Storytelling with Salient Stills

Michael J. Massey

Ph.D., Physics
The University of Michigan, Ann Arbor, MI
1992

B.A., Physics
Oberlin College, Oberlin, OH
1983

Submitted to the Program in Media Arts and Sciences, School of Architecture and Planning, in partial fulfillment of the requirements for the degree of
Master of Science in Media Arts and Sciences
at the Massachusetts Institute of Technology

September 1996

Author .....
Michael J. Massey
August 8, 1996

Certified by .....
Walter Bender
Principal Research Scientist, MIT Media Laboratory
Thesis Supervisor

Accepted by .....
Stephen A. Benton
Chairman, Departmental Committee on Graduate Students
Program in Media Arts and Sciences

Storytelling with Salient Stills

Michael J. Massey

Submitted to the Program in Media Arts and Sciences, School of Architecture and Planning, on August 8, 1996, in partial fulfillment of the requirements for the degree of Master of Science in Media Arts and Sciences at the Massachusetts Institute of Technology

Abstract

This thesis explores the application of machine vision technology to creating digital photographs from video sequences. This class of photographs, called "salient stills" can have various aesthetic looks depending on the particular video sequence from which they are derived, or from user specified estimation or rendering controls. As a narrative medium, photography is different from cinema. Cinema or video storytelling relies on the evolution over time to convey the story. The context of the story can be presented before, during, or after the content of the story is revealed. In conventional photography, the single image does not usually provide contextual information. The inherent ambiguity of conventional still photography is perhaps what makes photography so interesting to look at: It is an unnatural process to view time standing still. The salient still incorporates elements from many video frames, compositing them in a controlled statistical fashion. Thus, the salient still can preserve the context of video while presenting the content of a visual story.

Thesis Advisor: Walter Bender
Title: Principle Research Scientist, MIT Media Laboratory

The work reported herein was supported in part by a grant from International Business Machines and by the News in the Future Consortium.

Storytelling with Salient Stills

Michael J. Massey

The following people served as readers for this thesis:

Reader .....

Glorianna Davenport
Principal Research Associate
MIT Media Laboratory

Reader .....

José Azel
President
Aurora and Quanta Productions

Acknowledgments

Thanks to Walter for allowing me the opportunity and freedom to learn and explore the myriad tools and toys at the Media Lab. It was great fun!

I want to acknowledge Shawn Becker for his influence on my research and attitude on life and family over the last two years. I really enjoyed our conversations.

This thesis would not have become a reality without the support of Claudia Green. You are my shining light.

for Liana Rose

Contents

List of Figures

Chapter 1

Narrative Photography

1.1 Photography and the Photographer

When we look at a photograph, we can study the entire image. Our eyes dart across the image at its leisure, stopping for only a third of a second here and there. Where does it stop? How is the story "created" or perceived as we study a still photograph? Real images are often not perceived, but later revealed in a photograph. The role of the photographer is to edit reality; to anticipate the right moment, framing, lighting, and point of view in which to capture the essence of the idea or event he or she is trying to convey. As related by Wilson Hicks in his seminal book, *Words and Pictures*, the camera sees everything, but the photographer creates the image.

> Photojournalism demands of the camera that what it sees it see better than the eye; that, unlike the eye, it miss nothing important. To these demands the camera, inanimate object though it is completely capable of acceding. But photojournalism goes further. It demands that the camera be selective, or as selective as it can be made to be; that it discern as well as see, that it lay hold for what it sees with understanding. Obviously, with these latter demands the camera cannot cope, alone. This is where the photographer comes in, to exercise what discretionary powers are open to him in making use of the machine.[22]

These words are as applicable to digital photography in general and salient stills in particular as they are to conventional photography. To be sure, technological advances have provided the photographer a greater horizon. The camera is at once more forgiving and extensive in possibility. Still, it is the assessment of the situation, and anticipation thereof that no machine can provide.

As you will see later in this thesis, the salient still is a photograph. Suffice it to say that salient stills are a tool for digitally creating a variety of images that photographers have experimented with since the beginning of chemical photography. In much the same way that advances in camera technology have expanded the photographers creative potential *in the camera*, salient stills can expand the possibilities for post-production enhancement or special effects. What makes the salient still a special class of photograph is its ability to project the temporal evolution of a shot onto a single image.

1.2 Context

All visual narrative requires some form of context with which the viewer can make sense out of the what is seen. That context could be provided by as abstract a source as the memory of the viewer herself. "We tend to project life and expression onto the arrested image and supplement from our own experience what is not actually present."[26] In that extreme, the artist relies on some common experience or archetype to tie the piece together. On the other end of the spectrum is the completely literal narrative, exemplified by live news broadcasts. Photography is rather different from text or full motion video with sound.

Hicks thesis is that images and words complement each other, providing the reader infinitely more information than either alone. In this model, the basic narrative unit consists of a single captioned image. The caption provides just enough context to the image to tell a whole story. Either can logically be expanded. A sequence of captioned photographs makes up a photo essay.[22] This is precisely the working model for print photojournalism today. Hicks considers the text to be a sound or voice that is heard when it is read. Of course, what is lacking is the intonation, accent, and nuances of the speaker's actual voice or ambient sound. Current advances in multimedia reportage including salient stills provide the reader with the experience of audio captioning.

1.3 Representation

Narrative necessitates communication of a representation of human perception. But, what is a representation? Literally, "represent" is to re-present to an observer what was present in his or her visual sense on an earlier occasion. A painter though, can depict something never seen but only learned or imagined. Words as a representation can describe objects, places, animals, emotions and events, but are certainly not perceived. Perceptual psychologists[24,25,26] have approached the issue of representation in art and visual perception. M.A. Hagen discusses the relationship between actual surfaces and objects and their representation. A blank surface can be decorated, regularized textured, painted or embellished without producing a picture. The resulting pattern need not have any referential meaning. This is the distinction between decoration and depiction. Animals can perceive surfaces, only humans make and perceive pictures and symbols so as to communicate. Real surfaces are perceived in the course of human development by maturation and learning taken together, by encountering the surfaces in the natural environment, without being taught. On the other hand, referential meanings of marks on a surface get apprehended by children differently. At one extreme, photographs are independent of cultural conventions. Drawings and diagrams are somewhat conventional. Alphabetic writing is wholly conventional. A picture succeeds as a representation of ordinary objects and scenes because it contains the same kind of information for determinate perception as is provided by the light reflected from the ordinary environment.[25]

Photographs provide a partial second-hand perception. Does one have to read a picture as one has to read a script, or can one perceive a picture without learning? The Pioneer spacecraft were launched out into the galaxy carrying a videodisc containing hundreds of photographs of life on planet earth. The project designers, Carl Sagan in particular, believed that any sentient being that discovered the space probe would be able to perceive

our world through the literal representation afforded by pictures. Perhaps they would not be able to decipher our written language. Besides, any description would be in vain without a shared experience. But still photographs alone may not be sufficient to provide the full story.

The reading of spatial relationships on a canvas applies to the reconstruction of temporal relationships. We cannot estimate the passage of time in a picture without interpreting the event represented. Thus, representational art begins with the indications of meaning rather than with rendering nature. Narrative art makes use of symbolic gestures to convey the meaning of an event. The successful illustration of a narrative will always suggest and facilitate supplementing the past and anticipation of the future, the scanning backward and forward in time that comes from understanding an action.

> The photographic moment can be an illicit extrapolation of actions that unfold in phases. It can represent events which are assembled in one memory span and subordinate to one central perceptible meaning or events separated by hours or even years.[26]

The snapshot draws attention to the paradox of capturing life in a still, i.e., that we may never be aware of an arrested moment within the flux of events. According to Gombrich, herein lies the artificiality of pictorial art: The confinement of information to simultaneous cues.

> The cinematic shot or sequence succeeds when the snapshot fails. If it catches a person sneezing the sequence explains the resulting grimace which the corresponding snapshot may leave uninterpretable. The miracle is not that some snapshots capture an uncharacteristic act, but that the photograph or painting can abstract from movement and still produce meaningful expression.[26]

It is the lack of context that lends itself to this interpretation of still photography. Of course, a simple caption or title can resolve the ambiguity inherent in the arrested image.

The real challenge of any still photograph is to provide sufficient visual clues to be read without a caption. In some cases, salient stills can provide the tools to synthesize such an image.

In the introduction to Franz Masereel's 1919 woodcut novel, *Passionate Journey,* Thomas Mann warns that the reader may need multiple readings to understand the story. "Ambiguities in a single image require taking in pictures after and pictures before to understand the context of a particular image."[11] Each woodcut in *Passionate Journey* depicts a unique event in the protagonist's life. Taken alone, the images are simple and stark, crudely angular and graphic. Their true impact emerges, however when read in sequence.

1.4 Anti-Narrative

It has been argued among the critics that still pictures alone do not constitute narrative. McCloud expands upon Eisner's definition of comics as spatially juxtaposed *sequential* art.[6,10] Sontag says photography does not provide sufficient context to tell a story over time.[14] Gombrich argues that static signs can only represent static moments, never a moment which happens in time. "As soon as we assume a moment in time with no motion, movement as such becomes inexplicable."[26] Thus Zeno's Paradox: Achilles is in a race with the tortoise. The tortoise is given a head start. Once the faster Achilles has reached the previous position of the tortoise, the tortoise has already advanced. Achilles will never catch up to the tortoise no matter how fast he runs.

I believe a single still picture can tell a story, however short a story it may be. Context can be supplied by a caption or title or the viewer can read into the story whatever he sees. It is our expectation which derives from memory and experience that allows us to read the action in a still picture. The photographic moment is not still, but stopped.

> Photography's relation to time is also more complex than is often assumed. Roland Barthes wrote of photographs as representing the deaths of moments of life, while André Bazin believed that "photography ... embalms time, rescuing it simply from its proper corruption." ... Yet the Italian photographer Luigi Ghirri believed that each photograph represents the possibility of a renewal of the gaze, something which allows for the narrative possibilities of photographs to be explored.[16]

Conventional photography's inherently static nature is unnatural. Foveated vision prevents us from seeing an entire scene at once. We can only gaze at a small area at any instant. Visual perception is dynamic. Photography allows us to concentrate on a past moment in full detail. Eadweard Muybridge showed that a sequence reveals what was heretofore unperceived. Etienne-Jules Marey's cronophotography, while not as well known as Muybridge's studies, showed the progression of movement on a single film plate.

The difference between Muybridge and Marey is essential. Marey saw movement as a synthesized moment. He kept faithful to the notion of a point of view. He wanted to preserve the unity of the single image and realized the latter in relation to the bird in flight, in which he analyzed and synthesized the flight condensed in time. (Figure 1.1). On the other hand, Muybridge used a battery of plates, each one representing one instant of movement. In each unique plate, he kept the decomposition of a trajectory in successive movements. Utilizing various viewpoints, Muybridge was able to represent motion from a variety of different perspectives.

1.5 Subjective Narrative and Digital Manipulation

The conventional expectation is that the photo narrative is an accurate representation of actual events. (Mimetic narrative). But the photojournalist, as well as the documentary filmmaker must deal with the subjective nature of the profession. Fred Ritchin argues that

it is precisely the expression of a photographer's point of view that makes the narrative compelling. His point is a commentary against editorial manipulation of the story content. He argues that the facility of digital manipulation could create an environment where editors will modify an image to fit their version of the story. The risk is that editors will extract from an image another point of view, different from what the photographer intended. In reference to the photographer's subjective eye, Hicks comments,

> One hopes that it may soon be possible to begin to approach photography with a respect for what it is capable of, employing it to parallel words rather than illustrate them, to remain nuanced, saying more than one absolutely needs or even wants to say, to show in great detail what is happening without denying the photographer's point of view, and all the while encouraging people to read images carefully, aware of their propensity to be recontextualized. One wonders if there can be a rethinking of photography, an appreciation of its untapped powers of language once freed from its almost exclusive bondage in the media to "objective" transcription and preconceived illustration.[22]

The subjective story may also include the reader's creation of a story in her mind. Julio Cortázar's short story "Blow-Up" presents the magical situation of a photograph that seems to come alive. Unbeknownst to him at the time, a photographer witnesses a murder. He realizes what happened when examining the photograph he took just before the crime took place. The story is about his obsession with the photograph and the apparent murder story which evolves and comes to life as he examines the photo. After having "seen" the events unfold in the photograph before his eyes, the protagonist remarks,

> I had poked my nose in to upset an established order, interfering innocently in that which had not happened, but which was now going to happen, now was going to be fulfilled... My strength had been a photograph, that, there, where they were taking their revenge on me, demonstrating clearly what was going to happen. The photo had been taken, the time had run out, gone; we were so far from one another, the abusive act had certainly already taken place, the tears

> already shed, and the rest conjecture and sorrow.[12]

The narrative in that photograph which is the subject of "Blow Up" takes place in the viewer's own mind, in this case the mind of the photographer. At first he wonders if the murder actually took place, since he does not record that particular event, only perhaps a few seconds before it was to have taken place. He can only infer that a crime has occurred. He is frustrated at first by his inability to confirm what actually occurred, and finally, by his inability to prevent the supposed crime from being carried out. Sontag's and McCloud's definition of narrative requires a sequence of unique images in order to convey time through a series of discrete events. Cortázar is challenging the precept that the photograph is timeless. Frozen. Stopped forever. He is arguing that the narrative evolves through the ambiguity and anticipation that still photography presents.

Branigan discusses the relationship between the spectator and the "subjective camera" in film narrative. He suggests that there is a different kind of camera for each level of narration.

> The historical, profilmic camera (which usually rests on a tripod) is not quite the same camera which we reconstruct under the pressure of organizing data narratively in our effort, say, to imagine an organ grinder in *Hangover Square* as he discovers a fire in an antiques store, or Phillip Marlowe's ability to have a point of view on a lady.[9]

To define a broader context for a "camera", one would have to consider at least three variables (each of which could take a range of values): person (or nominal agent), time, and place. That is, the "camera" would be defined according to how pictures and sound condense at a given moment into a single hypothesis that stands for an event occurring in a time and place for a given person, such as an author, narrator, character, or spectator.

Branigan also defines the postfilmic camera which I interpret as the editing and post-production aspects of film narrative. Consider applying the principle of a postfilmic camera to still photography through digital techniques.

> A kind of photographic time travel is now possible. The "decisive moment," the popular Henri Cartier-Bresson approach to photography in which a scene is stopped and depicted at a certain point of high visual drama, is now possible to achieve at any time. One's photographs, years later, may be retroactively "rephotographed" by repositioning the photographer or the subject of the photograph, or by adding elements that were never there before but now are made to exist concurrently in a newly elastic scenes of space and time. The "decisive moment" may refer not to when the photographer took the picture, but when the image was modified.[17]

Ritchin may be right that the decisive moment looses some of it's romance and reality in digital photography in general. Through the salient still, the possibility of many decisive moments can be brought together in one image. It is no longer an issue to shoot a burst of 3-10 frames (in one second!) to capture *the* decisive moment. Action passes the camera faster than the photographer can react. To be sure, she still must compose the image and activate the shutter just before the moment. However, her profilmic and postfilmic tools have advanced to the point where the final image is sometimes created in the camera and sometimes afterwards.

1.6 Cronophotography, Futurism, Cinema, Comics, and the Salient Still

The Futurists were compelled by the fast pace of modern industrial society. Their work dealt with relating the visual perception of movement in static artwork. The philosophy of painting in the 18th and 19th century held that the artist was capturing a complete moment on the canvas.

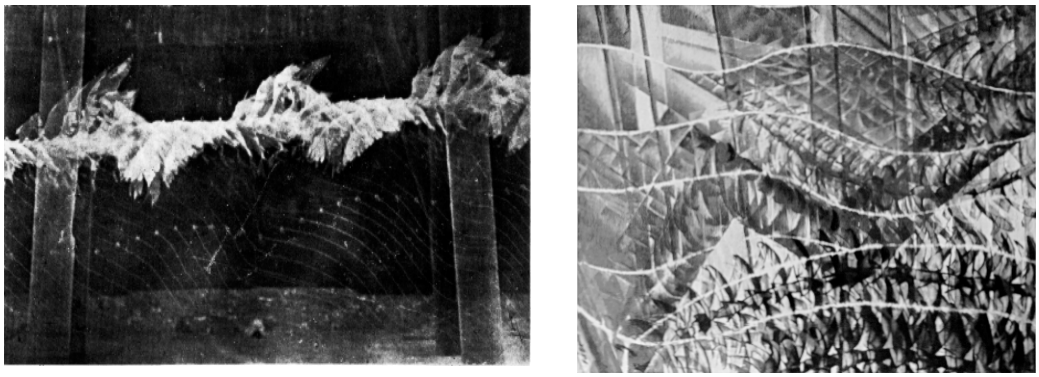

Figure 1.1: Marey's 1886 cronophotograph of a bird in flight[2] (left) with Balla's 1913 "Paths of Movement."[1] Marey's influence on the Futurists has been well documented.

In photography, motion studies by Marey, Doc Edgerton, and Muybridge allowed us to see the evolution of movement. Some of the techniques for illustrating movement from comics mimiced what the photographers' lens captured. The 'language' of comics relied on Iconographic representations for all manners of action and idea. Our understanding of comics as a storytelling medium succeeds by its ability to spatially represent temporal information. All media require a semantic or language to represent experience/perception.

In cinema, comics, classified ads, photography, and the manual arts, meaning is inferred from the visual semantics of the representation. Photography is more literal than drawing, painting, or comics, but still requires an understanding of the context. Cinema is the most literal, but specific techniques are utilized to enhance the plot, emotion, symbolism, etc.

Table 1 contrasts and compares photographic, cinematic and salient stills techniques, or semantics. Physical problems inevitably will arise when translating from one medium to another.

	Photography	*Video*	*Salient Stills*
Temporal	Moment, Mood, Scene Stop action.	Mood, Scene Action.	Mood, Scene Multiple stop action.
Quality	High image quality over entire image. Allows photographer control of focus, depth-of-field, cropping.	Poor image quality. Can use different shots to give detail.	Good image quality, esp. in static scenes or in case of zoom.
Perspective Projection	Plane projection: Up to 3 vanishing points. Spherical (fish-eye) projection Cylindrical (conic) projection: Multiple vanishing points. Take infinitesimal slice of camera image plane perpendicular to center of projection. Volpe projection	Multiple perspectives and points of view. Pan, tilt, or dolly shots.	Piece-wise planar projection. Anamorphic projection: 'Arbitrarily' warped image. Variable shear and/or non-isotropic scaling. Cylindrical (conic) projection: Multiple vanishing points.
Point of View	Fixed center of projection	Multiple c.o.p.'s: Trucks, tracks, dolly shots.	Fixed center of projection.
Special Effects	Motion blur, concurrent with moving foreground, background streaked.	Motion blur while moving.	Arbitrary camera motion: Background registration to reconstruct scene, moving features extracted.
	Motion blur, no camera motion.	Motion blur.	
	Multiple exposure (strobe)		Composites.

Table 1.1: Photography, Video and Salient Stills Compared

1.7 Cinematic Tools for Storytelling

The visual composition of an image sequence may comprise 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 viewers 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.[25] 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 infinite

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.[6]

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.[6,7,10]

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 which 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 a moving image to a still representation is a challenging task. Human visual perception has evolved to be extremely flexible for making sense of 2-dimensional static or moving representations of real or imagined scenes. Minsky notes that a viewer can understand the geography of an entirely foreign geometric space simply by knowing which pairs of points are close to one another.[29] The salient still process can be implemented with sophisticated image understanding techniques, that simulate actual cognitive processes. In lieu of that daunting task, I have used simple image registration algorithms along with a variety of user intervention. Thus, the

salient still is certainly a subjective tool for constructing a synthetic still image from an image sequence. As I will show with some examples, it can be utilized as a specialized creative tool for emphasizing a region of a larger scene or by isolating a particular action. In this respect, you can think of the salient still as similar to a certain cinematograper's trademark lens, or a still photographer's stylistic use of blur for emphasis.

Chapter 2

The Salient Still

2.1 Image Mosaics

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 areas 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[66] 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 multi-resolution 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.

2.2 Previous work

The salient still process utilizes image plane representations that derive from various research interests in imaging science and computer graphics. Abdei-Aziz and Karara[30] introduced a linear algebra approach to perspective modeling to the photogrammetry and image analysis communities.[31-39] Sutherland[40] introduced algebraic methods to the field of computer graphic modeling and rendering. Heckbert[41] applied linear algebra to texture-mapping polygons in perspective. Foley[42] discussed the matrix representation of 2-D transformations using homogeneous coordinates.

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.[45-49] Similar algorithms are used in the field of image enhancement.[52-60] 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,[61] Teodosio and Bender, Currin et al.,[67] Mann and Picard,[70] Hall,[72] Anandan et al.,[73] and Kang.[74] 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 depict the course of a day's events on the battle field 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.[15] 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."[1] Malevitch's works are formed by careful, deliberate compositing of graphic elements that lead to the perception of "rhythmic" movement in a static image.[1]

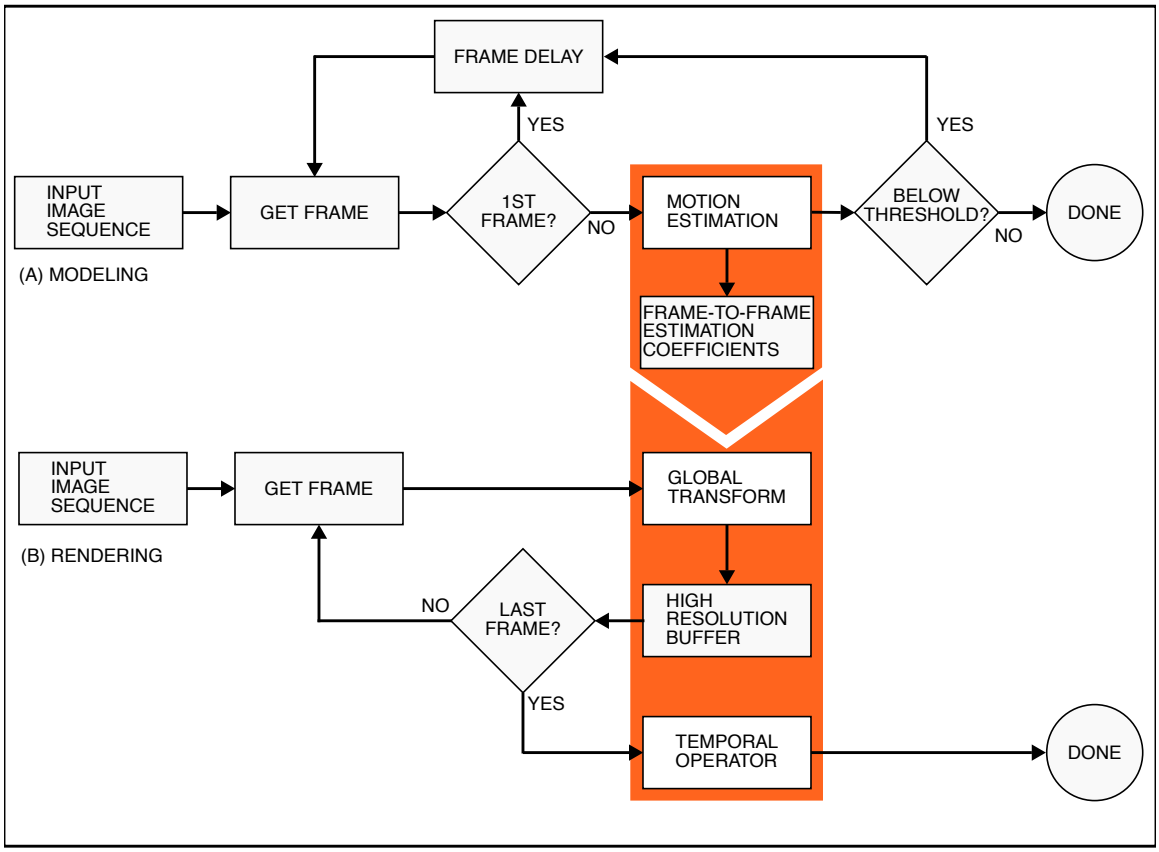

Figure 2.1: Salient still process: modeling and rendering

2.3 The Salient Still

The salient still process involves two stages in processing: modeling and rendering. The modeling stage establishes parameters that estimate correspondence amongst frames in a video sequence (Figure 2.1A). Individual frames are then fit to a global model of the sequence. Still images are rendered from this model (Figure 2.1B). 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 2A. The results of the salient still process is shown in Figure 2B. 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

2.4 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.

.

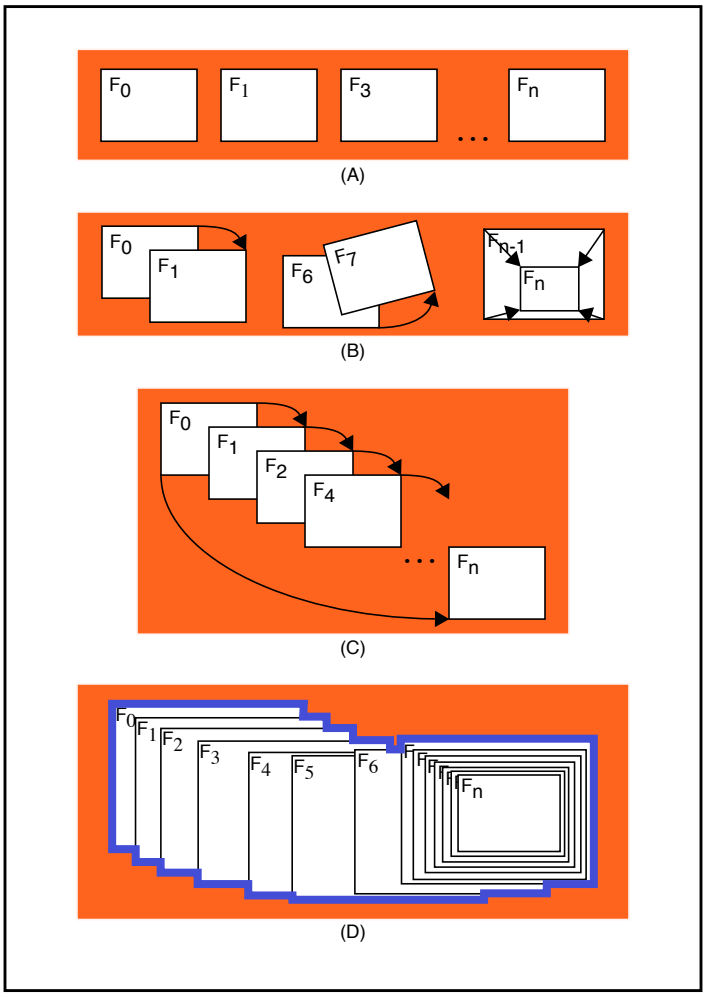

Figure 2.2: Creating a global model: (A) original sequence, (B) frame-to-frame correspondence, (C) cascading, (D) the "video orbit" with its irregular contour

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 non-isotropic 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 which can facilitate estimating the correct projection.

2.5 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 8 parameters that describe the transformation completely. Higher orders of the Taylor series involve N times 6 parameters, where N is the order of the expansion These expansion parameters are linear combinations of the 8 adjustable parameters from the perspective projection model.

If $|\vec{c} \cdot \vec{x}| < 1$ (from Eq. 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 *n*th order in $\vec{x}$. For example, the bi-quadratic 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. Eq. 5 has been utilized in a linear decomposition approach to estimating the perspective projection parameters.[70]

In the affine approximation, $\vec{c} = 0$ and the transformation takes a simpler form with only 6 adjustable parameters:

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

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, which account for shear in the transformed image, and (2) The diagonal elements which 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 the camera motion. Details of the camera model are found in Park et al.,[49] Becker and Bove,[36] Tan et al.,[38] McMillan and Bishop,[43] Aggarwal and Nandhakumar,[46] Melen,[35] and Szeliski.[44]

2.6 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[37] 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 camera's

rectangular field-of-view changing relative position, orientation, and scale as the camera is panned. Figure 2.2D 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 2.2B), 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 2.2C). 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 (Figure 2.3).

.

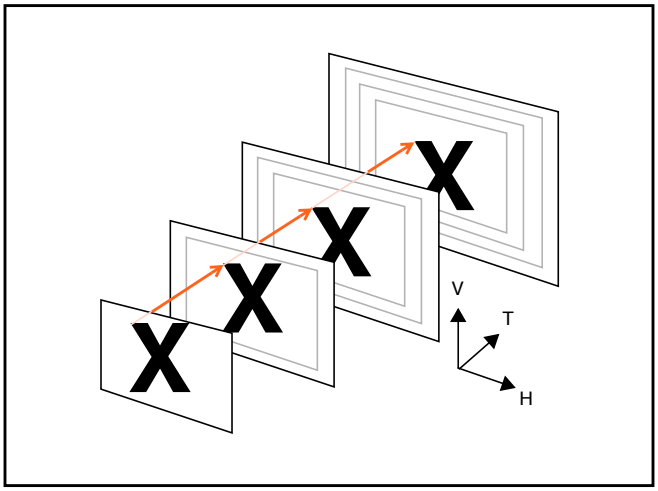

Figure 2.3: Frame-to-scene correspondence.

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 2.3)

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.[45] The optical flow field is modeled by a continuous variation of image intensity as a function of position and time:

$$I(x, y, t) = I(x + dx, y + dy, t + dt) \tag{8}$$

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.

Pyramids

In order to guarantee convergence, optical flow must be restricted to low velocity image sequences. To circumvent this restriction, it is advantageous to subsample 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.[47]

Block-matching

Kang[74] uses the vectors from the block-matching built into the ISO Motion Picture Experts Group (MPEG) digital video coder to generate estimation of frame-to-frame displacements. These estimates are used to generate salient stills, by-passing 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.[39] Motion sensing instruments attached to the camera body allow the camera to record its current position and acceleration. The data is 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 implementa-

tions of the salient still process have employed methods that simplify the computation, perhaps at the cost of generality.

2.7 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, by default the middle frame of the image sequence, determines the absolute coordinate system and consequently the orientation and perspective projection 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 2.4A), 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 2.4B).

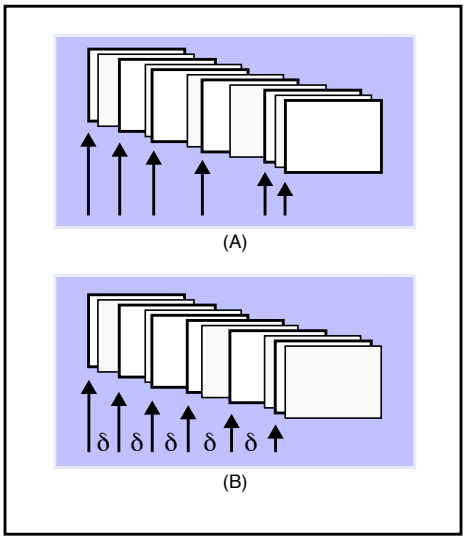

Figure 2.4: Temporal sub-sampling: (A) selecting every nth frame (n=2), and (B) selecting a frame whenever d > threshold

.

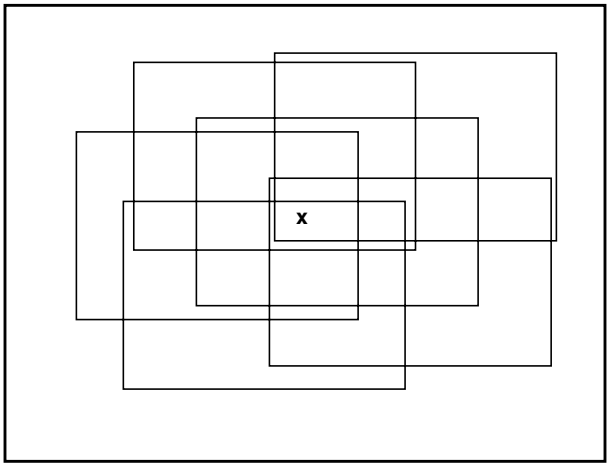

Figure 2.5: Point x is shared amongst multiple overlapping frames

Temporal operators

It is expected that multiple frames will overlap in the global model (Figure 2.5). 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 2.6A and B). These methods are non-causal. 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 2.6C). 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 2.6D). 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.

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 2.6E).

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 2.6F).

Figure 2.6: Temporal operators: (A) first, (B) last, (C) mean, (D) mode, (E) median, and (F) weighted median.

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 param-

eters used in encoding the original image sequence, such as the control parameters used by an MPEG coder.[74] Structured, or object-based coders have the potential of providing high-level information to the salient still rendering process, enabling ready manipulation of "actors."

2.8 Applications

The salient still applications discussed below emphasis 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[13] 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 only lasts for a few hundred milliseconds as 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 pix-

els over the width of the subject's face. For a full-body shot enlarged to 8 x 10 inches, this amounts to approximately 300 dpi effective resolution around the face. The effective resolution falls off rather quickly; there is 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 2.7). 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.

Figure 2.7: Salient still portrait of Marvin Minsky.

Story-boarding

The salient still process has been applied to the generation of a comic book based upon two episodes of the popular TV 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. 140 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.

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. Figure 11 shows a page from the resulting comic book.

Figure 2.8: *NPYD Blue* comics page. The image in the lower right and the four still frames on the left are from the same sequence. Note the perspective distortion and blur effects in the former.

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 was unnatural because the eye didn't perceive a discrete moment in time.[15] 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.

Effects that range from Marey's Chronophotography to Edgerton's high-speed strobe photography 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 appear quite naturally in different locations at different times. In Figure 2.9 a 35mm film sequence was used to capture an olympic diver's trajectory.

Figure 2.9: Salient still composite image of an olympic springboard diver.

2.9 Discussion

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 which 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.

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 than the language of photography). The salient still facilitates the transcoding 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.

There has been much resistance by the news industry to use salient still technology. Ritchin[19] argues for the need to distinguish between images that come directly from the capture device, and those that have been electronically manipulated. Reaves[20] 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, “It is like a reporter using a quote.”[21]

I believe 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. 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.

Chapter 3

Examples

3.1 Photography for salient stills

Previsualization

The key to making salient stills is to previsualize what the result will look like. There are specialized camera techniques that can be utilized to enhance certain features of the salient still. The photographer has to be aware of restrictions on camera movement and framing. In general the frame should contain approximately 60%-80% background features. This depends on the texture and contrast of the specific background. For example, architectural scenes in daylight should be the easiest to estimate, because they exhibit lots of edges and regular patterns. Featureless textures such as blank walls, cloudless sky, or grass may be more difficult to estimate. Thus, the scene should have recognizable features. In the following chapter I will discuss schemes for making the estimator more robust to pathological scenes.

The panning camera should stay fixed at the center of projection, or lens center. This restriction depends on the focal length and distance to the objects photographed. More care must be taken with near objects and wide-angle shots than with distant objects and medium shots. Of course, some very interesting blurring and distortion effects can be achieved by using short pathological sequences which violate the center of projection constraint.

There should be sufficient frame to frame overlap for the optical flow estimator to be linear. In general, the overlap should be greater than about 80%. There is good reason for using video at up to 30 fps. Also, I have experimented some with fast still cameras (~5 fps) with good results. The advantages of film are greater sharpness, dynamic range, and reso-

lution. The disadvantages are higher cost, more time for film development, and more computer processing time.

Editing

Once the video is digitized, it is edited down to just the right length. For example, in the *NYPD Blue* sequences, I found that the actor Jimmy Smits would move his head around much more than the other actors. If the shot consisted of a pan that ended up with a medium or close-up shot of Smits, I would have to carefully trim the sequence to avoid blurring of his head in the salient still. In addition, any scene breaks have to be trimmed since the estimator is not designed to detect scene breaks automatically. Scene boundary detection has been implemented by a number of researchers.[50,51]

It may be necessary to remove frames in which the segmentation is ambiguous. As you are shooting the sequence, someone walking in front of the camera may throw off the estimation. In order to correct for this you should stop moving the camera (pan or zoom) until the background can be seen again. Finally, you can remove the "noisy" frames from the sequence during the editing.

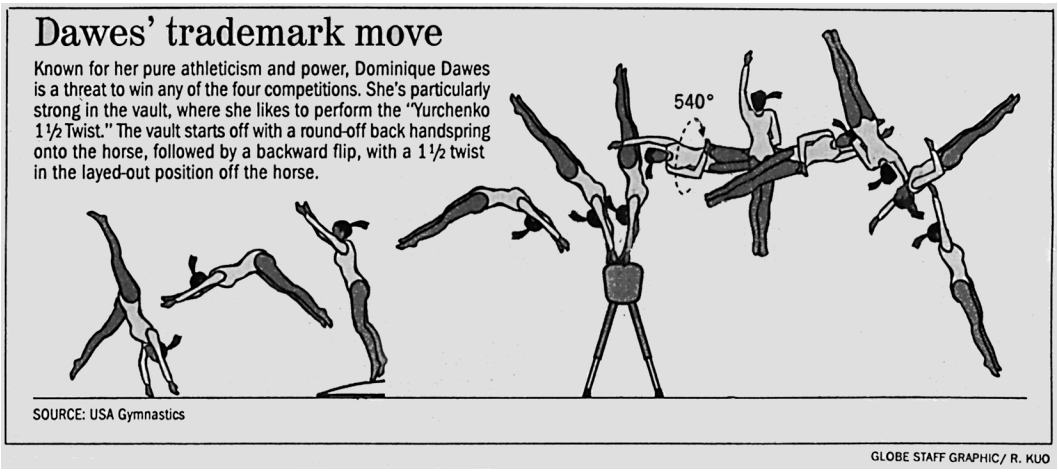

Figure 3.1: Newspaper illustration of a gymnast's trademark vault. [*Boston Globe* graphic, 25 June, 1996].

3.2 Some more salient stills

Sports

In Figure 2.9 a salient still that was created from a short sequence of seven frames from a 35mm camera captures the trajectory of an olympic diver over the course of her dive. Figure 3.1 is an illustration of a gymnast's trajectory during a vault from the *Boston Globe*. Salient stills opens the door for creative illustrations of sports action by compositing and blending several discrete moments of action.

Other dynamic sports action photo-illustrations can be achieved by shooting a fast zoom in or out toward the direction of action.

Figure 3.2: Road cycling salient still from a one-second zoom out sequence.

In Figure 3.2 the same cyclist appears in multiple temporal slices composed in the same salient still. The zooming action creates an active edge that is revealed or obscured as the lens zooms in or out.

Landscape

Narrative has been associated with landscapes because of the relationship between time and space; and location and identity:[16]

> Time and space are the axes of location, which plays a vital role in the formation of identity, and so once again the role of the reader/viewer and his or her own history and situation and how these affect the reader's interpretation of a text come to the fore ... The same can equally be said of both photographs and actual physical landscapes, and so one of the most effective ways in which we can compare written and photographic approaches to landscape is through tackling the question of how these representations function in relation to the reader/viewer.

Figure 3.3 is an example of a salient still panorama. The scene is a shot of Haymarket from the viewpoint of Boston's City Hall plaza. What makes this image compelling is that the activity of Haymarket is sharp and well resolved. The wider scene is Boston's skyline, with City Hall looming on the right. The buildings and street are defocused, so that the viewer is compelled to concentrate her gaze to the Haymarket scene. Thus, the story about Haymarket is emphasized, while the broader context of place is preserved.

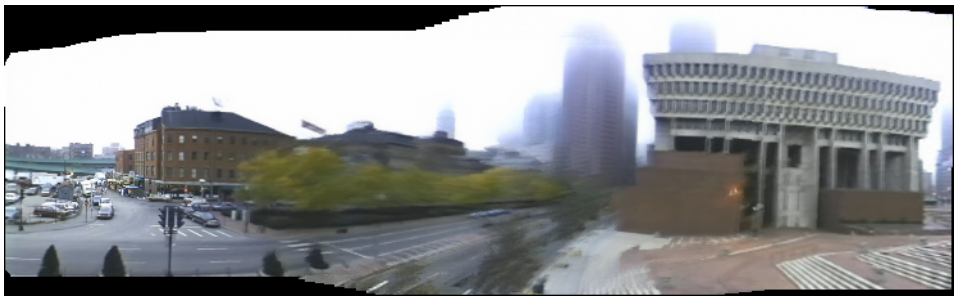

Figure 3.3: Panorama of Haymarket as seen from City Hall Plaza.

Figure 3.4: Detail of Figure 3.3. The full image spans 2900 pixels along its width.

To make the Haymarket salient still, I used an affine model constrained so that the zoom was isotropic and the shear could only follow a pure rotation. In order to do this, the shear components were forced to be equal in magnitude and opposite in sign.

Portraiture

I photographed Mark Berenson when he came to MIT to speak about his daughter's imprisonment in Peru. When I met him, I couldn't help starring into his sad, crystal-blue eyes. I knew my shot would emphasize those eyes. After the press conference, I was invited to photograph him under the banner calling for Lori's release. The shot was a simple pan down the banner followed by a zoom into his face (Figure 3.5). After digitizing and editing the sequence, I made a number of trial stills using various estimation options. Finally I settled on the same zoom/roll model which was used for the Haymarket sequence. The resulting still (Figure 3.5) appears to be in a quasi-cylindrical perspective projection. When the image is rendered with a median filter, the parts of the frame that are correlated over the sequence emerge in the salient still. The zoom-roll model approximates a cylindrical perspective projection for pans or tilts that are parallel or perpendicular to the horizon.

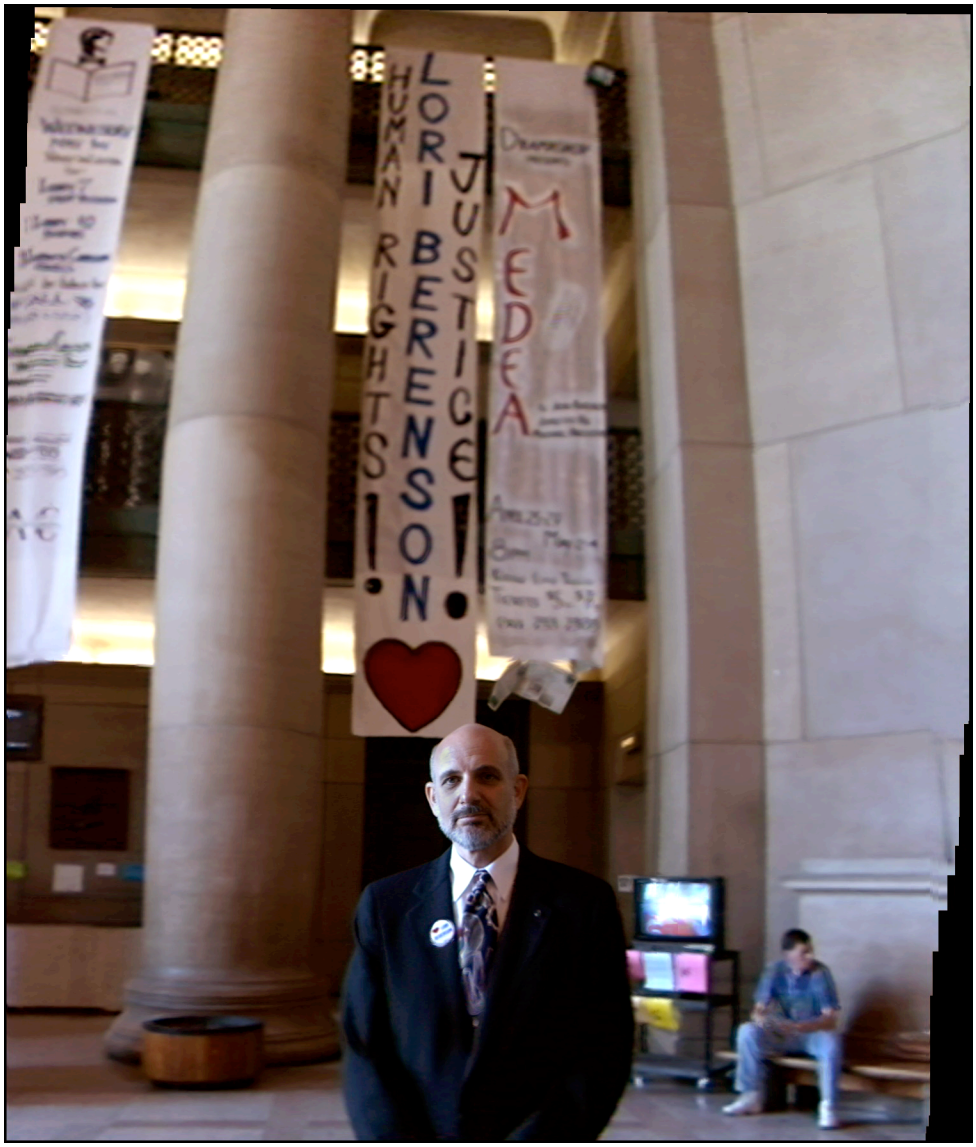

Figure 3.5: Salient still portrait of Mark Berenson.

Chapter 4

Future Work

4.1 Homogeneous Perspective Model

A number of authors have discussed the linear algebraic formalism for the homogeneous transformation that maps points in space to an image plane via a camera function.[35,43,44] Szeliski's[44] approach is arguably the clearest.

A Taylor series expansion of the perspective projection equation was utilized in section 2.5. There, I extended the affine approximation up to the bi-quadratic terms. Conceptually, the affine model does not extend very easily to a homogeneous perspective model with a fixed viewpoint. In the affine model, the coordinate system is taken relative to the image plane. The image plane can be translated, rotated, or scaled. But in a homogeneous perspective model, the operations are taken relative to a fixed (common) viewpoint. Translation *does not* occur in the image plane. The only operations are rotation about the viewpoint, non-isotropic scale, and skew (shear). It is only in the small angle approximation that the affine model is valid. The form of an affine transformation is appealing because it follows a physical model of the scene. That is, the rotation matrix which provides for rotation and scale and translation parameters can be read directly. But the perspective projection model is non-linear. It introduces a term in the denominator which affects the values of the rotation matrix and the translation components.

The correct projective transformation from 3-D objects to 2-D image points is facilitated by constructing a so-called homogeneous coordinate system. A simple point in homogeneous coordinates is composed of the Cartesian coordinates plus an additional component which represents the *projective depth* of the point. Thus, a 3-D object point, **p**, has four components, $(x, y, z, w)^{\mathrm{T}}$, where (x, y, z) are the Cartesian world coordinates, and

w is the normalized distance from the origin (viewpoint) to the image point. Assuming a fixed viewpoint, i.e. that the camera can rotate about the optical center only, the world image can be mapped onto the inside surface of a sphere, or alternatively, onto a cylinder.[43, 27] There exists a 3 by 4 camera matrix, $\boldsymbol{M}_{\text{cam}}$ which maps the world point onto an image plane,[44]

$$\mathbf{u} = \boldsymbol{M}_{\text{cam}}\mathbf{p}\,, \tag{9}$$

where $\mathbf{u} = (u, v, w)$ is the homogeneous coordinate of the image plane. The corresponding Cartesian coordinates are $(u/w, v/w)$.

4.2 Perspective Warps

Estimation could be modified to allow direct estimation of the camera parameters in a cylindrical projection or a piece-wise planar projection.

Directly estimating the camera parameters should be easier than modeling the planar projection because it only depends on five parameters: Pan, tilt, and roll angles, focal length, and zoom factor. Also, cylindrical perspective may be more appealing aesthetically, especially for ultra-wide panoramic images. On the other hand, a cylindrical projection model may not be appropriate for oblique (pan plus tilt) angles since straight lines are maintained only along the cylinder axis.

McMillan[43] does a good job of describing a method for estimating the intrinsic camera parameters, i. e. fixed focal length and roll angle for a 360° panoramic image. By assuming that the first and last frame overlap, the sum of all the individual pan angles must equal 2π radians. Utilizing this additional constraint the estimation is more robust and globally accurate. If the fixed focal length is known, then it is only a matter of estimating the translation in the image plane to determine the pan angle for each frame.

Mapping between two image planes with a common viewpoint can be achieved by a change of coordinates type of transformation of the form:

$$\mathbf{u}' = \boldsymbol{V}'\boldsymbol{R}\boldsymbol{V}^{-1}\mathbf{u}\,, \tag{10}$$

where $\mathbf{u}$ is an image plane point in homogeneous coordinates, $\mathbf{u}'$ is the new image plane point and $\boldsymbol{V}$ and $\boldsymbol{V}'$ are the viewing matrices. The viewing matrix projects 3-D points through the origin onto a 2-D projection plane a distance f along the z-axis:

$$\boldsymbol{V} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1/f & 0 \end{bmatrix}. \tag{11}$$

Note that Equation 10 includes both the reference $\boldsymbol{V}$ and the new $\boldsymbol{V}'$ viewing matrices which differ by the focal length of the frame (f). Equations 10 and 11 are a rather elegant representation of the projective transformation.

Another avenue that should be explored is a piece-wise planar perspective model.[44] Because a planar perspective model is limited to about 120° field of view, it should be reasonable to warp the planar perspective from a few fixed reference frames in the sequence. In that scenario the estimation is made between the 0^{th} frame and all subsequent frames. At some point, the n^{th} frame will not significantly overlap the reference frame. Then, the estimation should start with the n^{th} frame as the reference. The estimation and subsequent warping should be more accurate for long sequences. This method is reminiscent of MPEG predictive coding except that the key frames are separated by a fixed spatial distance instead of a fixed temporal spacing.

4.3 Non-linear Least Squares Estimation

Szeliski[44] was very successful in implementing a Levenberg-Marquart (LM) non-linear least squares estimation for the planar perspective projection model. An LM method is ideal for salient stills because it allows for minimizing a generalized cost function, which could be implemented in a Bayesian *a priori* classification scheme. In such a method, the user could directly influence the estimation to favor certain *a priori* conditions about the sequence. Thus, the same minimizations algorithm could be used with a variety of cost functions for each of the possible category of sequences or camera motion. The estimation would be based on prior knowledge of the state of the system for that sequence, i.e. "pan n degrees," "zoom in," etc. The user selects a particular cost function for each group of frames (video orbit).

4.4 Masking

One of the most difficult problems for the optical flow estimation is image segmentation, that is, to distinguish between a moving actor or object and a moving background. I have proposed that the scene should contain about 70% background features for the most accurate estimation. It should be possible to improve on this figure if something is already known about the segmentation. I'd like to implement a simple method of utilizing an approximate mask to eliminate the moving actor from the estimation. The current version of the salient stills package allows the user to supply a rectangular estimation window. This feature could be modified to allow the user to select an estimation mask, either from a pre-defined set or drawn by hand. The mask is simply a one-bit (black and white) alpha-channel-type image that overlays the frame. Only the masked regions of the image would be used in the estimation. For example, consider a shot panning along with a runner or gymnast. The runner will be in the center of the frame, so you want to use the edges of the frame for estimation. A photographer can shoot with a particular mask in mind or create a

mask for his own shooting style.

4.5 Temporal Filters

Commercially available image editing software allows an artist to create a number of layers for compositing. She can select the opacity of layers and the type of filter that is applied through the layer. However, the filters only act on the layers below. Rendering a salient still involves statistical filtering operations over all the spatially coincident layers. I mentioned the specific filtering methods currently implemented in Section 2.7. New varieties of temporal filters such as “lighten”, “darken”, and “difference” can aid in creative control of the compositing.

A lighten filter could be implemented as follows. Align the frames and create a background image using a mean or median temporal filter. Pick the lighter pixel value for each coincident pixel. The set of lighter images would be saved separately in order to composite the moving actor onto the background.

References

Photography, art and narrative theory.

1. R. Carrieri, **Futurism,** Edizioni Del Milone, Milan, (1966).
2. J. Claire, **E. J. Marey 1830/1904**, Centre Georges Pompidou, Paris, (1977).
3. W. S. Rubin, **Dada and Surrealist Art,** Harry N. Abrams, Inc., New York, (1968).
4. B. Newhall, **The History of Photography: From 1839 to The Present,** Museum of Modern Art, New York, (1982).
5. D. Hockney, **Hockney on Photography: Conversations with Paul Joyce,** J. Cape, London, (1988).
6. W. Eisner, **Comics & Sequential Art**, Poorhouse Press, Tamarac, Florida, (1985).
7. S. Katz, **Film Directing Shot by Shot**, Michael Wiese Productions, Studio City, California, (1991).
8. R. Lowell, **Matters of Light and Depth**, Broad Street Books, Philadelphia, (1992).
9. E. Branigan, **Narrative Comprehension and Film**, Routledge, London, (1992).
10. S. McCloud, **Understanding Comics**, Harper Collins, New York, (1993).
11. F. Masereel, **Passionate Journey**, City Light Books, San Francisco, (1994).
12. J. Cortázar, "Blow-Up", in **End of the Game and other stories**, Harper and Row, New York, pp. 128-130, (1967).
13. R. Solso, **Cognition and the Visual Arts,** MIT Press, Cambridge, (1994).
14. S. Sontag, **On Photography**, Farrar, Straus & Giroux, New York, (1977).
15. A. Scharf, **Art and Photography**, Penguin Books, London, (1968).
16. S. Hill, "Landscape, Writing, and Photography," in *Deep South*, **2**, No. 1, University of Otago, New Zealand, [http://elwing.otago.ac.nz:889/dsouth/vol2no1/sally.html], (1996).
17. F. Ritchin, **In Our Own Image**, Aperture, New York, (1990).
18. F. Ritchin, "The End of Photography as We Have Known It," **PhotoVideo: Photography in the age of the computer**, Paul Wombell, ed., Rivers Oram Press, London, (1991).
19. F. Ritchin, "Electronic Word," *WIRED* **3**, No. 1, (1994).
20. S. Reaves, "The Unintended Effects of New Technology (And Why We Can Expect More)," *News Photographer* **50,** No. 7**,** National Press Photographers Association, (1995).
21. M. Frankel, conversation with author, (1995).
22. W. Hicks, **Words and Pictures**, Arno Press, New York, (1973).

23. Daniel Boorstin, **The Discoverers**, Random House, New York, (1983).

Psychology of visual perception.

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

25. M.A. Hagen, editor, **The Perception of Pictures,** Vol. 1, Academic Press, New York (1980).

26. E.H. Gombrich, **The Image and The Eye,** Cornell University Press, Ithaca (1982).

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

28. A. Valva, "Qualitative Research on Salient Still Technology," unpublished paper, MIT Media Laboratory, (1995).

29. M. Minsky, **The Society of Mind**, Simon and Schuster, New York, (1985).

Image geometry.

30. 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, (1971).

31. S. Ganapathy, "Decomposition of Transformation Matrices for Robot Vision," *Proceedings of the 1st IEEE Conference on Robotics,* Atlanta, (1984).

32. B. K. P. Horn, **Robot Vision,** The MIT Press, Cambridge, (1986).

33. 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).

34. S. F. Ray, **Applied Photographic Optics,** Focal Press, London, (1988).

35. T. Melen, "Extracting Physical Camera Parameters From 3 by 3 Direct Linear Transformation Matrix," *Optical 3-D Measurement Techniques II,* Gruen/Kahmen (Eds.), Wichmann, (1993).

36. Becker and V. M. Bove, "Semi-Automatic 3-D Model Extraction From Uncalibrated 2-D Camera Views," **Proceedings of SPIE Image Synthesis, (**1995).

37. 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, (1995).

38. Y. P. Tan, S. R. Kulkarni, and P. J. Ramadge, "A New Method for Camera Motion Parameter Estimation," *IEEE ICIP*, Washington D.C., (1995).

39. Verplaetse, "Inertial Proprioceptive Devices: Self-Motion-Sensing Toys and Tools,"

to appear in *IBM Systems Journal*, **35**, Nos. 3&4, (1996).

Computer graphics rendering.

40. I. E. Sutherland, "Three-Dimensional Data Input By Tablet," *Proceedings of the IEEE,* (1974).

41. Heckbert, "Survey of Texture-Mapping," **IEEE Computer Graphics and Animation,** (1986).

42. J. D. Foley, A. van Dam, S. K. Feiner, and J. F. Hughes, **Computer Graphics**, second edition, Addison-Wesley Publishing, Reading, MA, (1992).

43. McMillan and G. Bishop, "Plenoptic Modeling: An Image-Based Rendering System," *Proceedings ACM SIGGRAPH 95,* Los Angeles, pp. 39-46, (Aug. 1995).

44. R. Szeliski, "Image Mosaicing for Tele-Reality Applications," CRL Technical Report 94/2, Digital Equipment Corporation, (1994).

Motion estimation and image segmentation.

45. B. K. P. Horn and B. G. Schunk, "Determining Optical Flow," *AI* 17 (1981).

46. K. Aggarwal and N. Nandhakumar, "On the Computation of Motion Sequences of Images - A Review," *Proceedings IEEE* 76/8, (1988).

47. J. Bergen and R. Hingorani, "Hierarchical Motion-Based Frame Rate Conversion," *David Sarnoff Research Center Technical Report, (*April 1990).

48. 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).

49. 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, pp. 288-296, (June 1994).

50. E. Elliot, "Thinking with Motion Images via Streams and Collages," Master's thesis, MIT Program in Media Arts and Sciences, (1992).

51. K. Karahalios, "Salient Movies," Master's thesis, MIT EECS Department, (1995).

Image enhancement.

52. Y. Suenaga, "Super-resolution: Getting a Sharp Image From a Set of Multiple Frames," unpublished paper**,** MIT Media Laboratory, (1982).

53. W. Hannan, "Imaging System With Enlarged Depth of Field," *U.S. Patent 4,404,594,* (1983).

54. P. Burt, et al. "The Laplacian Pyramid as a Compact Image Code," *IEEE Transactions on Communications,* (1984).

55. P. Bennett, and S. Gabriel, "Spatial Transformation System Including Key Signal Generator," *U.S. Patent 4,463,372,* (1984).

56. B. Ferren, "Spatial Imaging System," *U.S. Patent 4,584,704,* (1986).

57. E. Adelson, "Depth-of-focus Image Processing Method," *U.S. Patent 4,661,986,* (1987).

58. W. Glenn, "Television Camera and Recording System for High Definition Television Having Imagers of Different Frame Rate," *U. S. Patent 4,652,909,* (1987).

59. P. Burt, "Pyramid Processor For Building Large-area, High-resolution Image by Parts," *U.S. Patent 4,797,942,* (1989).

60. 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).

61. P. McLean, "Structured Video Coding," MAS M.S. Thesis, MIT, (1991).

62. M. Irani and S. Peleg, "Improving Resolution by Image Registration," *CVGIP: Graphical Models and Image Processing,* 53/3, (1991).

63. A. M. Tekalp, M. K. Ozkan, and M. I. Sezan, "High-Resolution Image Reconstruction from Lower-resolution Image Sequences and Space-varying Image Restoration," *ICASSP,* (1992).

64. R. Ginosar, O Hilsenrath, and Y. Zeevi, "Wide Dynamic Range Camera," *U.S. Patent 5,144,442,* (1992).

65. R. Tinkler, "System and Method for Fusing Video Imagery from Multiple Sources in Real time," *U.S. Patent 5,140,416,* (1992).

66. L. Teodosio and W. Bender, "Salient Video Stills: Content And Context Preserved," *ACM Multimedia,* Anaheim, (1993).

67. 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).

68. P. Migliorati, F. Pedersini, L. Sorcineli, S. Turbaro, "Semantic Segmentation Applied to Image Interpolation in the Case of Camera Panning and Zooming," *ICASSP,* (1993).

69. K. Aizawa, T. Komatsu, T. Saito, and M. Hatori, "Subpixel Registration for a High Resolution Imaging Scheme Using Multiple Imagers," *ICASSP,* (1993).

70. S. Mann and R. Picard, "Virtual Bellows: Constructing High Quality Stills from Video," *IEEE Image Proceedings,* Austin, (1994).

71. H. S. Sawhney, S. Ayer, and M. Gorkani, "Dominant and Multiple Motion Estima-

tion for Video Representation," *IEEE,* (1995).

72. 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).

73. P. Anandan, M. Irani, R. Kumar, J. Bergen, "Video as an Image Data Source: Efficient Representations and Applications," *IEEE,* (1995).

74. J. Kang, "Generating Salient Stills Using Block-based Motion Estimation," EECS M.S. Thesis, MIT, (1995).

Mathematical modeling and parameter estimation.

75. A. N. Tikhonov and V. Y. Arsenin, **Solutions of Ill-Posed Problems**, Winston and Sons, Wash. D.C., (1977).

76. C. W. Therrien, Decision, **Estimation, and Classification,** John Wiley and Sons, New York, (1989).

77. W. H. Press, et al., **Numerical Recipes in C: The Art of Scientific Computing,** Cambridge University Press, New York, 2nd edition, (1992).