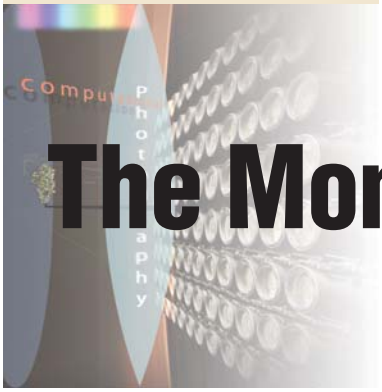

The Moment Camera

Michael F. Cohen and Richard Szeliski
Microsoft Research

Future cameras will let us "capture the moment," not just the instant when the shutter opens. The moment camera will gather significantly more data than is needed for a single image. This data, coupled with automated and user-assisted algorithms, will provide powerful new paradigms for image making.

Before the advent of the camera, artists were tasked with recording events and providing a visual history of their world. Although a great deal of early art recorded religious or mythical stories, by the 16th century, artists in the Netherlands began depicting scenes of normal life, typified by Pieter Bruegel's paintings (www.ibiblio.org/wm/paint/auth/bruegel/). Although no one believes that all the action depicted in these scenes took place at the same instant, Bruegel successfully captured the moment.

The moment provides a key concept, both in our article title and in the preceding sentence. What might we mean by *a moment* in this context?

To illustrate this concept, we can construct an axis that runs from the objective to the subjective, as Figure 1 shows. At the objective end, a photograph provides some semblance of an event's objective visual record. That same visual event evokes a different internal experience in each of us. At the subjective end of the axis, personal experiences of external stimuli are often referred to as qualia in philosophical discussions.[1] Somewhere, close to the objective end of the axis but still subjective, lies a point we call a moment. While a quale is by definition both subjective and personal, a moment is subjective but universal.

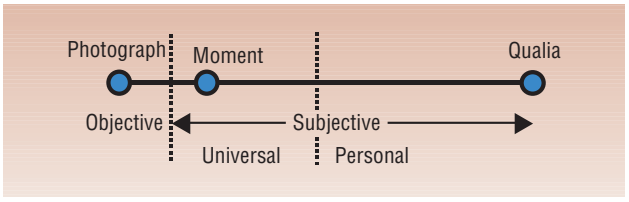

Figure 1. A moment. Although subjective, a moment lies close enough to the objective axis to represent a shared experience of a scene.

For example, people spend about 10 percent of their waking life with their eyes closed[2]—a person's normal, resting blink rate being 20 closures per minute, with the average blink lasting one-quarter of a second. Yet, when looking at our friends, we universally do not *see* them as having their eyes closed unless we consciously concentrate on their blinking.

On the other hand, taking a photograph of a friend often surprises us because the picture reveals closed or partially closed eyes, as Figure 2 shows. The rather awkward expression of half-closed eyes clearly does not capture the moment, because it does not correspond to what we experience when looking at our friend.

With the advent of the camera in the mid-19th century, art began to move away from realistic depiction into the more abstract realms of Impressionism, Cubism, and more pure Abstraction. The camera, although capable of capturing instants in time, cannot on its own—except in rare instances—truly record moments.

When coupled with computation and a user interface, digital cameras can bring back the ability to capture moments as opposed to just instantaneous snapshots. Such computational cameras or computational photography systems can provide a wealth of opportunities for both professional and casual photographers.

Published by the IEEE Computer Society
0018-9162/06/$20.00 © 2006 IEEE

Our hypothetical *moment camera* contains new light-capture modalities that can leverage several recent research developments in computer graphics, computer vision, and the subfield at their intersection, image-based rendering.

THE MOMENT CAMERA

When turned on, current digital cameras constantly scan the scene they are pointed at, responding to changing lighting conditions by modifying their speed or aperture and setting the focus to adapt to depths in the scene. Meanwhile, the user points the camera, trying to frame a shot, and waits for that elusive instant to push the button to record the light entering the aperture and landing on the sensor. At that instant, the camera might decide to fire the flash, at which time the total light then landing on the sensor during a fixed exposure interval is mapped to a raw image. This image typically receives further processing from a demosaicing algorithm before being compressed into a JPEG image for transfer to the permanent memory medium.

Imagine a modification in the camera's underlying functionality that keeps it always recording, somewhat like a DV camera in record mode. Thus, rather than only recording a snapshot, the camera constantly records time slices of imagery. Let's assume one frame every 100th of a second or less, depending on the mode. Let's also assume a finite round-robin buffer of perhaps 500 frames, or 5 seconds, resulting in a spacetime slab in memory at all times. We can think of this most easily as a short video sequence.

We refer to this new device as a moment camera. When coupled with computational photography algorithms and an appropriate user interface, this somewhat unremarkable change in functionality provides many new possibilities. To demonstrate the technology today, we simulate the moment camera with either a still camera taking multiple photographs in succession, or with a current DV camera at 30 frames per second and, unfortunately, at a significantly lower resolution.

MOMENT CAMERA PROCESSING STEPS

Although the input to the moment camera creates a spacetime slab, the moment's output typically consists of a single image. Thus, the processing primarily selects the color for each output pixel given the set of input images in the spacetime slab. This processing typically includes the following steps:

1. Align or warp the input images so that at any single output pixel, all input pixels represent the same point in the world as best as possible.
2. For each output pixel, from all input images that map to that pixel, select the best one to use for the output.
3. Adjust the selected pixel's color to blend seamlessly with its neighbors.

Figure 2. The blink of an eye. Although these two photographs were taken a fraction of a second apart, only the second one captures the moment.

The first step, aligning images, is most often done by finding features in the images, then matching features across images to determine transformations for each image and align them in a global space.[3] Alternatively, dense correspondence fields can be computed and used to perform the alignment.[4]

The second step involves an optimization that, for each pixel, tries to locally make the best selection based on predefined or interactively defined criteria, while globally trying to maintain smoothness. We often refer to the local criteria for selecting any particular pixel as the *data cost*, while the cost for transitioning from a pixel of one time slice to another as the *smoothness cost*.

In early work, *Image Stacks* (http://research.microsoft.com/research/pubs/view.aspx?tr_id=666) relied on the user to make most decisions. More recent applications, including Photomontage[1] and Seamless Image Stitching,[5] explore the definition of the data and smoothness costs, either by the user or automatically. To achieve the trade-off between optimizing each pixel individually and creating a seamless result, applications often use graph cut techniques[6] as the optimization method.

In the third and final step, the pixel value can be modified either to adjust the virtual exposure or to compensate for other differences between images. For example, gradient domain blending modifies pixel values to match across seams while trying to maintain local gradients.[7,8]

We rely on these three steps in the examples that follow.

STILL CAMERA MODES

The moment camera can be used in a variety of modes. Each mode determines some aspects of the actual capture, but perhaps more importantly, it guides the user interface. We do not describe the details of each UI here because any real-world implementation will require much more thought and experimentation.

Point and shoot

In its simplest mode, from a user's perspective, the moment camera operates much like a current point-and-shoot camera. The user simply frames the shot and presses a button. However, unlike a current camera, the

Figure 3. Flash versus no flash. (a) A noisy, no-flash image and (b) a low-noise flash image combine to produce (c) a low-noise image with good lighting.

moment camera records images continuously, not just at the instant the user presses the shutter button.

As it records, the camera rapidly varies the exposure times, bracketing the neutral setting. The camera tests multiple points in the scene for focus and records images with varying focus settings. If low light is an issue, the flash can fire during a subset of the exposures.

Meanwhile, time is inexorably marching forward, so the images vary during the time they are taken. When the user pushes the button, the camera records a slab of spacetime beginning a couple of seconds in the past until perhaps a second or two in the future for further processing.

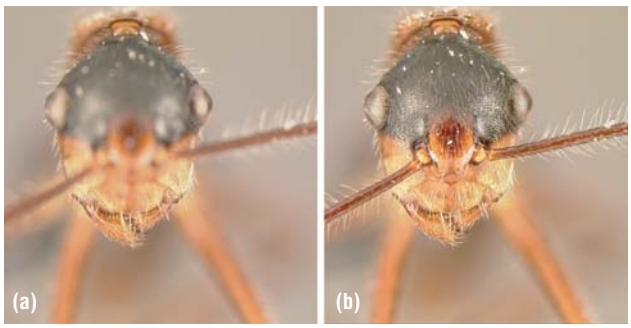

Figure 4. Expanded depth of field. The (a) single focal plane image is less detailed than (b) a composite of multiple focal plane images.

Figure 5. High-dynamic-range imagery. The moment camera can merge multiple exposures—bracketed shots—to get a wider dynamic range comparable to nondigital film techniques: (a) exposures merged without motion compensation versus (b) those with motion compensation.

The point-and-shoot moment camera supports several relatively simple application scenarios, including the following:

- *Wind time backward or forward*. Often the camera misses that fleeting expression at the instant the button push captures the image. Selecting a better frame as in Figure 2 more accurately captures the moment.
- *Flash/no flash*.[9,10] Low-light situations often lead to very noisy results, as Figure 3a shows. Using a flash can reduce the noise, but at the cost of ruining the subtle lighting, as Figure 3b shows. Because the spacetime slab contains both flash and no-flash images, the high-frequency details from the flash image can be combined with a smoothed version of the no-flash image to obtain a desired low-noise image while maintaining the original lighting, as Figure 3c shows.
- *Expanded depth of field*. Particularly when taking close-up shots, getting the whole object in focus simultaneously can be difficult. While the autofocus seeks to find a consensus depth on which to focus, the moment camera records multiple images with different focus settings. Thus, for every pixel location, the slab contains multiple versions of the same point with varying focus, as Figure 4 shows. Maximizing the focus involves detecting which pixel has the highest local contrast and selecting it, while simultaneously maintaining coherence using a smoothness term in the optimization criterion.

High-dynamic-range imagery and tone mapping

Current digital cameras suffer from limited dynamic range: They cannot image both very bright areas and dark areas in the same exposure. To compensate for this, multiple exposures—bracketed shots—can be merged to get a wider dynamic range.[11] Inside a moment camera, this kind of bracketing can be performed automatically, taking additional underexposed and overexposed shots when the camera detects that it is not adequately capturing the full dynamic range in a single shot. Global alignment followed by local optic flow can compensate for possible motion in the scene, as Figure 5 shows.[4]

Once a wide-dynamic-range image has been assembled, the camera can store it either in an extended dynamic-range image format for further processing or tone-map it back to a displayable 8-bit gamut. A more intelligent moment camera not only performs this processing onboard, but also lets the user interactively steer the tone-mapping process by indicating at a high level

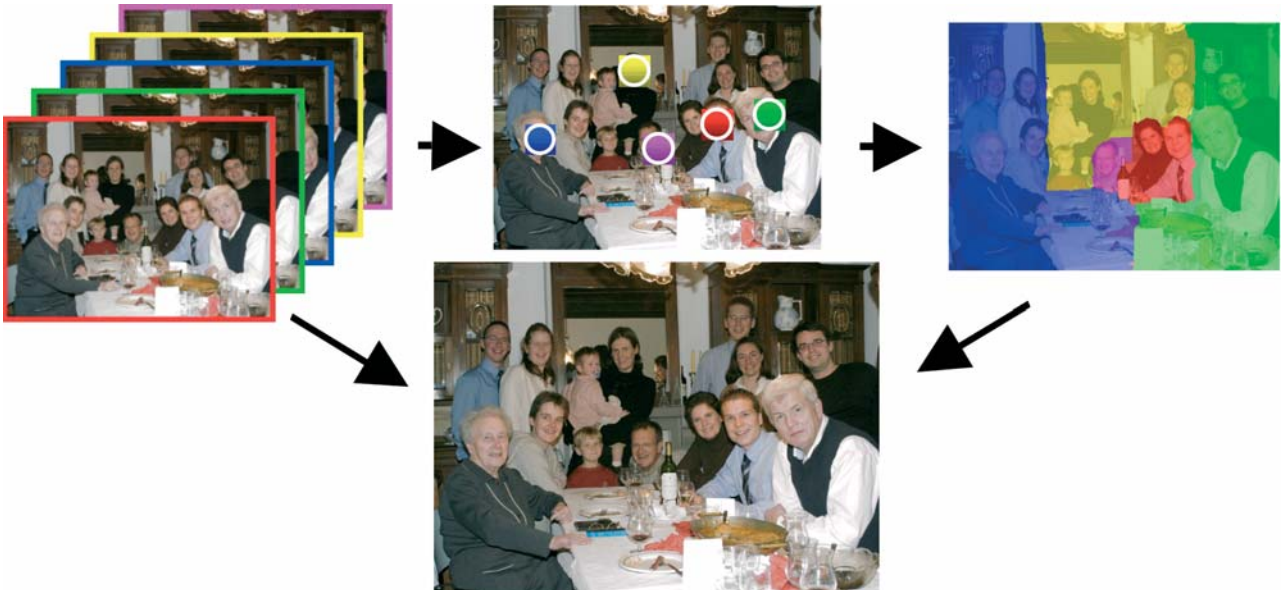

Figure 6. Group Shot. Working with stored images, the user indicates when each person photographed looks best. The system automatically finds the best regions around each selection to compose into a final group shot.

which regions should be brighter or darker or more or less saturated.[12]

Group Shot

When taking a picture, we often catch a person with their closed eyes. Taking a picture of a group of people exponentially increases the difficulty of avoiding this—it becomes almost impossible to capture an instant when everyone is smiling with their eyes open.

With an application such as Group Shot (http://research.microsoft.com/projects/GroupShot/), a user can assemble an ideal group photograph from multiple shots. The user indicates the best instance of each person, and the system finds the best jigsaw-puzzle-like regions that it can compose to create a seamless final image, as Figure 6 shows.

The moment camera can perform this operation in-camera to help ensure the creation of a successful composite. While viewing the scene, the user points at each person when they smile and look at the camera. Graph cut picks out a region around each selection to cut into the final composite and records a thin spacetime slab for that region. This can be repeated until a successful shot is created. Slight time shifts can be made on each region independently to perfect the result.

Panoramas: Widening the field of view

We are often confronted with a majestic scene—think of the Grand Canyon—that will not fit into the view finder. Multiple overlapping images can be stitched into a single panoramic image. Several applications can now do this after the fact. Many problems remain, however, that a moment camera could remedy.

The first problem is *coverage*. Without careful planning, we often miss parts of the scene. This happens most often in large sky areas or when the interesting parts of the scene lie at different heights in different directions. The results often have gaps or a snakelike shape rather than being a rectangular panorama.

By providing on-the-fly alignment and stitching, the user can literally paint the panorama, examining the coverage to ensure capturing the complete scene.[13] At the same time, allowing the exposure to vary between overlapping frames can create high-dynamic-range panoramas. Using shorter or longer exposures can adjust areas that appear too light or dark.

Finally, the world usually does not stand still during a panorama's capture. Focusing the graph cut criteria on selecting commonly seen and most likely static pixels can avoid including ghostlike figures in the panorama, as Figure 7 on the next page shows.

DEPICTING MOTION

While the previous examples purposefully remove transient events to create a consistent still, at times a user might want to explicitly depict motion in a single image. This type of representation dates back to the 19th century. Unless taken under careful conditions, stroboscopic imagery often results in ghostlike representations of the dynamic elements.

Stroboscopic-like stills

Leveraging graph cut, however, we can create stroboscopic-like images. By specifying in the objective function that we want to retain dynamic elements, as opposed to removing them as in the bottom half of Figure 7, the result resembles Figure 8, which shows a girl swinging across a set of monkey bars.

Figure 7. Panoramic composite. (a) The overlapping images are aligned and blended together, resulting in ghosted figures; (b) graph cut finds regions in each image to stitch together to create a consistent scene.

Figure 8. Stroboscopic-like images. Dynamic scenes can be represented by optimizing for dynamic elements while also maintaining consistency.

Cliplets

A spacetime slab is, by definition, the same as a short video sequence. Sometimes, a very short subsequence, or cliplet, can capture the moment, while still allowing the imagination to fill in what happened just before or after the bit of action.

Just as a still image forces the viewer's imagination to fill in what is left out, such short cliplets serve a similar purpose. These short sequences are best viewed by, for example, holding on the first frame for 3 to 4 seconds, then playing the short sequence and holding again on the final frame. Figure 9 provides an example that covers less than one-third of a second.

Motion loops

Some types of motion are more stochastic or repetitive. Examples range from flowing or rippling water to a person sitting still, breathing, and blinking. These motion types are amenable to the creation of looping video textures, which stochastically jump from one frame to a matching frame either forward or backward in time.[14] This work has also been extended to panoramic video textures constructed with video taken from a slowly panning camera.[15] The spacetime slab that the moment camera captures provides the input needed for these kinds of experiences.

ARTISTIC EXPRESSION

Many of our examples use the moment camera to first capture a spacetime slab and then choose portions of time slices from the slab to construct a final output image. The goal has been to create a seamless result that "captures the moment." However, more artistic tools can easily be created to combine pixels in the slab in interesting ways. In Figure 10, we have modified the selection mechanism to create surprising artistic effects. Very simple criteria can be modified in real time to provide a wide variety of expressive results.

Figure 9. Spacetime slab. About one-third of a second separates these three time slices of the slab. A cliplet that holds on the first frame, plays the intervening 10 frames, then holds on the last, viscerally depicts the moment.

Future cameras might have even more advanced capabilities than those we've described. For example, cameras that notice when someone is smiling are already being developed. Future cameras could suggest better ways to frame a scene and indicate that we should back up or point the camera just a bit higher. Cameras might someday even learn our habits and help develop a style of their own based on how we use them. In our own work, we are building a moment camera prototype to continue our research in this promising new area.

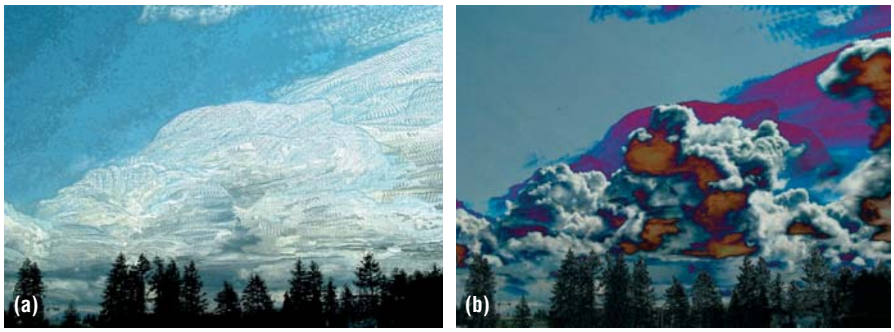

Figure 10. Artistic imaging tools. Researchers used a single time-lapse slab of clouds drifting across the sky to create these images. (a) An algorithm picked out for each pixel location in the time slice with the highest local contrast. (b) A more complex difference function of multiple time slices creates unusual colors when a channel wraps around to indicate colors above 255 or below 0.

Acknowledgments

This work represents a sampling of years of research at Microsoft Research and the University of Washington. Our colleagues who helped in this work include Aseem Agarwala, Maneesh Agrawala, Matthew Brown, Patrick Baudisch, R. Alex Colburn, Brian Curless, Mira Dontcheva, Steven Drucker, Hugues Hoppe, Daniel Lischinski, Georg Petschnigg, David Salesin, Drew Steedly, Kentaro Toyama, Matt Uyttendaele, Jue Wang, and Simon Winder.

References

1. "Qualia," *The Stanford Encyclopedia of Philosophy*, M. Tye and E.N. Zalta, eds.; http://plato.stanford.edu/archives/sum2003/entries/qualia/.
2. A. Agarwala et al., "Interactive Digital Photomontage," *ACM Trans. Graphics*, Aug. 2004, pp. 292-300.
3. M. Brown and D. Lowe, "Recognising Panoramas," *Proc. Int'l Conf. Computer Vision* (ICCV 03), IEEE CS Press, vol. 2, Oct. 2003, pp. 1218-1225.
4. S.B. Kang et al., "High Dynamic Range Video," *ACM Trans. Graphics*, July 2003, pp. 319-325.
5. A. Eden, M. Uyttendaele, and R. Szeliski, "Seamless Image Stitching of Scenes with Large Motions and Exposure Differences," *Proc. IEEE Computer Soc. Conf. Computer Vision and Pattern Recognition* (CVPR 2006), IEEE CS Press, 2006, pp. 2498-2505.
6. Y. Boykov, O. Veksler, and R. Zabih, "Fast Approximate Energy Minimization via Graph Cuts," *IEEE Trans. Pattern Analysis and Machine Intelligence*, Nov. 2001, pp. 1222-1239.
7. P. Pérez, M. Gangnet, and A. Blake, "Poisson Image Editing," *ACM Trans. Graphics*, July 2003, pp. 313-318.
8. A. Levin et al., "Seamless Image Stitching in the Gradient Domain," *Proc. 8th European Conf. Computer Vision* (ECCV 2004), vol. 4, Springer-Verlag, 2004, pp. 377-389.
9. E. Eisemann and F. Durand, "Flash Photography Enhancement via Intrinsic Relighting," *ACM Trans. Graphics*, vol. 23, no. 3, 2004, pp. 673-678.
10. G. Petschnigg et al., "Digital Photography with Flash and No-Flash Pairs," *ACM Trans. Graphics*, Aug. 2004, pp. 664-672.
11. P. Debevec, and J. Malik, "Recovering High Dynamic Range Radiance Maps from Photographs," *Proc. Siggraph 97*, ACM Press, 1997, pp. 369-378.
12. D Lischinski et al., "Interactive Local Adjustment of Tonal Values," *ACM Trans. Graphics*, to appear Aug. 2006.
13. P. Baudisch et al., "Panoramic Viewfinder: Providing a Real-Time Preview to Help Users Avoid Flaws," *Proc.* OZCHI *2005*, *ACM Int'l Conf. Proc. Series*, ACM Press, 2005.
14. A. Schödl et al., "Video Textures," *Computer Graphics*, July 2000, pp. 489-498.
15. A. Agarwala et al., "Panoramic Video Textures," *ACM Trans. Graphics*, July 2005, pp. 821-827.

Michael F. Cohen *is a principal researcher for Microsoft Research. His research interests include image-based rendering, animation, camera control, more artistic nonphotorealistic rendering, linked-figure animation, and computational photography applications. Cohen received a PhD in computer science from the University of Utah. Contact him at mcohen@microsoft.com. Further publications can be found at www.research.microsoft.com/~cohen.*

Richard Szeliski*, a principal researcher, leads the Interactive Visual Media Group at Microsoft Research. His research interests include digital and computational photography, video scene analysis, 3D computer vision, and image-based rendering. Szeliski received a PhD in computer science from Carnegie Mellon University. Contact him at szeliski@microsoft.com.*