Salient stills: Process and practice

by M. Massey
W. Bender

Unlike a photograph, which represents a discrete moment of time, a salient still reflects the aggregate of the temporal changes that occur in a moving image sequence. The salient still image may have multiresolution patches, a larger field of view, or higher overall resolution than any individual frame in the original image sequence. The salient still process is reviewed in the context of resolution enhancement, motion estimation, segmentation, and model-based coding. Applications of salient stills, including portraiture, storyboarding, and database search, are discussed. Subjects' reactions to salient still images are presented.

The application of image registration to the enhancement of image resolution, the creation of image mosaics, and the prediction of frame-to-frame correspondence for compression is an active area of research in image processing. Few researchers have applied these results to the emulation of dynamic images created by manual artists and photographers. Image registration is a potent means for creating photographic effects that can convey a sense of space, time, and motion. Variables include image size, aspect ratio, contour, element repetition, and variations in spatial resolution and image focus.

Teodosio and Bender[1] described a class of images called *salient stills*: multiple frames of an image sequence that may include variations in focal length or field of view are combined to create a single still image. The still image may have multiresolution patches, a larger field of view, or higher overall resolution than any individual frame in the original image sequence. It may also contain selected salient objects from any one of the sequence of video frames. The still can be created automatically or with user intervention.

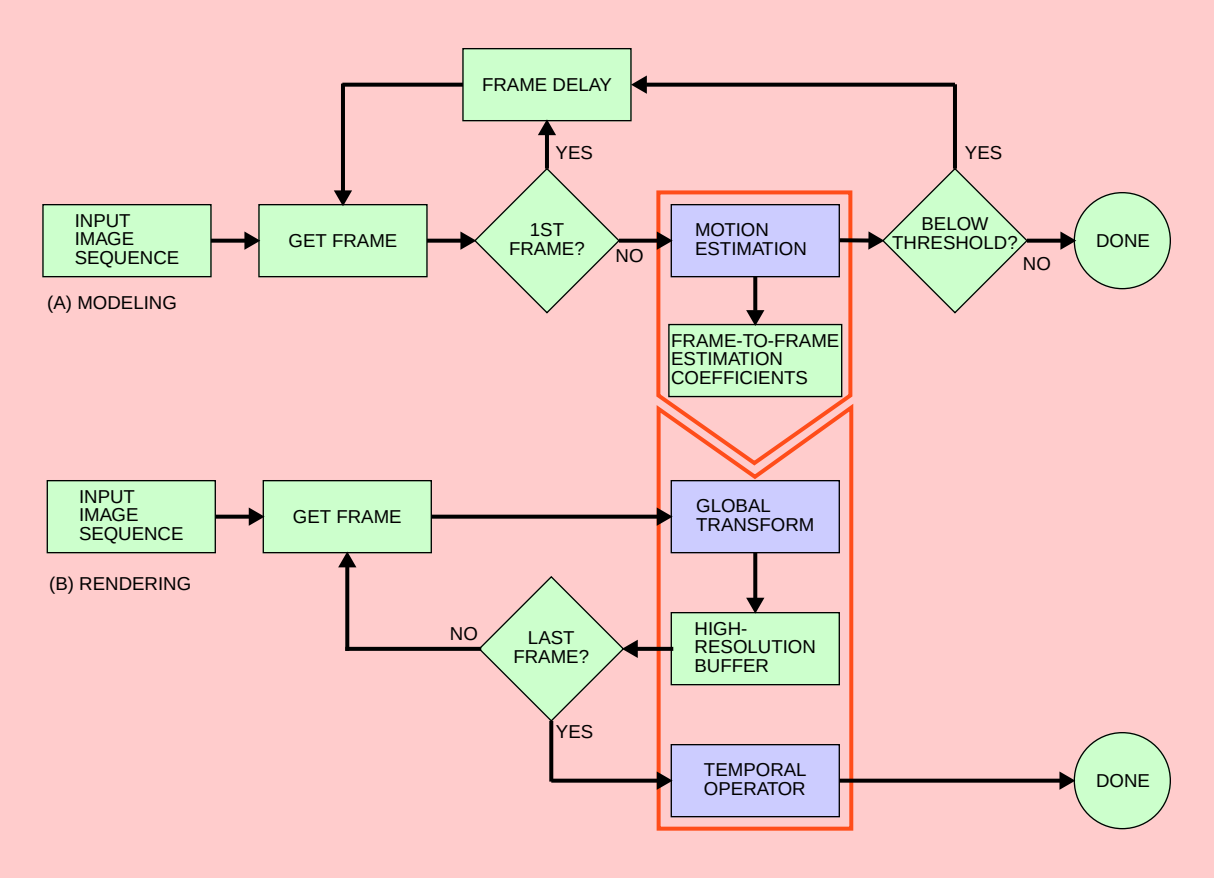
Previous work

The salient still process utilizes image representations that derive from various research interests in imaging science and computer graphics. Abdei-Aziz and Karara[2] introduced a linear algebra approach to perspective modeling to the photogrammetry and image analysis communities.[3-12] Sutherland[13] introduced algebraic methods to the field of computer graphic modeling and rendering. Heckbert[14] applied linear algebra to texture-mapping polygons in perspective.

Enhancement. Much of the work in the field of motion estimation and image segmentation addresses the problems of modeling small changes between frames as part of predictive encoding for image compression or image segmentation mechanisms for machine vision.[15–19] Similar algorithms are used in the field of image enhancement.[20-41] For the most part, whether or not the image model is inclusive of a perspective transformation, these algorithms consider only incremental changes between small numbers of images. Exceptions include McLean,[29] Teodosio and

0018-8670/96/$5.00 © 1996 IBM

Figure 1 Salient still process: modeling and rendering



Bender, Currin et al.,[34] Mann and Picard,[37] Hall,[39] Anandan et al.,[40] and Kang.[41] These latter approaches model an entire "scene" of images, while taking into account a variety of *large* camera motions.

Narrative and perspective. The salient still process utilizes narrative techniques that derive from such diverse sources as Giotto di Bondone, Paolo Uccello, the artists of the Late Heian Period in Japan, Muybridge, Marey, Duchamp, Boccioni, and Malevitch.

Giotto reintroduced the Western world to perspective. Subsequently, artists and engineers have been engaged in a study of representation on a surface of the spatial relation of objects as they might appear to the eye. However, Giotto's interest in perspective was as an expressive tool rather than a rendering tool. He used perspective to draw the subject's attention to various elements in the composition and to contain the narrative. Uccello, in his depiction of *La Bataglia di San Romano*, used perspective to show the course of events of a day on the battlefield in the space of the picture plane. The Fujiwara scroll paintings, most notably the Tale of the Heiji War Scroll, also use orthographic projection in a similar manner, spreading a temporal narrative across a panoramic view. The viewer is given a god's-eye view, infinitely far away, and is able to see the events of an entire day at once.

Visual dynamics. Photography provides a means for visualizing high-speed motion, which is difficult to perceive due to the persistence of vision. Muybridge was mechanical in his "stop-action" capture and representation of movement. Marey used his "photographic gun" in order to combine contiguous and superimposed images to depict movement in proper spatial registration.[42] Duchamp, drawing upon Marey, used an ensemble of discrete instants that flow across

the image plane in order to represent permutation and motion in painting. Boccioni and the Futurist painters were interested in creating "realistic" still images that reflected "virtual dynamism of the objects in a static state."[43] Malevitch's works are formed by careful, deliberate compositing of graphic elements that lead to the perception of "rhythmic" movement in a static image.[43]

The salient still

The salient still process involves two stages in processing: modeling and rendering. The modeling stage establishes parameters that estimate correspondence among frames in a video sequence (see Figure 1A). Individual frames are then fit to a global model of the sequence. Still images are rendered from this model (Figure 1B). In this rendering stage, once a projection is chosen, both automatic and manual methods are used to establish what portions of the image sequence are salient. Selected frames from an example image sequence are shown in Figure 2 (bottom). The result of the salient still process is shown in Figure 2 (top). The salient still process is the synthesis of imaging technology and cinematographic narrative techniques. The process utilizes representations of time and space that are sympathetic to both image and story.

Cinematic tools for storytelling

The visual composition of an image sequence may consist of: (1) camera motions, including pan and dolly shots; (2) lens effects, such as change in focal length (zoom), depth of field, and focus (focus-pull); (3) objects or characters moving relative to the frame; (4) changes in light source or shadows from moving objects or characters; and (5) effects such as fades, inserts, overlays, etc. Filmmakers have additional tools at their disposal: setting, sound, film type, shot composition and juxtaposition, editing, and acting. The current implementations of the salient still process are useful for extracting and preserving the narrative elements that embody selective focus, camera motion, and shot composition.

The zoom lens permits the camera operator to situate someone in space and then isolate details. A fast zoom accentuates action or drama, while a slow zoom serves to bring the viewer in (or out) imperceptibly during a long monologue. The focus-pull technique directs the viewer's attention to characters or actions that are spatially separate, usually at different depths relative to the camera. Selective focus can also be achieved by split-field lenses or by special lenses that rotate the plane of focus.

Pans and tilts involve camera rotations around a fixed axis perpendicular to the optical axis of the lens. One of the visual and perceptual consequences of these techniques is a change in perspective.[44] Pans can be disorienting if the scene is an extreme long shot in a small space. Vanishing points move drastically, imparting a sense of vertigo. Whenever the center of projection is not held nearly constant, i.e., the camera is not merely rotated but physically moved, objects in the foreground may occlude objects in the background. Changes in perspective occur when the camera moves in or out of the scene. Moving the camera in while zooming out can give the viewer the sensation of running down an infinitely receding corridor. When the camera moves to track a character or object in motion, the viewer has the sensation of moving along with the action.

The film director's control over what appears in the frame and how events are staged for the camera is known as *mise-en-scene*, literally "staging an action." Camera angle accentuates a particular viewpoint. Eye-level shots give a sense of presence with the action. Shots from below convey a feeling of tension or distortion. Shots from above are useful for establishing context.[45]

The video or film image is bounded by framing. The frame makes a finite slice from an implicitly continuous world. When the shot changes in a particular scene, leaving an object or actor outside of the frame, it is assumed that the object or actor is still there. Off-screen space exits in the mind's eye.[45-47]

Object or actor movement plays a variety of roles in cinematic narrative and perception. Movement can draw viewer attention to very small areas. Movement can also disambiguate depth clues for planes and volumes. Compositions that emphasize movement are "time-bound" because the viewer's glance is directed from place to place by the variety of velocities, directions, and rhythms of movement. A shot composed of discordant objects in motion is dynamic; the viewer's attention is forced from one object to another. Translating the psychophysical phenomenon of the moving image to a still representation is a challenging task. The salient still can be thought of as the visual equivalent of a shot sequence distilled to its essence.

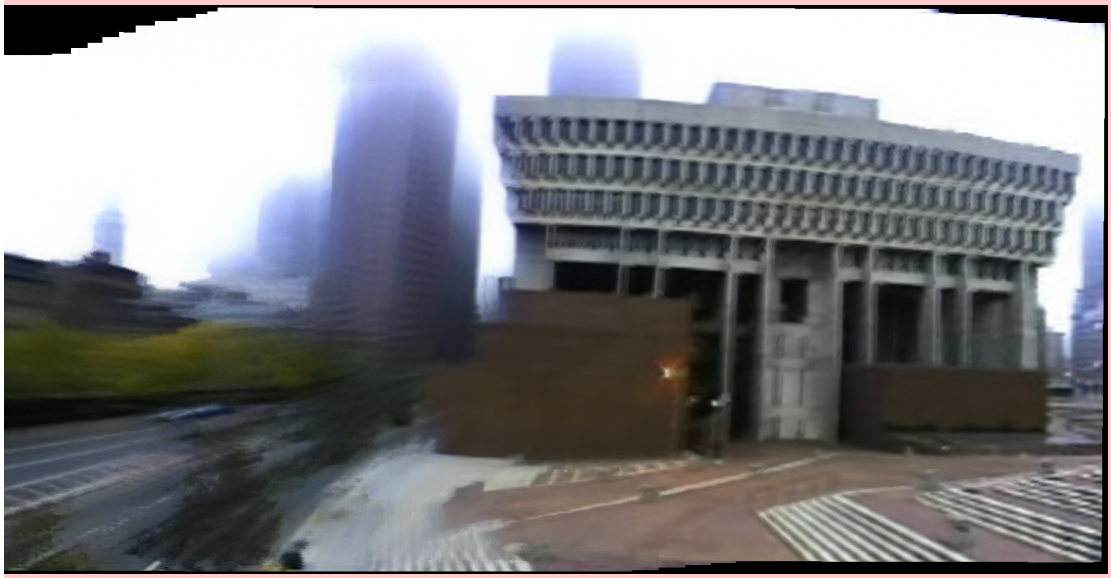
Figure 2 An example salient still

Modeling scenes

There are many choices to make in representing a scene. A general approach is to consider the changes that occur over time in the image plane of the camera. These changes are analogous to variations of the intensity distribution on the retina of our eyes. As we roll our eyes or move our head, the image on our retina changes. A scene model must quantify these changes.

A trivial representation of the image plane is that of the static shot, i.e., to assume that there are no changes over time. A scene is represented by a single frame.

A more sophisticated representation accounts for translation of the camera. Translation alone may be suitable in some situations. It is adequate for representing an extremely distant shot, where the camera is panning over a small angle on a level tripod. It is also



an adequate representation when the camera is moving over a flat surface, perpendicular to the image plane. The images in these sequences can be made coincident simply by translation.

A still more comprehensive approach utilizes an affine transformation applied to the entire image plane. An affine transformation can account for translation, scaling, rotation, or shear of the image. Affine transformation results in an orthographic projection. Thus, the camera is restricted to rendering distant objects. An entire scene is represented as a planar surface. Changes in the focal length of the lens and camera roll can be modeled as well.

The convergence of vanishing points and the commensurate nonisotropic scaling across the image plane are not accounted for by the affine transformation. A perspective projection is required to model these image attributes. There are a number of linear

approximations to the perspective projection that can facilitate estimating the correct projection.

Mathematics of the image plane

All of the aforementioned representations can be derived from the Taylor series expansion of the expression for the perspective projection model.

Simply stated, the perspective projection transformation can be written as:

$$\vec{x}' = \frac{\underline{A}\vec{x} + \vec{b}}{\vec{c} \cdot \vec{x} + 1} \tag{1}$$

where $\vec{x}'$ is the transformed coordinate, $\vec{b}$ is the affine translation, $\vec{c}$ contains the pan-tilt coordinates, and $\underline{A}$ is the affine rotation matrix:

$$\underline{A} = \begin{bmatrix} a_{xx} & a_{xy} \\ a_{yx} & a_{yy} \end{bmatrix} \tag{2}$$

The perspective projection model has eight parameters that describe the transformation completely. Higher orders of the Taylor series involve N times six parameters, where N is the order of the expansion. These expansion parameters are linear combinations of the eight adjustable parameters from the perspective projection model.

If $|\vec{c} \cdot \vec{x}| < 1$ (from Equation 1) then the Taylor series expansion amounts to the infinite series expansion of the denominator:

$$\vec{x}' = (\underline{A}\vec{x} + \vec{b})(1 - (\vec{c} \cdot \vec{x}) + (\vec{c} \cdot \vec{x})^2 - (\vec{c} \cdot \vec{x})^3 + \ldots) \tag{3}$$

where the expansion is to the nth order in $\vec{x}$. For example, the biquadratic expansion will be:

$$\vec{x}' = \underline{A}\vec{x}(1 - (\vec{c} \cdot \vec{x})) + \vec{b}(1 - (\vec{c} \cdot \vec{x}) + (\vec{c} \cdot \vec{x})^2) \tag{4}$$

which upon expanding the vector and matrix algebra gives:

$$\begin{aligned} x' &= b_x + q_{xx}x + q_{xy}y + q_{xxx}x^2 + q_{xyy}y^2 + q_{xxy}xy \\ y' &= b_y + q_{yx}x + q_{yy}y + q_{yxx}x^2 + q_{yyy}y^2 + q_{yxy}xy \end{aligned} \tag{5}$$

where:

$$q_{xx} = a_{xx} - b_x c_x, \tag{5a}$$

$$q_{xy} = a_{xy} - b_x c_y, \tag{5b}$$

$$q_{xxx} = b_x c_x^2 - a_{xx} c_x, \tag{5c}$$

$$q_{xyy} = b_x c_y^2 - a_{xy} c_y, \tag{5d}$$

$$q_{xxy} = 2b_x c_x c_y - a_{xy} c_x - a_{xx} c_y, \tag{5e}$$

$$q_{yx} = a_{yx} - b_y c_x, \tag{5f}$$

$$q_{yy} = a_{yy} - b_y c_y, \tag{5g}$$

$$q_{yxx} = b_y c_x^2 - a_{yx} c_x, \tag{5h}$$

$$q_{yyy} = b_y c_y^2 - a_{yy} c_y, \text{ and} \tag{5i}$$

$$q_{yxy} = 2b_y c_x c_y - a_{yx} c_y - a_{yy} c_x, \tag{5j}$$

and a, b, c are the affine-rotation, translation and pan-tilt parameters, respectively. Equation 5 has been utilized in a linear decomposition approach to estimating the perspective projection parameters.[37]

In the affine approximation, $\vec{c} = 0$ and the transformation takes a simpler form with only six adjustable parameters:

$$\begin{aligned} x' &= a_{xx}x + a_{xy}y + b_x \\ y' &= a_{yx}x + a_{yy}y + b_y \end{aligned} \tag{6}$$

The models discussed above are general transformations of the image plane. In "non-view-camera" photography, the image plane is fixed relative to the optical axis of the camera (the focal length of the lens may change) and camera motion is restricted to rotations about the fixed center of projection. Under these restrictions, some of the perspective parameters are unnecessary: (1) the off-diagonal elements of the affine rotation matrix that account for shear in the transformed image, and (2) the diagonal elements that account for scaling in the horizontal and vertical directions. Equation 6 can be simplified to a single rotation matrix times a scale factor:

$$\underline{A} = F \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \tag{7}$$

When restricted to a fixed center of projection, it may be better to model the image plane based solely on the

camera motion. Details of the camera model are found in Park et al.,[19] Becker and Bove,[9] Tan et al.,[11] McMillan and Bishop,[48] Aggarwal and Nandhakumar,[16] and Melen.[7]

Modeling the salient still

Motion within the discrete visual field of video (or film) may be modeled by frame-to-frame correspondences. A real-time system is constrained to sequential pairs of temporal neighbors. "Off-line," it is possible to include frames that are not necessarily adjacent in time in the evaluation of the frame-to-frame correspondences. Mann and Picard[10] call the set of frames that map to a reference frame the "video orbit" of the reference frame. The contour defined by the video orbit may be irregular. This is a consequence of the rectangular field of view of the camera changing relative position, orientation, and scale as the camera is panned. Figure 3D illustrates the video orbit from a combination of a pan and zoom.

Establishing a correspondence model is the first stage of the salient still modeling process. We have experimented with a number of methods for determining frame-to-frame correspondence (Figure 3B), including optical flow field, pyramid, block-matching, and instrumentation.

A global model. The estimations of frame-to-frame correspondence are cascaded together to construct a global model (Figure 3C). This model enables each individual frame to be mapped to each frame in the image sequence, i.e., a frame-to-scene correspondence. A three-dimensional space-time continuum is built for the video sequence. The result is a video volume where spatial location in the world is on the horizontal (H) and vertical (V) axes, and time is on the T axis (see Figure 4). A vector passing through the volume perpendicular to the first image plane will pierce the same spatial location in the world of each image. For example, the second image of a pan-left sequence is adjusted right so that the two frames line up; the second image of a zoom sequence is scaled so that it appears that all of the frames were captured at the same focal length (Figure 4).

Optical flow. Even a complex moving scene will appear as a single distribution of intensity undergoing a simple translation when viewed over a sufficiently small time and through a sufficiently small image plane. This is the basic assumption of the optical flow field.[15] The optical flow field is modeled by a continuous variation of image intensity as a function of position and time:

$$I(x, y, t) = I(x + dx, y + dy, t + dt) \quad (8)$$

Modeling an arbitrary optical flow field is indeterminate unless objects within the frame are continuous and moving slowly relative to each other. The precision of the technique is limited when used to extract camera motion from an arbitrary image sequence. However, an assumption of small displacements allows one to model the optical flow field as a two-dimensional displacement for each element of the flow field.

Figure 3 Creating a global model: (A) original sequence, (B) frame-to-frame correspondence, (C) cascading, (D) the "video orbit" with its irregular contour

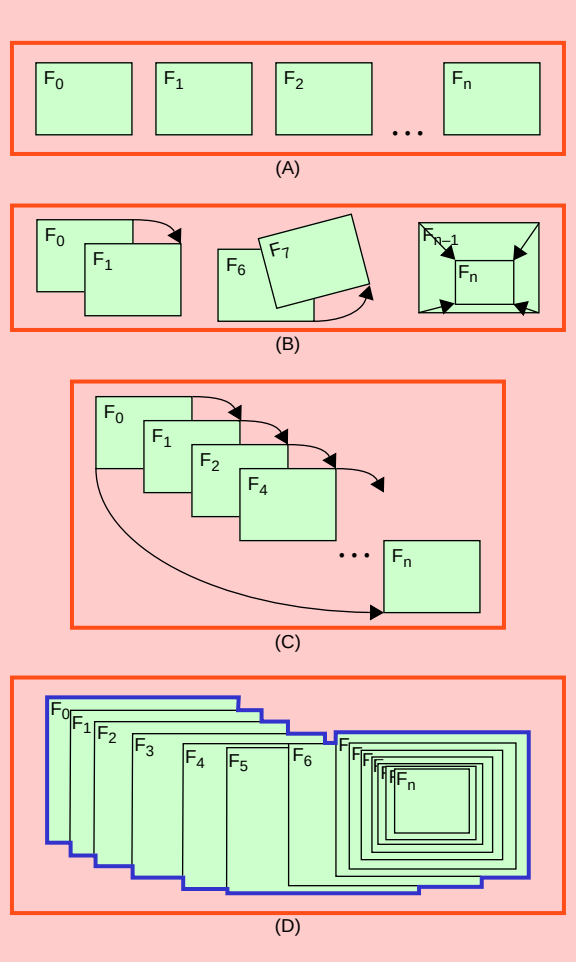


Figure 4 Frame-to-scene correspondence

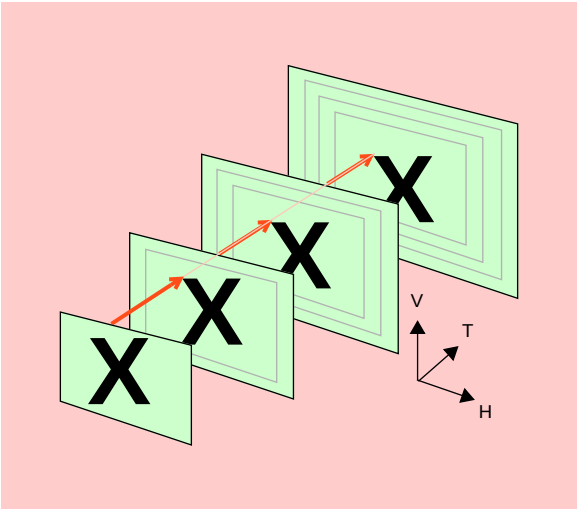
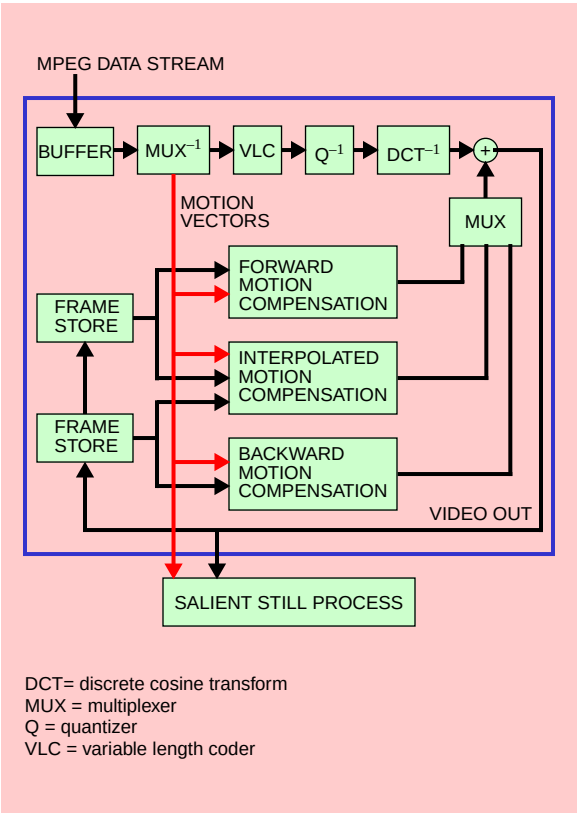


Figure 5 Extracting the motion vectors from the MPEG decoder



Pyramids. In order to guarantee convergence, optical flow must be restricted to low velocity image sequences. To circumvent this restriction, it is advantageous to sub-sample the image sequence to lower spatial resolutions before estimating the optical flow field: the optical flow between neighboring frames can be kept within the required limits by reducing the size of the overall image. The estimates made at lower resolutions are used as the initial guess when calculating optical flow fields at higher resolutions.[17]

Block-matching. Kang[41] and Kermode[49] used the vectors from the block-matching built into the International Organization for Standardization (ISO) Motion Picture Experts Group (MPEG) digital video coder to generate estimation of frame-to-frame displacements (Figure 5). These estimates are used to generate salient stills, bypassing the need for motion estimation within groups of frames (GOF). It is still necessary to use motion estimation to determine the relationship between GOFs.

Smart cameras. Relative camera motion can be measured directly. Verplaetse uses inertial guidance.[12] Motion sensing instruments attached to the camera body allow the camera to record its current position and acceleration. The data are extracted during the motion estimation processing to provide an initial guess of the motion parameters, reducing the search space of the estimation. Supplemental techniques are applied to further refine the measured parameters.

Segmentation. When objects are moving relative to the camera, motion estimation has to distinguish between camera motion and movement of characters and objects relative to the frame. Most estimation techniques do not extract discrete objects for identification. Intelligent segmentation of objects and scenes is useful for both improving the estimation of camera motion and facilitating manipulation of individual characters and objects in the rendering process.

These tools and techniques have been mentioned as an introduction to the range of current research that has been applied to the motion estimation problem. Current implementations of the salient still process have employed methods that simplify the computation, perhaps at the cost of generality.

Rendering the salient still

There are several parameters to consider in the rendering of a salient still: the frame of reference, which

frames to be rendered, the temporal operator to be applied, and how objects moving relative to the frame will be handled. Defaults can be chosen for each of these parameters, resulting in automatic rendering, or each parameter can be adjusted manually.

Frame of reference. While the global model of an image sequence establishes a possible mapping between each frame, the resulting coordinate system is relative. During the rendering process, an absolute coordinate system has to be chosen in order to map the image sequence to the output matrix. The choice of a reference frame (Figure 6), by default the middle frame of the image sequence, determines the absolute coordinate system and consequently the orientation of the resultant still.

Temporal sub-sampling. It is not necessary to include every frame used in creating the global model in rendering the still. As a rule of thumb, the more dense the temporal sampling, the more accurate the global model. For reasons such as reduced computation or storage, it may be desirable to discard or ignore frames during rendering. Frames to include (or exclude) can be chosen manually or by algorithm. Temporal sub-sampling (Figure 7A), e.g., using every fourth frame, is a crude but generally effective method. Applying a threshold on change in the global estimation parameters ensures more uniform sub-sampling (Figure 7B).

Temporal operators. It is expected that multiple frames will overlap in the global model (Figure 8). The mapping from the output image raster to the global scene model is not isometric. Statistical methods for determining a unique value at each point in the output image include: replace first, replace last, mean, mode, median, and weighted median. The first two methods place frames on top of each other sequentially, replacing pixels in regions of overlap (Figures 9A and 9B). These methods are noncausal. The other methods utilize an analysis of all pixels that map to the same point in the output raster.

The mean operator samples all the "overlapping" pixels at each point in the global scene model and outputs an average of the commensurate pixel values (Figure 9C). The net effect is to eliminate any temporal noise inherent in video. It is similar to a long exposure in conventional photography, because the photographic film is taking an average of the exposed light at any point over the entire exposure time. The difference is that the photographic image will be blurred if the camera is not stabilized, but the salient still allows the camera to move.

The mode operator outputs the "most popular" pixel value at any location in the global model (Figure 9D). The operator also results in a reduction in temporally induced noise, but the contours of objects moving relative to the global model are very noisy.

Figure 6 The choice of the frame of reference: (A) the first frame, (B) the middle frame, or (C) the last frame, determines the orientation of the still.

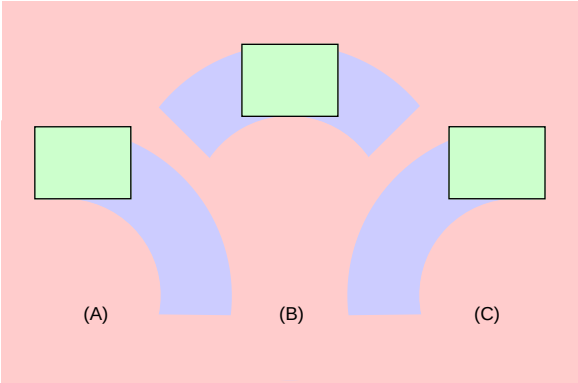


Figure 7 Temporal sub-sampling: (A) selecting every n^{th} frame (n=2), and (B) selecting a frame whenever δ > threshold

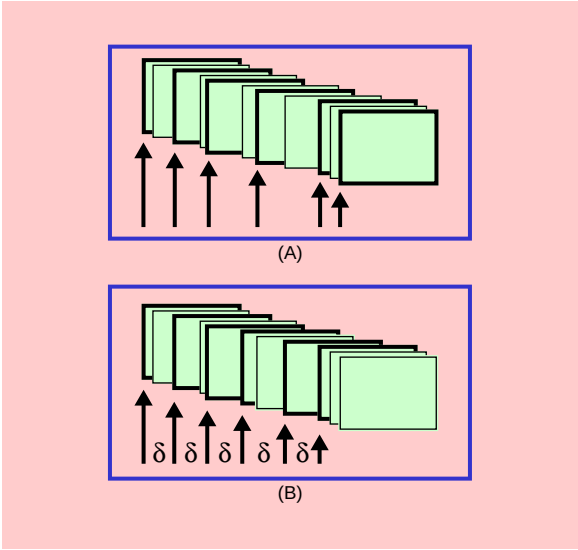
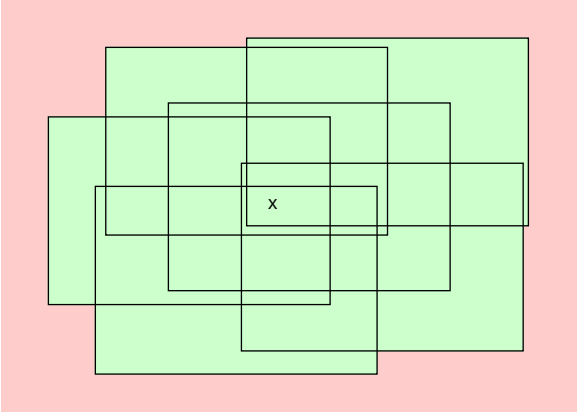


Figure 8 Point x is shared among multiple overlapping frames



The median operator has the advantages beyond the mean operator. A median filter selects only components of the global model that are correlated. Thus, the median operator is less subject to loss of detail and "bleeding" when there are objects moving relative to the global model (Figure 9E).

The weighted median operator is most useful for sequences where there is a change in focal length (e.g., zoom sequences). With such sequences, the relative resolution of the individual source frames is a function of the focal length. By applying a weighting factor to the median operator that is proportional to the inverse of the zoom factor, high-resolution patches in the output image result (Figure 9F).

Many other temporal operators are possible, including operators that examine local image features, such as dynamic range, activity, resolution, gain, and bias; and global model features, such as foreground vs background. These operators may be tied to parameters used in encoding the original image sequence, such as the control parameters used by an MPEG coder.[41] Structured, or object-based coders have the potential of providing high-level information to the salient still rendering process, enabling ready manipulation of "actors."

Applications

The salient still applications discussed below emphasize the transfer of temporally salient information rather than resolution enhancement.

Portraiture. Rembrandt was renowned for dramatic use of lighting and detail in painting faces, thus pulling the viewer's interest to those regions of the painting. To be sure, portrait painting relies on the innate human instinct to look at a person's face. Solso[50] asserted that visual attention is motivated by a variety of cognitive factors, including the interest and previous knowledge of the viewer and the context of image. Full attention is assigned to a salient feature of the image by moving the eyes in such a way as to focus that part of the image on the fovea. The detailed examination lasts only for a few hundred milliseconds, because the eye is continually moving from one region of interest to another. Eye movement studies show clearly that people spend most of their attention on the eyes and mouth of the figures in paintings, drawings, and photographs.

The salient still process is directly applicable to the creation of portraits with enhanced sharpness around the features that demand the viewer's attention, e.g., the subject's face. High-resolution regions can be added to the still image by making judicious use of zooms when shooting the input video sequence. The typical duration of a video sequence for this application is less than two seconds (less than 60 frames). This is the time that is necessary to mechanically adjust the focal length of the lens. Resolution is limited to about 640 pixels over the width of the subject's face. For a full-body shot enlarged to 8 by 10 inches, this amounts to approximately 300 dpi (dots per inch) effective resolution around the face. The effective resolution falls off rather quickly; there are approximately 50 dpi at the edges. The resulting image appears to have been shot with a shallow depth of field, since the face is sharp, but the rest of the image is relatively soft (Figure 10). The image is distinct from the photographic analog because the sharpness appears only in a small region. The entire focal plane of a photograph taken with a shallow depth of field is sharp.

Storyboarding. The salient still process has been applied to the generation of a comic book based upon two episodes of the popular television series, *NYPD Blue*. The comic book format is used as a medium for transcoding the video to text and expressive still imagery. *NYPD Blue* was an ideal source for this project because the cinematography relies on short, fast camera motion. One hundred forty carefully edited video sequences, ranging in duration from 10 to 120 frames, were digitized. The editing was geared toward maintaining a coherent narrative while utilizing sequences that maximize camera motion.

Figure 9 Temporal operators: (A) first, (B) last, (C) mean, (D) mode, (E) median, (F) weighted median, and (G) the resultant still

Figure 10 Salient still portrait of Marvin Minsky

The sequences were first batch-processed under similar constraints: translation only, reference to the middle frame of the sequence, and use of the temporal median filter. Adjustments were made after reviewing the results. Both in order to avoid excessive motion by the actors and to remove segments that do not obey affine constraints, some of the sequences were edited more tightly. The final weighting of selected frames was adjusted in order to emphasize narrative or aesthetic qualities: apparent motion cues, multiple imagery, and blur. Groups of individual stills were laid out as pages, often using the irregular framing as a narrative device. Dialog was added manually. Although a page from the comic book cannot be reprinted in this paper, a sample of a salient still resulting from the same process is shown in Figure 11.

Database search. Temporal media such as video must be viewed in order to be evaluated. A still image representation of a video sequence is only useful to the extent that it conveys information about the sequence. Much can be surmised from individual salient stills, including actor and camera motion. A moving character may appear blurred or in multiple images. A pan between two actors or tilt of the camera is evident as an extended or irregular framing. Salient still images facilitate the access of video from databases and over data networks.

Photo illustration. Conventional photography was considered vulgar and unworthy of artistic merit when it was introduced to the art world in the 19th century. The critics contended that photography was merely a literal mapping of the physical world, a technical manifestation devoid of expressive creativity. Some argued that the new photography was artificial because the human visual system could never perceive an entire scene at once. Others complained that the frozen images resulting from a fast shutter speed were unnatural because the eye did not perceive a discrete moment in time.[42] But photography has evolved in myriad ways, technically as well as aesthetically. The salient still is a subset of photography. It can mimic photographic special effects or result in wholly unique imagery.

An effect similar to Marey's Chronophotography[42] can be achieved by rendering selective portions of a video sequence. Individual frames can be emphasized by manipulating the parameters of the weighting function used in the temporal median operator. Masking regions within discrete frames can also be used for emphasis and visual dynamics. Once the estimator stage has placed the characters in their proper relative spatial positions, the illustrator can accurately clone temporal doubles across a scene. Furthermore, the sequence can be directed in such a way that the actors

Figure 11 Salient still storyboard. Irregular framing results from camera motion.

appear quite naturally in different locations at different times (Figure 12).

Evaluation

Valva[51] performed a preliminary study to consider reactions to images made with the salient still process. The study posed several questions: Are people disturbed by the perspective distortions found in some images? Is the region-specific sharpness inherent in these images confusing? The study also sought to determine if subjects could glean more or less information about the story than they could otherwise get from a conventional still image.

Valva's hypothesis was that salient stills are perceived as photographs. Subjects are distracted by areas of poor image quality that are due to excessive interpolation of low-resolution regions of the image. High resolution areas force the subject to concentrate on the sharpest area of the image regardless of its particular content, in a manner similar to a short depth of field.

Methodology. The survey of five people, who varied in age and visual arts experience, was anecdotal and qualitative. Twenty images were mounted on black boards (reference Figure 13). Participants were asked to describe what they thought was the main point or story line of each image. They were instructed to:

1. Indicate the part of the image that portrayed the main idea or most significant part of the image
2. Write a caption for the image
3. Describe the image as if to a person who had lost his or her sight

Results. Subjects were consistently drawn to the area of the image that was the sharpest. They were inter-

Figure 12 Salient still "Chronophotograph"

ested in the human facial features, even if those features were blurry. Also, areas of bright color and high contrast elicited a response in the subjects. In one image, the sharpest part of the image occurs in the background, while the foreground is dominated by bright yellow chairs. Everyone in the group marked the lower-resolution, bright chairs as the most important part of the image. In some cases, interest was drawn toward areas where the characters in the image were directing their gaze. The center of the image was important to the subjects. The parts of the image that elicited the perception of motion also drew the subjects' attention. The subjects' response to these areas seemed to be dependent on their experience in the visual arts.

The more visually literate subjects tended to disparage some images. The subject most well trained in photography found the overall image quality of salient stills inferior as compared with conventional photography. When this subject considered a low-resolution region of the image as important, then the image as a whole was considered to be out of focus. "But it is a very poor picture so it's hard to see … The picture is very blurred … It is an interesting photo, but it is out of focus."

Participants who were less experienced with photography paid more attention to image content than image resolution, except when the lack of resolution grossly interfered with interpretation. This was especially true with regard to facial expression. All subjects looked for facial expression and interpreted the images according to the social interactions of the people depicted. In some instances, facial expression attracted the subject's attention more than the region that was of highest resolution. Subjects were consistently disturbed by images where they could not recognize people or decipher their facial expression, as depicted in this response: "I circled his face. I was considering circling the head of lettuce … the lettuce seems to be more in focus, but I think the human face since birth is what people tend to focus on since that's what the first object of focus is for newborns in their early … development. I just find his face more interesting than the head of lettuce. That's where my eye is pulled even though [there are] brighter parts or more in focus."

Blurring occurs in the test images either because of motion in the source video or because of interpolation of low-resolution regions. Subjects were disturbed both when a person's face was more blurry than the rest of the image and when a person's features were distorted. "I immediately noticed that there was a difference in the lighting and my eye goes to his eye because one was out of focus and one was focused." By far, the most distortion in the test images arises from rendering motion, rather than interpolation. As during a long, tripod-mounted exposure in photography, stationary features remain sharp and moving objects are blurred.

Some subjects noticed artifacts of the salient still process such as irregular borders. These were interpreted by one subject as an obstruction in front of the camera. Another subject described the edges as "weird framing." Still, some subjects did not comment on the jagged edged borders at all.

Discussion. The small sample size and lack of control images, i.e., conventional photographs, were the main shortcomings of this study. Although not scientifically rigorous, this work provided some observations of the way that people perceive salient still images.

Subjects naturally directed their gaze to areas of high contrast, bright colors, and what was in focus. However, this reflex competed with the stronger impulse to scan a human face. Even the most abstract forms can

be seen as a human face.[47,50,52] In many cases it should be sufficient to render the human face with more effective resolution than the rest of the scene. Of course, it is not always possible to extract the necessary resolution from standard video. In general, excellent results are achieved with zoom sequences.

It can be argued that Valva misstated the problem by evaluating salient stills within the context of photography. Her study could have asked: How does a single frame of video compare to a salient still? But people, accustomed to viewing photographs, seem to have an expectation that still images should have all the detail and sharpness that film affords. To date, the salient still cannot compete with the superior image quality of conventional photography, but it can improve and embellish video stills.

Conclusion

Photography spans the range of quality from poorly composed, blurry scenes of a family outing to the wonderfully detailed expressive images of the great masters. Conventional photography can depict a (decisive) moment, which is captured in only an instant, then ceases to exist. It can reveal a mood, not so much an event, as in a portrait that spans some time. And photography can show something timeless, as in a scene or location, which may appear at a certain time of day, under certain lighting, but which is relatively invariant over time. Salient stills lie somewhere in the regime of the latter two.

The principal aim of this paper was to acquaint the reader with the art and technology of salient stills. The mathematics of the image-plane projection were introduced to illustrate the various levels of complexity associated with the estimation problem. Rendering the salient still was discussed with an emphasis on creative control.

Media transcoding is the process of translating from one medium to another. In the case of video-to-still-image transcoding, there are two problems that need to be addressed: resolution enhancement (video puts resolution in time, while stills put resolution in space) and narrative (the language of cinematography is different from the language of photography). The salient still facilitates the transcoding of both the content and context of the video story. It provides an automated tool for compositing the individual frames of a video sequence into a single still image that portrays the camera motion and the relative position of the subjects in the image.

Figure 13 Garlic image from Valva's study

There has been much resistance by the news industry to use salient still technology. Ritchin[53] argues for the need to distinguish between images that come directly from the capture device, and those that have been electronically manipulated. Reaves[54] further argues that the use of photo illustrations in news stories is inappropriate. "Mixing photo illustration into a news story places the unnecessary burden on the reader for making the appropriate cognitive switch to 'symbolic' interpretation as envisioned by the editor." Max Frankel of *The New York Times* has a more balanced view. While acknowledging the need to wait for the "next generation" of editors before seeing a salient still accompanying a news article, he said,[55] "It is like a reporter using a quote."

It is the authors' view that the utility and credibility of an image lies in the hands of the image creator and editor. The technology is neutral when it comes to truth. To the extent to which salient still technology is used to distort temporal and spatial relationships, it has the potential of harm. Its use as a tool that provides context to temporal and spatial relationships is beneficial.

Acknowledgments

This work was supported in part by the News in the Future research consortium at the MIT Media Laboratory and International Business Machines Corporation.

Cited references

1. L. Teodosio and W. Bender, "Salient Video Stills: Content and Context Preserved," *ACM Multimedia*, Anaheim, CA (1993).
2. Y. I. Abdei-Aziz and H. M. Karara, "Direct Linear Transformation from Comparator Coordinates into Object Space Coordinates in Close-Range Photogrammetry," *Proceedings ASP/UI Symposium on Close-Range Photogrammetry*, Urbana, IL (1971).
3. S. Ganapathy, "Decomposition of Transformation Matrices for Robot Vision," *Proceedings of the 1st IEEE Conference on Robotics*, Atlanta, GA (1984).
4. B. K. P. Horn, *Robot Vision*, MIT Press, Cambridge, MA (1986).
5. R. Y. Tsai, "A Versatile Camera Calibration Technique for High-Accuracy 3-D Machine Vision Metrology Using Off-the-Shelf TV Cameras and Lenses," *IEEE Journal of Robotics and Automation (*1987).
6. S. F. Ray, *Applied Photographic Optics*, Focal Press, London (1988).
7. T. Melen, "Extracting Physical Camera Parameters from 3 by 3 Direct Linear Transformation Matrix," *Optical 3-D Measurement Techniques II*, A. Gruen and H. Kahmen, Editors, Herbert Wichmann Verlag GmbH (1993).
8. R. Szeliski, "Image Mosaicing for Tele-Reality Applications," Digital Equipment Corporation, Cambridge Research Laboratory Technical Report Series CRL 94/2 (May 1994).
9. S. Becker and V. M. Bove, "Semi-Automatic 3-D Model Extraction from Uncalibrated 2-D Camera Views," *Proceedings of SPIE Image Synthesis (*1995).
10. S. Mann and R. W. Picard, *Video Orbits of the Projective Group: A New Perspective on Image Mosaicing*, MIT Media Lab Perceptual Computing Section Technical Report No. 338, Cambridge, MA (1995).
11. Y. P. Tan, S. R. Kulkarni, and P. J. Ramadge, "A New Method for Camera Motion Parameter Estimation," *IEEE International Conference on Image Processing (ICIP)*, Washington, D.C. (1995).
12. C. Verplaetse, "Intertial Proprioceptive Devices: Self-Motion-Sensing Toys and Tools," *IBM Systems Journal* **35**, Nos. 3&4, 639–650 (1996, this issue).
13. I. E. Sutherland, "Three-Dimensional Data Input by Tablet," *Proceedings of the IEEE* (1974).
14. P. Heckbert, "Survey of Texture-Mapping," *IEEE Computer Graphics and Animation* (1986).
15. B. K. P. Horn and B. G. Schunk, "Determining Optical Flow," *AI* **17** (1981).
16. J. K. Aggarwal and N. Nandhakumar, "On the Computation of Motion Sequences of Images—A Review," *Proceedings of the IEEE* **76**, No. 8 (1988).
17. J. Bergen and R. Hingorani, *Hierarchical Motion-Based Frame Rate Conversion*, Technical Report, David Sarnoff Research Center, Princeton, NJ *(*April 1990).
18. J. Bergen, P. Burt, R. Hingorani, and S. Peleg, *Three-Frame Technique for Analyzing Two Motions in Successive Image Frames Dynamically*, U.S. Patent 5,067,014 (1991).
19. J. Park, N. Yagi, K. Enami, K. Aizawa, and M. Hatori, "Estimation of Camera Parameters from Image Sequence for Model-Based Video Coding," *IEEE Transactions: Circuits and Systems for Video Technology* **4**, No. 3, 288–296 (June 1994).
20. Y. Suenaga, "Super-Resolution: Getting a Sharp Image from a Set of Multiple Frames," unpublished paper, MIT Media Laboratory, Cambridge, MA (1982).
21. W. Hannan, *Imaging System with Enlarged Depth of Field*, U.S. Patent 4,404,594 (1983).
22. P. Burt, et al. "The Laplacian Pyramid as a Compact Image Code," *IEEE Transactions on Communications* (1984).
23. P. Bennett and S. Gabriel, *Spatial Transformation System Including Key Signal Generator*, U.S. Patent 4,463,372 (1984).
24. B. Ferren, *Spatial Imaging System*, U.S. Patent 4,584,704 (1986).
25. E. Adelson, *Depth-of-Focus Image Processing Method*, U.S. Patent 4,661,986 (1987).
26. W. Glenn, *Television Camera and Recording System for High Definition Television Having Imagers of Different Frame Rate*, U. S. Patent 4,652,909 (1987).
27. P. Burt, *Pyramid Processor for Building Large-Area, High-Resolution Image by Parts*, U.S. Patent 4,797,942 (1989).
28. J. J. Campbell, Y. C. Faroudja, and T. C. Lyon, *Television Scan Line Doubler Including Temporal Median Filter*, U.S. Patent 4,967,271 (1990).
29. P. McLean, "Structured Video Coding," master's degree thesis, MIT, Cambridge, MA (1991).
30. M. Irani and S. Peleg, "Improving Resolution by Image Registration," *CVGIP: Graphical Models and Image Processing* **53**, No. 3 (1991).
31. A. M. Tekalp, M. K. Ozkan, and M. I. Sezan, "High-Resolution Image Reconstruction from Lower-Resolution Image Sequences and Space-Varying Image Restoration," *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (1992).
32. R. Ginosar, O Hilsenrath, and Y. Zeevi, *Wide Dynamic Range Camera*, U.S. Patent 5,144,442 (1992).
33. R. Tinkler, *System and Method for Fusing Video Imagery from Multiple Sources in Real Time*, U.S. Patent 5,140,416 (1992).
34. B. L. Currin, A. A. Abdel-Malek, and R. I. Hartley, *Forming, with the Aid of an Overview Image, a Composite Image from a Mosaic of Images*, U.S. Patent 5,187,754 (1993).
35. P. Migliorati, F. Pedersini, L. Sorcineli, S. Turbaro, "Semantic Segmentation Applied to Image Interpolation in the Case of Camera Panning and Zooming," *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (1993).
36. K. Aizawa, T. Komatsu, T. Saito, and M. Hatori, "Subpixel Registration for a High Resolution Imaging Scheme Using Multiple Imagers," *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (1993).
37. S. Mann and R. Picard, "Virtual Bellows: Constructing High Quality Stills from Video," *IEEE Image Proceedings*, Austin, TX (1994).
38. H. S. Sawhney, S. Ayer, and M. Gorkani, "Dominant and Multiple Motion Estimation for Video Representation," *Proceedings of IEEE International Conference on Image Processing*, Volume 1, Washington, D.C. (October 1995), pp. 322–325.
39. J. Hall, *Imaging Apparatus for Providing a Composite Digital Representation of a Scene Within a Field of Regard*, U.S. Patent 5,394,520 (1995).
40. P. Anandan, M. Irani, R. Kumar, and J. Bergen, "Video as an Image Data Source: Efficient Representations and Applications," *Proceedings of IEEE International Conference on Image Processing*, Volume 1, Washington, D.C. (October 1995), pp. 318–321.
41. J. Kang, "Generating Salient Stills Using Block-based Motion Estimation," master's degree thesis, MIT, Cambridge, MA (1995).
42. A. Scharf, *Art and Photography*, Penguin Books, London (1968).
43. R. Carrieri, *Futurism*, Edizioni Del Milone, Milan (1966).
44. *The Perception of Pictures*, Volume 1, M. A. Hagen, Editor, Academic Press, New York (1980).
45. W. Eisner, *Comics & Sequential Art*, Poorhouse Press, Tamarac, FL (1985).

46. S. Katz, *Film Directing Shot by Shot*, Michael Wiese Productions, Studio City, CA (1991).
47. S. McCloud, *Understanding Comics*, A Kitchen Sink Book for Harper Perennial, New York (1993).
48. L. McMillan and G. Bishop, "Plenoptic Modeling: An Image-Based Rendering System," *Proceedings ACM SIGGRAPH 95*, Los Angeles, CA (August 1995), pp. 39–46.
49. R. G. Kermode and A. B. Lippman, "Coding for Content: Enhanced Resolution from Coding," *1995 IEEE International Conference on Image Processing* III, Washington, D.C. (October 23–26, 1995), pp. 460–463.
50. R. Solso, *Cognition and the Visual Arts*, MIT Press, Cambridge, MA (1994).
51. A. Valva, "Qualitative Research on Salient Still Technology," unpublished paper, MIT Media Laboratory, Cambridge, MA (1995).
52. E. H. Gombrich, *The Image and the Eye*, Cornell University Press, Ithaca, NY (1982).
53. F. Ritchin, "Electronic Word," *Wired* **3**, No. 1 (1994).
54. S. Reaves, "The Unintended Effects of New Technology (and Why We Can Expect More)," *News Photographer* **50**, No. 7, National Press Photographers Association, Durham, NC (1995).
55. M. Frankel, conversation with author (1995).

General references

E. H. Adelson and J. R. Bergen, "The Plenoptic Function and the Elements of Early Vision," *Computational Models of Visual Processing*, Chapter 1, M. Landy and J. A. Movshon, Editors, MIT Press, Cambridge, MA (1991).

J. J. Gibson, *The Ecological Approach to Visual Perception*, Houghton Mifflin Co., Boston, MA (1979).

D. Hockney, *Hockney on Photography: Conversations with Paul Joyce*, J. Cape, London (1988).

R. Lowell, *Matters of Light and Depth*, Broad Street Books, Philadelphia, PA (1992).

B. Newhall, *The History of Photography: From 1839 to the Present*, Museum of Modern Art, New York (1982).

W. S. Rubin, *Dada and Surrealist Art*, Harry N. Abrams, Inc., New York (1968).

Accepted for publication April 8, 1996.

Michael Massey *MIT Media Laboratory, 20 Ames Street, Cambridge, Massachusetts 02139-4307 (electronic mail: mmassey@media.mit.edu).* Dr. Massey is a recent graduate of MIT's Media Lab, where he learned digital video production and applied video image processing techniques to create expressive still images. He has worked as a freelance photojournalist and documentary photographer for ten years. He received a Ph.D. in Raman spectroscopy and solid state physics from the University of Michigan in 1992.

Walter Bender *MIT Media Laboratory, 20 Ames Street, Cambridge, Massachusetts 02139-4307 (electronic mail: walter@media.mit.edu).* Mr. Bender is a principal research scientist at the MIT Media Laboratory and principal investigator of the laboratory's News in the Future consortium. He recieved the B.A. degree from Harvard University in 1977 and joined the Architecture Machine Group at MIT in 1978. He received the M.S. degree from MIT in 1980. Mr. Bender is a founding member of the Media Laboratory.

Reprint Order No. G321-5624.