Abstract

We introduce a new method to describe, in a single image, changes in shape over time. We acquire both range and image information with a stationary stereo camera. From the pictures taken, we display a composite image consisting of the image data from the surface closest to the camera at every pixel. This reveals the 3-d relationships over time by easy-to-interpret occlusion relationships in the composite image. We call the composite a shape-time photograph.

Small errors in depth measurements cause artifacts in the shape-time images. We correct most of these using a Markov network to estimate the most probable front surface, taking into account the depth measurements, their uncertainties, and layer continuity assumptions.

January, 2002. Part of this work was done while both authors were at Mitsubishi Electric Research Laboratories, Cambridge, MA 02139.

Shape-Time Photography

William T. Freeman
MIT Artificial Intelligence Laboratory
and Mitsubishi Electric Research Labs

Hao Zhang
U. C. Berkeley Computer Science Department

1 Introduction

With a 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?), action summarization, and photography of physical phenomena.

How might one convey changes in shape with a still image? A photograph depicts the object, of course, but not its change over time. Multiple-exposure techniques, pioneered in the late 1800's by Marey and Murbridge [Braun 1992; Muybridge 1985] can give beautiful static descriptions of motion. However, they have two drawbacks: (1) The control of image contrast is a problem; the image becomes over-exposed where objects at different times overlap. Backgrounds need to be dark. (2) The result doesn't tell how the various shapes relate to each other in three-dimensions. What we see is like an X-ray, showing only a flattened comparision between the 2-d shapes.

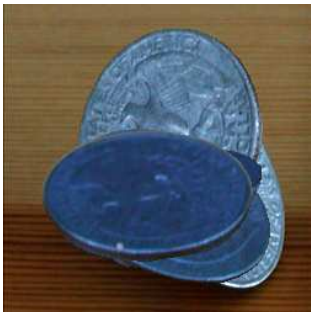
Using background stabilization techniques from computer vision, researchers have developed video summarization tools which improve on the multiple-exposure techniques. Researchers at both Sarnoff Labs [Sawhney and Kumar 2001] and Salient Stills [Salient 2001] have shown video sumaries where the foreground image at each time overwrites overlapping portions of the previous foreground images, composited onto a stabilized background. We will refer to this compositing as the "layer-by-time" algorithm, since time, not 3-d shape, determines object visibility. The layer-by-time method avoids the contrast reduction of multiple exposures, but unfortunately cannot describe the shape relationships between foreground objects at different times.

Our solution for displaying shape changes over time makes use of 3-d information, captured along with the images. We form a composite image where the pixels displayed are those which were in front of the pixels at all other times at the same location. The effect is to display a surface that is the union of the surfaces in all the photographs (without mutual shading). This allows objects to occlude themselves at different times, revealing the 3-d shape relationships.

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

(a)

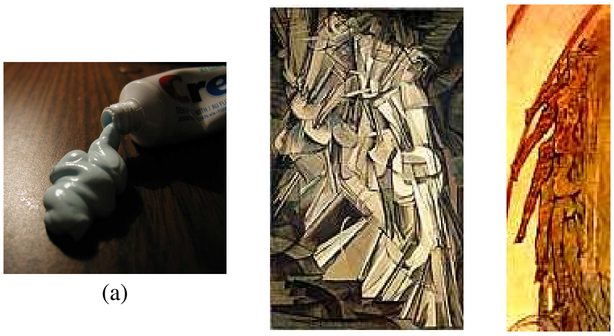
(b)



(c)

(d)

Figure 1: **(a) Image sequence of rolling coin. (b) Multiple exposure summary. (c) Layer-by-time summary. (d) Shape-time summary. (The same color-based foreground masks were used in (c) and (d)).**



(a) (b) (c)

Figure 2: **(a) Extruded shapes leave a shape-time summary of motion. (b) Nude descending a staircase, by Duchamp, describes shape relationships over time, similar to our shape-time summaries. (c) Cover of Nogegon, showing hand-drawn shape-time images.**

describe how the coins of different times relate spatially. (d) is our proposed summary of the sequence. The composite 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. To emphasize that the technique describes shapes over time, we call it "shape-time photography".

Fig. 2 shows other images related to shape-time photography. (a) shows squeezed toothpaste; we are able to look at such shape-time images of extruded material and infer the motion histories. Shape-time photographs have some resemblance to Duchamp's "Nude Descending a Staircase", (b) (the classic depiction of motion and shape in a static image). The comic book Nogenon uses drawn shape-time outlines, (c), in its story [Schuiten and Schuiten 2001].

2 Problem Specification

To make a shape-time photograph, we need to record both image and depth information. Many 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 that of more sophisticated technologies, a stereo camera is very portable, allowing a broader range of photographic subjects. Stereo also avoids the problem of registering range and image data, since disparities are computed from the image data. Fig. 3 shows the stereo camera we used. The beam-splitter system allowed us to capture left and right images using a single shutter, giving synchronized images.

The simplest version of shape-time photography assumes a stationary camera, which photographs N stereo image pairs. (Background stabilization techniques such as [Tao et al. 2000] should be useful in generalizing the results of this paper to non-stationary cameras). At each pixel, we need to decide which of the N pixel values captured there to display. We can generate a single-frame composite, from one camera's viewpoint (left, for our examples), or a a composite stereo image.

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

$$I_k^L = L_k(\text{argmin}_t {d^L}_k(t)) \quad (1)$$

We call $\text{argmin}_t {d^L}_k(t)$ the layer assignment at pixel k, since it indicates which time layer is displayed there in the composite shape-time image.

Figure 3: **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.**

For perfect depth data, this is trivial to compute: at every position, display the pixel of the layer for which the depth is smallest. However, substituting the measured depth, $\hat{d}_k^L$ for the true depth d_k^L in Eq. (1) gives unacceptable artifacts, illustrated in Fig. 4. Shape-time rendering involves comparisons between nearly equal depth values from different frames and can reveal even small errors in depth. We will need to estimate the proper layer assignments, starting from the measured depth and its uncertainty. Other depth measuring technologies may also need the processing steps we describe below.

3 Algorithm

One could create an algorithm which estimated the layer assignments directly from all the image data. Motion coherence over time, as well as stereo, could be used to estimate depth, as in [Sawhney et al. 2001]. However, we are often interested in large motions between frames, which doesn't benefit from that integrated approach, so instead we chose a modular architecture. Thus, we first measure stereo disparity, $\hat{d}_k^L(t)$, and its uncertainty, $\sigma_k^L(t)$, independently at each time, t. (Since we are only interested in ordinal relationships, we treat stereo disparity like depth). This approach lets us incorporate improved stereo algorithms as they are developed. We could also change to other depth measurement methods without changing our algorithm.

Small errors in depth estimates lead to the islands where the selected layer switches in the shape-time composite, in Fig. 4. To remove those spurious layer switches, we add assumptions: (1) that a pixel's layer assignment is likely to be the same as its neighbors, and (2) that layer transitions are more likely to occur at image edges, because they may be occluding edges where a layer switch should occur. These are analogous to assumptions about disparity used in stereo [Kolmogorov and Zabih 2001]. A probabilistic formulation can fold in these assumptions with the measurements of disparity and uncertainty.

We assume that the layer assignments at each pixel form a Markov random field (MRF) [Geman and Geman 1984; Kolmogorov and Zabih 2001]. Let $\vec{t}$ denote the layer assignments at all the pixel locations, ie, $\vec{t} = [t_1, t_2, \ldots, t_k, \ldots, t_M]$, where M is the

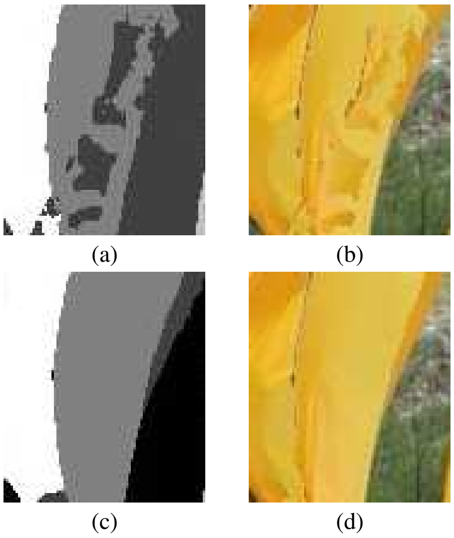


Figure 4: **(a) Layer assignments, without MRF processing. (b) Shape-time image based on those assignments. (c) Most probable layer assignments, computed by MRF. (d) Resulting shape-time image.**

number of pixels in the image.

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

where Z is a normalization constant.

$\phi_{(jk)}(t_j,t_k)$ is an NxN matrix determining the probability of a layer transition from layer t_j to layer t_k between the (neighboring) pixels j and k. We set these probabilities based on edge information at the pixels j and k at times t_k and t_j. 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 $\max(E_j(t_j),E_j(t_k),E_k(t_j),E_k(t_k))$.

$\psi_k(t_k)$ is an N-vector describing how probable it is that each of the N layers is in front, based on the depth and uncertainty measurements $d_k^L(t)$ at each time at pixel k. We assume that each layer's depth measurement is an independent Gaussian random variable of mean $\hat{d}_k^L(t)$ and standard deviation $\sigma_k^L(t)$. The probability that the depth at any layer is smaller than that of all the other layers is the product of N 1-d Gaussian integrals, which we evaluate analytically.

We have constructed Eq. (2) so that the $\vec{t}$ which maximizes $P(\vec{t})$ represents our best estimate for the desired layer assignments for the shape-time composite image. Exact maximization of $P(\vec{t})$ is NP-hard, but good approximate methods exist [Yedidia et al. 2001; Kolmogorov and Zabih 2001]. We found good results using belief propagation [Pearl 1988; Yedidia et al. 2001] (see [Murphy 2001] for code), which imposes no constraints on the form of the matrices $\phi_{(jk)}(t_k,t_j)$. Twenty iterations of passing messages between all pairs of pixels yielded an estimate for the belief $b_k(t_k)$, the marginal probability that layer t_k is in front at pixel k. We selected the t_k maximizing $b_k(t_k)$.

Our stereo camera is uncalibrated. We found the fundamental matrix by using the web-based point matching algorithm of Zhang [Zhang 1996]. We rectified the image so epipoles are along scan lines using the algorthm of [Pollefeys et al. 1999].

We obtained improved stereo disparity results if we bandpass filtered and contrast normalized the rectified images before calculating disparities [Freeman et al. 2000]. This lessened the effect

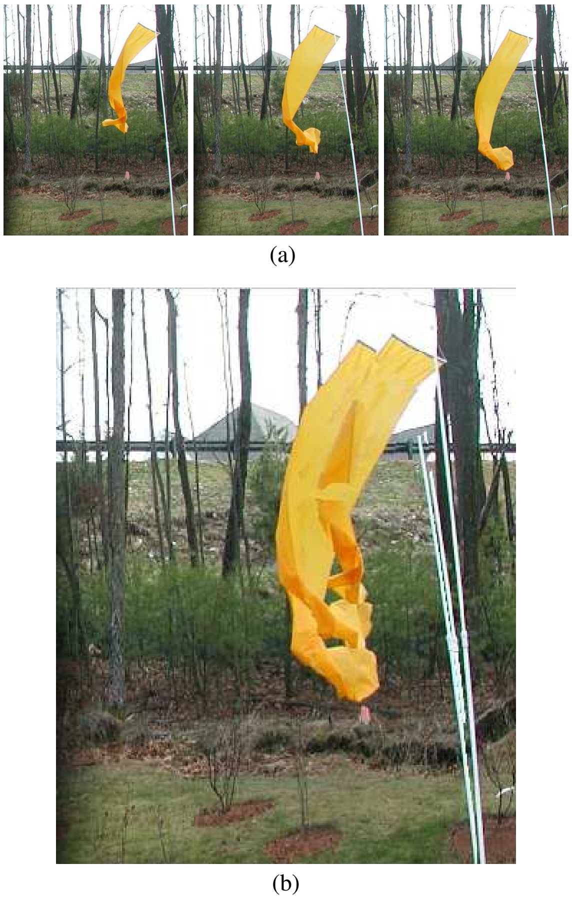


Figure 5: **(a) Component frames of banner in wind. (b) Shape-time composite, showing the evolving flag shapes in relation to each other.**

(a)

(b)

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

of brightness variations within our stereo camera and of 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 hand-drew 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.

4 Results

We show results indicating possible applications of the shape-time technique. Fig. 5 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, allowing visualization of one's actions over time (Fig. 6 and 7) or the relationship between different shapes on the body or face (Fig. 8).

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.

5 Conclusions

We proposed a new method for summarizing short video sequences. We point out the usefulness of shape-time photography, and show

(a)

(b)

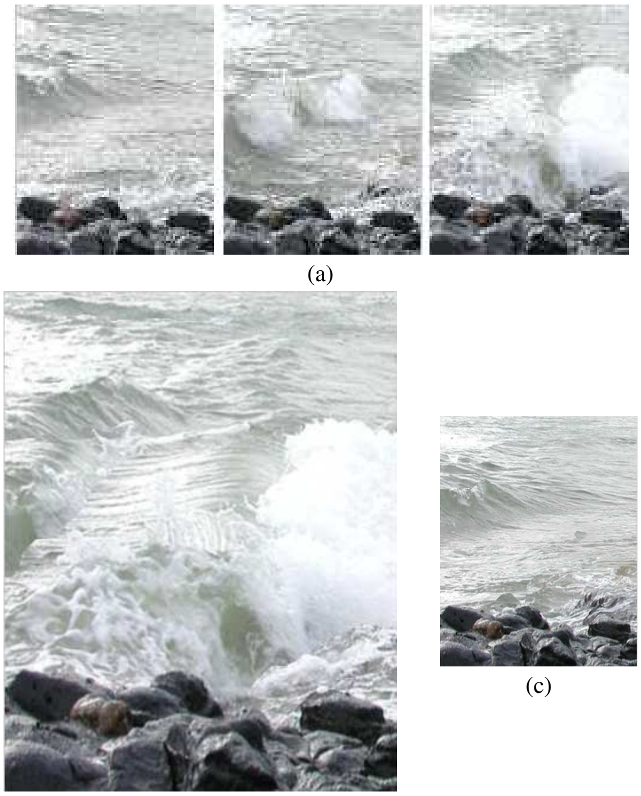
Figure 7: **(a) Frames of sewing stitch example. (b) Shape-time rendering of the sewing stitch, illustrating the hand's movement.**

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

a 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. It could be useful for summarizing action, for instructional photographs, or physics illustrations. It can describe in a picture how things move.

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



(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).**

References

BRAUN, M. 1992. *Picturing Time*. University of Chicago.

FREEMAN, W. T., PASZTOR, E. C., AND CARMICHAEL, O. T. 2000. Learning low-level vision. *Intl. J. Computer Vision 40*, 1, 25–47.

GEMAN, S., AND GEMAN, D. 1984. Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Pattern Analysis and Machine Intelligence 6*, 721–741.

KOLMOGOROV, V., AND ZABIH, R. 2001. Computing visual correspondence with occlusions using graph cuts. In *International Conference on Computer Vision (ICCV)*.

MURPHY, K., 2001. www.cs.berkeley.edu/ murphyk/Bayes/bnt.html.

MUYBRIDGE, E. 1985. *Horses and other animals in motion*. Dover.

PEARL, J. 1988. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann.

POLLEFEYS, M., KOCH, R., AND GOOL, L. V. 1999. A simple and efficient rectification method for general motion. In *International Conference on Computer Vision (ICCV)*, 496–501.

SALIENT, 2001. www.salientstills.com.

SAWHNEY, H. S., AND KUMAR, R., 2001. Presentation describing Sarnoff Labs Vision Group.

SAWHNEY, H. S., GUO, Y., HANNA, K., KUMAR, R., ADKINS, S., AND ZHOU, S. 2001. Hybrid stereo camera. In *ACM SIGGRAPH*. In *Computer Graphics* Proceedings, Annual Conference Series.

SCHUITEN, L., AND SCHUITEN, F. 2001. *Nogegon*. Humanoids Publishing, www.humanoids-publishing.com.

TAO, H., SAWHNEY, H., AND KUMAR, R. 2000. Dynamic layer representation with applications to tracking. In *Proc. of the IEEE Computer Vision and Pattern Recognition*.

YEDIDIA, J. S., FREEMAN, W. T., AND WEISS, Y. 2001. Generalized belief propagation. In *Adv. in Neural Information Processing Systems*, MIT Press, vol. 13, 689–695.

ZHANG, Z. 1996. Determining the epipolar geometry and its uncertainty: A review. Tech. Rep. 2927, Sophia-Antipolis Cedex, France. see http://www-sop.inria.fr/robotvis/demo/f-http/html/.