Spacetime Stereo: A Unifying Framework for Depth from Triangulation,” *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* pp. II-359-II-366, 2003.

[15] U. Dhond and J. Aggarwal, “Structure from Stereo—A Review,” *IEEE Trans. Systems, Man, and Cybernetics,* vol. 19, no. 6, 1989.

[16] O. Hall-Holt and S. Rusinkiewicz, “Stripe Boundary Codes for Real-Time Structured-Light Range Scanning of Moving Objects,” *Proc. Int'l Conf. Computer Vision,* pp. 359-366, 2001.

[17] S. Inokuchi, K. Sato, and F. Matsuda, “Range-Imaging for 3D Object Recognition,” *Proc. Int'l Conf. Pattern Recognition,* pp. 806-808, 1984.

[18] R. Jarvis, “A Perspective on Range Finding Techniques for Computer Vision,” *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 5, no. 2, pp. 122-139, 1983.

[19] T. Kanade, A. Gruss, and L. Carley, “A Very Fast VLSI Rangefinder,” *Proc. IEEE Int'l Conf. Robotics and Automation,* pp. 1322-1329, 1991.

[20] S. Kang, J. Webb, C. Zitnick, and T. Kanade, “A Multibaseline Stereo System with Active Illumination and Real-Time Image Acquisition,” *Proc. Int'l Conf. Computer Vision,* pp. 88-93, 1995.

[21] G. Medioni and J. Jezouin, “An Implementation of an Active Stereo Range Finder,” *Optical Soc. Am. Technical Digest Series,* vol. 12, pp. 34-51, 1987.

[22] Y. Ohta and T. Kanade, “Stereo by Intra- and Inter-Scaline Search Using Dynamic Programming,” *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 7, no. 2, 1985.

[23] M. Okutomi and T. Kanade, “A Locally Adaptive Window for Signal Matching,” *Int'l J. Computer Vision,* vol. 7, no. 2, 1992.

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

[25] K. Pulli, H. Abi-Rached, T. Duchamp, L. Shapiro, and W. Stuetzle, “Acquisition and Visualization of Colored 3D Objects,” *Proc. Int'l Conf. Pattern Recognition,* pp. 11-15, 1998.

[26] S. Rusinkiewicz, O. Hall-Holt, and M. Levoy, “Real-Time 3D Model Acquisition,” *ACM Trans. Graphics,* vol. 21, no. 3, pp. 438-446, 2002.

[27] P. Saint-Marc, J. Jezouin, and G. Medioni, “A Versatile PC-Based Range Finding System,” *IEEE Trans. Robotics and Automation,* vol. 7, no. 2, pp. 250-256, 1991.

[28] D. Scharstein and R. Szeliski, “A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms,” *Int'l J. Computer Vision,* vol. 47, no. 1, pp. 7-42, 2002.

[29] D. Scharstein and R. Szeliski, “High-Accuracy Stereo Depth Maps Using Structured Light,” *Proc. Computer Vision and Pattern Recognition,* 2003.

[30] E. Shechtman, Y. Caspi, and M. Irani, “Increasing Space-Time Resolution in Video,” *Proc. European Conf. Computer Vision,* 2002.

[31] T.C. Strand, “Optical Three-Dimensional Sensing for Machine Vision,” *Optical Eng.,* vol. 24, no. 1, pp. 33-40, 1985.

[32] L. Zhang, B. Curless, and S. Seitz, “Rapid Shape Acquisition Using Color Structured Light and Multi-Pass Dynamic Programming,” *IEEE 3D Data Processing Visualization and Transmission,* 2002.

[33] L. Zhang, B. Curless, and S. Seitz, “Spacetime Stereo: Shape Recovery for Dynamic Scenes,” *Proc. Computer Vision and Pattern Recognition,* 2003.