Shape-Time Photography

William T. Freeman*

EECS Dept.
Massachusetts Institute of Technology
Cambridge, MA 02139

Hao Zhang*

EECS Dept.
U.C. Berkeley
Berkeley, CA 94720

IEEE Computer Vision and Pattern Recognition (CVPR), Madison, WI, June, 2003.

We introduce a new method to describe shape relationships over time in a photograph. We acquire both range and image information in a sequence of frames using a stationary stereo camera. From the pictures taken, we compute a composite image consisting of the pixels from the surfaces closest to the camera over all the time frames. Through occlusion cues, this composite reveals 3-D relationships between the shapes at different times. We call the composite a shape-time photograph.

Small errors in stereo depth measurements can create artifacts in the shape-time images. We correct most of these using a Markov network to estimate the most probable front-surface pixel, taking into account (a) the stereo depth measurements and their uncertainties, and (b) spatial continuity assumptions for the time-frame assignments of the front-surface pixels.

1 Introduction

With a single still image, we seek to describe the changes in the shape of an object over time. Applications could include artistic photographs, instructional images (e.g., how does the hand move while sewing?), action summarization, and photography of physical phenomena.

How might one convey, in a still image, changes in shape? A photograph depicts the object, of course, but not its relationship to objects at other times. Multiple-exposure techniques, pioneered in the late 1800's by Marey and Murbridge [1, 9] can give beautiful depictions of objects over time. They have two drawbacks, however: (1) The control of image contrast is a problem; the image becomes over-exposed where objects at different times overlap. Backgrounds may need to be dark to avoid over-exposure. (2) The result doesn't show how the various shapes relate to each other in three-dimensions. What we see is like an X-ray photograph, showing only a flattened comparision between 2-d shapes.

Using background stabilization techniques from computer vision, researchers have developed video summarization tools which improve on multiple-exposure methods. Researchers at both Sarnoff Labs [13] and Salient Stills [7] have shown single-frame composites where the foreground image at each time overwrites the overlapping portions of all the previous foreground images, over a single, stabilized background. We will refer to this compositing as the "layer-by-time"algorithm, since it is time, not 3-D shape, which determines object visibility. The layer-by-time method avoids the contrast reduction of multiple exposure techniques, However, since temporal order, not shape, determines the occlusion relationships, this method cannot describe the shape relationships between foreground objects at different times. Video cubism [5] is a less structured approach to rendering video information into a single frame, and also does not incorporate shape information into the composite.

Our solution for displaying shape changes over time makes use of 3-D information which is captured along with the images. We form a composite image where the pixels displayed are those showing the surfaces closest to the viewer among all surfaces seen over the entire sequence. The effect is to display a photograph of the union of the surfaces in all the photographs (without mutual illumination and shading effects). This allows occlusion cues to reveal the 3-D shape relationships between objects seen over different times in the original video sequence.

Figure 1 illustrates these summarization methods for the case of a familiar motion sequence: the rattling spiral of a coin as it rolls to a stop on a table. (a) shows the individual frames of the sequence. (To avoid motion blur, we placed the coin in those positions, using clay underneath). The multiple-exposure summary, (b), shows the loss of image contrast where foreground objects overlap. The layer-by-time algorithm, (c), shows more detail than (b), but doesn't reveal how the coins of different times relate spatially. (d) is our proposed summary of the sequence. The composite

*This work was initiated when both authors were at Mitsubishi Electric Research Labs (MERL), WTF as a researcher and HZ as a student intern.

(a)

(b)

(c)

(d)

Figure 1: (a) Image sequence of rolling coin. (b) Multiple exposure summary. (c) Layer-by-time summary. (d) Shape-time summary. (Color-based foreground masks were used in (c) and (d) to isolate the foreground coins from the background: in (c) to specify the foreground object and in (d) to remove the unreliable stereo depth for the featureless background.)

image is constructed to make sense in 3-D. We can see how the coin occludes itself at other times; these occlusions let us picture the 3-D relationships between the different spatial configurations of the coin. To emphasize that the technique describes shapes over time, we call it "shape-time photography".

1.1 Related effects

In some special cases of natural viewing, we are accustomed to viewing shape-time images. Extrusion processes, such as squeezed blobs of toothpaste or shaving cream, leave a shape-time history of the motion of the extrusion source. Shape-time photographs have some resemblance to Duchamp's "Nude Descending a Staircase", the classic depiction of motion and shape in a static image. The comic book Nogenon uses drawn shape-time outlines in its story [14]. In unpublished independent work, researchers at Georgia Tech have made graphical displays of data from a motion-capture system using a shape-time style rendering, but not using visual input [2].

2 Problem Specification

To make a shape-time photograph, we need to record both image and depth information. Various technologies can measure depth everywhere in a scene, including shape-from-defocus, structured light systems, and stereo. While stereo range can be less accurate than others, a stereo camera is quite portable, allowing a broad range of photographic subjects in different locations. Stereo also avoids the problem of registering range and image data, since disparities are computed from the image data itself. Fig. 2 shows the stereo camera we used. The beam-splitter system allowed us to capture left and right images using a single shutter, assuring temporally synchronized images.

The simplest version of shape-time photography assumes a stationary camera which photographs N time-frames of stereo image pairs. (Background stabilization techniques such as [16] might be used to generalize the results of this paper to non-stationary cameras). At each position, we need to select for display a pixel from one of the N frames captured over all times at that position. We can then generate a single-frame composite, from one camera's viewpoint (left, for our examples), or a composite stereo image.

Let $L_k(t)$ and $R_k(t)$ denote the values at the kth pixel at time frame t recorded in the left and right images, respectively. Let $d_k(t)$ be the distance to the surface imaged at the kth pixel (of the left camera) at frame t. Pixel k of the left view shape-time image, I, is simply

$$I_k = L_k(\text{argmin}_t d_k(t)) \quad (1)$$

Figure 2: The apparatus for taking synchronized stereo image sequences: Olympus Camedia C-3040 camera, and a Pentax stereo adapter (connected using a Kenco 41mm - 52mm adapter ring). The digital camera can take 5 full-resolution shots in a row at 1/3 second intervals. The L/R split-screen image is visible in this photo on the camera's LCD display. Insert: a typical split-screen image recorded by the camera.

We call $\text{argmin}_t d_k(t)$ the frame assignment at pixel k, since it indicates which time-frame's pixel is displayed at position k in the shape-time composite image.

With perfect depth data, the frame assignment is trivial to compute; we simply find the frame of minimum depth for each position. However, substituting the measured stereo depth, $\hat{d}_k$ for the true depth d_k in Eq. (1) typically gives unacceptable artifacts, illustrated in Fig. 3. Equation (1) for shape-time rendering involves comparisons between very similar depth values from different frames and can reveal even small errors in stereo depth. We will need to estimate the proper frame assignments for each pixel in much the same way as one approaches other low-level vision problems: combining local evidence (the measured stereo depths and their uncertainties) with regularization constraints (penalties for inconsistent frame assignments over space). Shape-time composites made using depth measurements other than stereo may also benefit from the processing steps below.

3 Algorithm

One might design an algorithm to estimate the time-frame assignments directly from image data without first computing stereo depth. Motion coherence over time, as well as stereo, could then be used to estimate depth, as in [12]. However, we are often interested in sequences with large motions between frames, which are not amenable to that integrated approach. To address that more general case, we chose a modular architecture. We first measure stereo disparity, $\hat{d}_k(t)$, and its uncertainty, $\sigma_k(t)$, independently at each time, t. (Since we are only interested in ordinal relationships, we treat stereo disparity like depth.) We then assign the time frame to be displayed at each pixel based on those depth estimates. This modular approach will also let us incorporate improved stereo algorithms into our shapetime system as they are developed in the future. We can also substitute other depth measurement methods instead of stereo.

Our stereo camera is uncalibrated. We found the fundamental matrix by using the web-based automatic point matching algorithm of Zhang [18]. We rectified the image so epipoles are along scan lines using the algorthm of [11].

We used the stereo algorithm of Zitnick and Kanade (ZK) [19] which constructs a 3-dimensional array of match values in disparity space. The iterative algorithm enforces global constraints of uniqueness (one disparity value per point) and continuity (neighboring pixels have similar disparities) by diffusing or inhibiting support among neighboring match values. After the algorithm has converged, occluded areas are explicitly identified. A version of the code is available for download [19].

3.1 Probabilistic formulation

Small errors in depth estimates lead to islands of pixels where the selected frame switches in the shape-time composite, illustrated in Fig. 3. To remove those spurious frame switches, we add two assumptions: (1) that a pixel's time-frame assignment is likely to be the same as its neighbors, and (2) that time-frame assignment transitions are more likely to occur at image edges, because they may be occluding edges where a frame switch should occur. These are analogous to assumptions about disparity used in stereo [6, 15], but we are applying this to estimate the frame of minimum depth from a collection of frames, not the disparity.

A probabilistic formulation can combine these assumptions with the noisy stereo disparity data to compute the most probable frame of minimum disparity for each pixel. We assume that the frame assignments at each pixel form a Markov random field (MRF) [4, 6, 3, 15]. Let the vector $\vec{t}$ denote the frame assignments at all the pixel locations, ie, $\vec{t} = [t_1, t_2, \ldots, t_k, \ldots, t_M]$, where M is the number of pixels in the image. We seek to define a probability $P(\vec{t})$ such that the time-frame assignments which maximize $P(\vec{t})$ result in an artifact-free shape-time composite image.

Toward this end, we write $P(\vec{t})$ as a product of (a) local evidence terms, $\psi_k(t_k)$, incorporating stereo disparity information, and (b) neighbor compatibility matrices, $\phi_{jk}(t_k, t_j)$, encouraging spatially consistent pixel frame as-

signments for the composite image:

$$P(\vec{t}) = \frac{1}{Z} \prod_{(jk)} \phi_{jk}(t_k, t_j) \prod_k \psi_k(t_k). \quad (2)$$

Z is a normalization constant, independent of $\vec{t}$. The parentheses in $\prod_{(jk)}$ mean the product is taken over only neighboring pixels j and k.

3.1.1 Local evidence terms

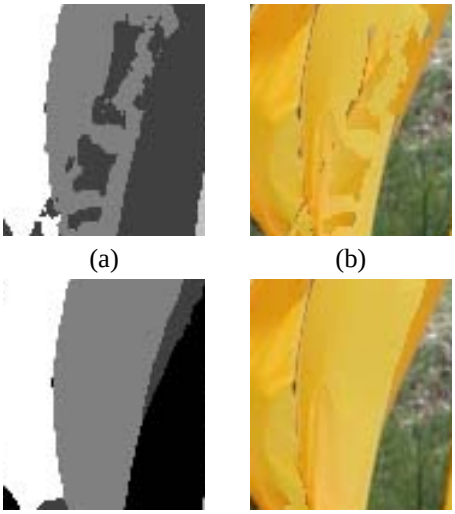
$\psi_k(t_k)$ is an N-vector describing the probability, based on stereo depth and uncertainty measurements, that time-frame t_k is in front of the others, at pixel k. We assume that each frame's depth measurement $d_k(t)$, at pixel k, is an independent Gaussian random variable of mean $\hat{d}_k(t)$ and standard deviation $\sigma_k(t)$. We want to transform these uncertain depth measurements into a set of probabilities that each time-frame corresponds to the front-most surface, see Fig. 5. Let $d(1), ..., d(N)$ be the N independent, but not identically distributed Gaussians (dropping the subscript k for brevity here). We want to compute the probability that $d(1)$ (say) is the smallest among all the $d(t)$'s, namely, $P(d(1) < d(2), ..., d(1) < d(N))$. We condition on the value of $d(1)$:

$$\begin{aligned} & P(d(1) < d(2), ..., d(1) < d(N)) \\ = & \int_x P(d(2) > d(1), d(3) > d(1), \\ & ..., d(N) > d(1) \mid d(1) = x) P(d(1) = x) \quad (3) \\ = & \int_x P(d(2) > x) P(d(3) > x) \\ & ... P(d(N) > x) P(d(1) = x) \\ & \text{because the } d(t)\text{'s are independent} \quad (4) \end{aligned}$$

This lets us evaluate the integration over the joint probability in N-dimensions using a 1D integral. We evaluate that integral numerically using simple trapezoid quadrature.

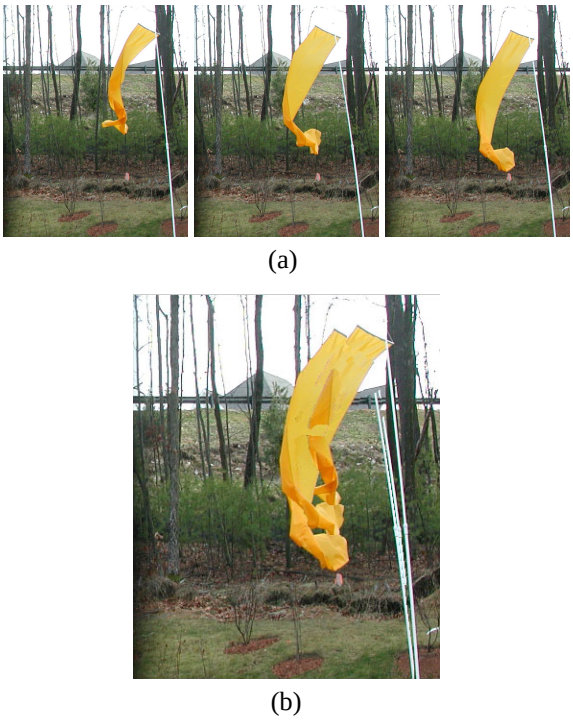
The ZK stereo algorithm returns a confidence value, $c_k(t)$, for every disparity estimate. We used the confidence value to estimate a standard deviation, $\sigma_k(t)$, of the Gaussian distribution used to model each pixel's true disparity. We used a heuristically selected function, mapping ZK confidence 0.05 to Gaussian disparity $\sigma = 0.03$; confidence 0.002 to $\sigma = 0.2$, linearly interpolating in between. We clamped σ to the extremal values outside that interval.

In Eq. (5) below for belief propagation, the resulting local evidence vectors, describing the probability that each time-frame corresponds to the front-most surface pixel, are treated as a message coming into the node at that pixel.



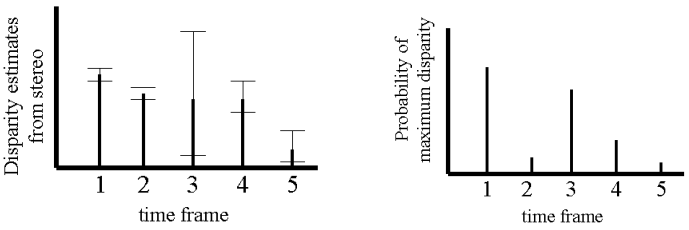
(a) (b) (c) (d)

Figure 3: (a) Time-frame assignments for the front-most surface pixels, based on stereo depth measurements alone, without MRF processing. Grey level indicates the time-frame assignment at each pixel. (b) Shape-time image based on those assignments. (c) Most probable time-frame assignments, computed by MRF. (d) Resulting shape-time image. Note that the belief propagation in the MRF has removed spurious frame assignment changes.



(a) (b)

Figure 4: (a) Component frames of banner in wind. (b) Shape-time composite, showing the evolving flag shapes in relation to each other.



(a) output from stereo (b) local evidence input to MRF

Figure 5: (a) From the stereo algorithm, for any given pixel, we find the disparity estimate and its uncertainty for each time frame. (b) We convert these to the probability that each time frame has the maximum disparity at the given pixel. Note that the probability is not necessarily monotonic with the mean disparity (illustrated by the synthetic data plotted here, which are monotonically decreasing in mean disparity over the time frames shown, but not monotonically decreasing in the probability of maximum disparity). These probabilities (in (b)) are the input to the MRF which computes an approximation to the optimal time frame assignment at each pixel.

3.1.2 Neighbor compatibility matrices

The compatibility matrix, $\phi_{jk}(t_j, t_k)$, between neighboring nodes is an $N \times N$ matrix determining the probability of a minimum disparity frame transition from frame t_j to frame t_k between the (neighboring) pixels j and k. We set these probabilities to more easily allow frame assignment breaks in regions of large image gradients. Let the squared magnitude of the image gradient at time t and pixel k be $E_k(t)$ (scaled to range from 0 to 1). The diagonal entries of $\phi_{(jk)}(t_j, t_k)$ are 1. The off-diagonal entries are $\max(E_j(t_j), E_j(t_k), E_k(t_j), E_k(t_k))$.

3.1.3 Belief propagation

We have constructed Eq. (2) so that the $\vec{t}$ which maximizes $P(\vec{t})$ represents our best estimate for the desired time-frame assignments for the shape-time composite image. Exact maximization of $P(\vec{t})$ is NP-hard, but good approximate methods exist [17, 6]. We found good results using belief propagation [10, 17, 3] (see [8] for code), which imposes no constraints on the form of the matrices $\phi_{(jk)}(t_k, t_j)$. (See [15] for stereo depth estimation using belief propagation in a Markov random field.)

In belief propagation, each node sends a "message", $\vec{m}(i)$, to all the neighboring nodes. The messages are N-vectors, initialized to all ones. The iterative update equation for the message from node r to node s is:

$$\vec{m}_s^r(i) \leftarrow \sum_j \phi_{(rs)}(i,j) \prod_{(pr)} \vec{m}_r^p(j) \qquad (5)$$

Upon convergence, the marginal probability that time-frame i has the maximum disparity is contained in the "belief" at a node, s, $\vec{b}_s(i)$, obtained from the messages through:

$$\vec{b}_s(i) = \prod_{(pr)} \vec{m}_r^p(i) \qquad (6)$$

Twenty iterations of passing messages between all pairs of pixels yielded an estimate for the belief $b_k(t_k)$, the marginal probability that time-frame t_k is in front at pixel k. We selected the $\hat{t}_k$ maximizing $b_k(t_k)$; the displayed shape-time composite image was $I_k = L_k(\hat{t}_k)$.

We obtained improved stereo disparity results if we bandpass filtered and contrast normalized the rectified images before calculating disparities [3]. This lessened the effect of brightness variations within our stereo camera and fixed some matching problems in low-contrast regions in the image. For Figs. 6 and 8, the stereo disparity values in the distant background were too noisy to be useful, so we generated a mask isolating the foreground person from the image. Both the disparity calculation and the shape-time computation took roughly 90 seconds to compute for typical images shown here, computed on a 500 MHz machine.

4 Results

We show results indicating possible applications of the shape-time technique. Fig. 4 shows a blowing flag where fluid dynamics controls the shape evolution over time. The method allows a new way to visualize those shape changes over time.

Figs. 6, 7 and 8 show shape-time applied to people. Figs. 6 and 7 give instructional single-frame summaries of short actions, such as someone's style of throwing, or a particular sewing stitch. Fig. 8 shows relationship between different shapes on the body or face.

Fig. 9 examines the water height at different phases of a wave breaking on the shore, revealing the surge in water height relative to the other frames at the final frame of the sequence, which dominates in the shape-time composite image. Fig. 1 (d) shows shape over time as the coin falls.

Some artifacts are visible in the composite images. The face is broken-up slightly in Fig. 6, and some background pixels are seen attached to the index finger of the left hand in Fig. 7. These are caused by edge artifacts in the depth data from the stereo algorithm. The MRF processing can tolerate small stereo depth errors, but the approach cannot work for imaging conditions where stereo depth gives no useful information, or when such regions cannot be masked out of the image.

One feature of our present implementation is that the composite image does not depict the temporal ordering; we treat all time frames identically. This could be easily

(a)

(b)

Figure 6: (a) Frames of girl throwing snowball. (b) Shape-time photograph showing the girl's throwing form.

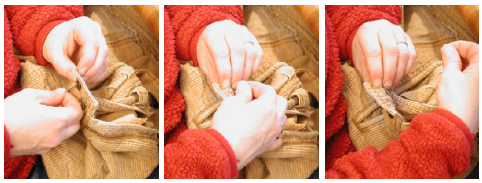
changed by altering each displayed pixel by some function of its time frame: for example, by darkening, reducing the color saturation or reducing the opacity of pixels further back in time. We chose the current implementation to emphasize the depiction of shape relationships over the depiction of temporal dynamics.

5 Conclusions

We proposed a new method for showing the shape relationships between objects at different times. We point out the usefulness of shape-time photography, and show a method to implement it. We developed an algorithm to reduce the artifacts resulting from noise in stereo disparity measurements. Since this rendering method is sensitive to small errors in depth, the algorithm may be needed for shape-time renderings from other depth modalities, as well.

The method occupies a a special-effects niche. Shape-time photography could be useful for summarizing action, for instructional photographs, or physics illustrations. It can describe in a picture how things move. It may reveal a pattern or spatial relationship in the world that is not clear from the video sequence or its individual frames.

The shape-time rendering of this paper is a special case



(a)

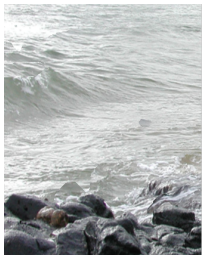
(b)

Figure 7: (a) Frames of sewing stitch example. (b) Shape-time rendering of the sewing stitch, illustrating the hand's movement.

Figure 8: Portrait of a man, from frontal and profile views. Intersection contours in the shape-time image describe his face shape.

(a)

(b)



(c)

Figure 9: (a) Images in wave sequence. (b) Shape-time composite image of ocean wave breaking. (c) Inside-out rendering of wave (furthest surface shown at every point).

of a more general problem: given a stack of images captured from one viewpoint, use computer vision analysis to select which pixels to display in a composite image. The pixel selection could depend on object motion (show where the objects moved fastest, or where something moved toward you), or on the orientation of a face (show wherever the dancer looked back). As one example of this generalization, in Fig. 9 (c) we show the wave rendered "inside out": we display the surfaces *furthest away* from the camera. This gives a picture of the lowest water in the breaking wave during its cycle.

Acknowledgments

We thank Steve Seitz for helpful conversations about this research.

References

[1] M. Braun. *Picturing Time*. University of Chicago, 1992.

[2] I. Essa, 2002. Personal communication.

[3] W. T. Freeman, E. C. Pasztor, and O. T. Carmichael. Learning low-level vision. *Intl. J. Computer Vision*, 40(1):25–47, 2000.

[4] S. Geman and D. Geman. Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

[5] A. Klein, P. Sloan, A. Colburn, A. Finkelstein, and M. F. Cohen. Video cubism. Technical Report MSR-TR-2001-45, Microsoft Research, 2001.

[6] V. Kolmogorov and R. Zabih. Computing visual correspondence with occlusions using graph cuts. In *Intl. Conf. on Computer Vision (ICCV)*, 2001.

[7] M. Massey and W. Bender. Salient stills: process and practice. *IBM Systems Journal*, 35(3&4):557–574, 1996.

[8] K. Murphy, 2001. www.cs.berkeley.edu/ murphyk/Bayes/bnt.html.

[9] E. Muybridge. *Horses and other animals in motion*. Dover, 1985.

[10] J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, 1988.

[11] M. Pollefeys, R. Koch, and L. V. Gool. A simple and efficient rectification method for general motion. In *Intl. Conf. on Computer Vision (ICCV)*, pages 496–501, 1999.

[12] H. S. Sawhney, Y. Guo, K. Hanna, R. Kumar, S. Adkins, and S. Zhou. Hybrid stereo camera. In *ACM SIGGRAPH*, 2001. In *Computer Graphics* Proceedings, Annual Conference Series.

[13] H. S. Sawhney and R. Kumar, 2001. Personal communication.

[14] L. Schuiten and F. Schuiten. *Nogegon*. Humanoids Publishing, www.humanoids-publishing.com, 2001.

[15] J. Sun, H. Shum, and N. Zheng. Stereo matching using belief propagation. In *Proc. ECCV*, 2002.

[16] H. Tao, H. Sawhney, and R. Kumar. Dynamic layer representation with applications to tracking. In *Proc. of the IEEE Computer Vision and Pattern Recognition*, Hilton Head, SC, 2000.

[17] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Generalized belief propagation. In *Adv. in Neural Info. Proc. Systems*, volume 13, pages 689–695. MIT Press, 2001.

[18] Z. Zhang. Determining the epipolar geometry and its uncertainty: A review. Technical Report 2927, Sophia-Antipolis Cedex, France, 1996. see http://www-sop.inria.fr/robotvis/demo/f-http/html/.

[19] C. L. Zitnick and T. Kanade. A cooperative algorithm for stereo matching and occlusion detection. *IEEE Pattern Analysis and Machine Intelligence*, 22(7), July 2000.