Salient Stills

by

Laura A. Teodosio

B. A., Engineering Sciences
Yale University, New Haven, CT
1986

Submitted to the Media Arts and Sciences Section,
School of Architecture and Planning, in Partial
Fulfillment of the Requirements for the degree of

Master of Science

at the Massachusetts Institute of Technology

June 1992

Author _____

Laura A. Teodosio
February 14, 1992

Certified by _____

Walter Bender
Principal Research Scientist, MIT Media Laboratory
Thesis Supervisor

Accepted by _____

Stephen A. Benton
Chairperson
Departmental Committee on Graduate Students

Salient Stills

by

Laura A. Teodosio

Submitted to the Media Arts and Science Section, School of Architecture and Planning, on February 14, 1992 in partial fulfillment of the requirements for the degree of Master of Science at the
Massachusetts Institute of Technology

Abstract

The transformation of a sequence of moving images to one still image presents several problems: a loss of perceived image quality, since the eye is more susceptible to many image artifacts when they are stationary; and a loss of context, since no individual frame can capture everything conveyed over the temporal dimension of a moving image sequence. These problems have been addressed by the creation of a new type of image called salient stills. Though discrete images, they represent not a discrete moment of time, but rather the aggregate changes that occur in a moving image sequence. They retain much of the original content (detail) and context (spatial extent) of the original sequence.

Thesis: Supervisor: Walter Bender
Title: Principal Research Scientist, MIT Media Laboratory

The work reported herein was supported in part by a grant from International Business Machines.

Contents

1 Introduction

Photography is a discrete medium. It suspends the world for one instant, rendering it silently within the two-dimensional boundaries of an image frame. There exists only one visual point of view. Film and video are temporal mediums. Strings of photographs suggest motion, objects move and the viewer's vantage point changes. But what is the relationship between a single frame and a series of frames?

Sometimes it is desirable to transform a sequence of moving images to one still frame - for iconic representation, for data compression or for publication in a print medium. This transformation presents two problems: a loss of perceived quality since the eye is more susceptible to many image artifacts when images are stationary; and a loss of context, since no individual frame can capture everything conveyed over the temporal dimension of a moving image sequence.

These problems are addressed in the creation of a new type of image called *salient stills*. Derived from an intersection of photography and moving image, salient stills attempt to retain both the content and context of the moving visual material. Though discrete images, they represent not a discrete moment of time, but rather the aggregate changes that occur in a moving image sequence. The portrayal of temporal image changes in one still frame requires the use of techniques not normally found in still photography. As a result, the perspective of these images may be warped, aspect ratios altered, resolution varied, multiple vantage points depicted and individual objects in the frame manipulated.

The production of salient stills from moving images requires the ability to move beyond the frame and into its elements - the individual carriers of content. Elements such as background, moving characters and camera movement must be extracted and manipulated. This is accomplished by creating a structured video representation. Objects and movements within frames are modeled rather than whole images. Through the manipulation of these video elements, the salient still emerges.

The determination of what is important in an image is essential. What are the objects that are the carriers of content? What items are needed to provide context? How do these elements interact? Currently, much of this information must be provided externally by editors or other human observers. However, once a paradigm is developed, a computer may be able to execute these tasks unaided. For example, in some contexts such as an interview, faces may be salient and furniture not. A computer can attempt to find an image or part of an image portraying a particular face [Turk90]. In other contexts, like architectural video, furniture may be important but people and faces not important at all.

An obvious place at which to encode saliency into a media representation is at the originating source. Editors and producers know the images or parts of images that are the most evocative, that 'say' something. Admittedly, that saliency of objects will change depending on the context of their usage. The work of this thesis does not address the ramifications of the use of video objects for wholly new appropriations. Nor does the computer attempt to discover what objects are salient; rather it manipulates the salient (or non-salient) objects to simulate a new image.

There are many ways to build a structured representation of video signal: extract any information that is embedded in the signal (i.e. closed captioning of television broadcasts); retrieve it directly from the signal; laboriously log the content by watching the signal, or have a user assist the computer in identification. None of these methods is ideal or adequate by itself, so a combination is explored.

Searching for salient information in an image sequence is only the first half of its

reappropriation, transformation and republication. The eventual manifestation of the information into a perceptual signal must be considered in any representation.

This thesis explores methods to extract video structure and tools to perform the composition of salient stills.

Chapter 2 provides background material. The pertinent information processing and image processing precedents are discussed. Techniques used to show changing information in stationary images are reviewed.

Chapter 3 explains the implementation of the tools for creating salient stills including a new method to create high resolution images from variable focal length frames.

Chapter 4 discusses some of the potential applications of salient stills and their intermediate technology.

Chapter 5 discusses some specific images created.

2 Background

The foundations and impetus for this work come from two worlds: information processing, particularly image processing; and the history of images - painting and photography. Information processing offers the ability to decouple a signal from its packaging and manipulate it in a structured manner. Various forms of painting and photography offer technologies to transform temporal data into a discrete two-dimensional form. Salient stills lay at the intersection of these two domains.

2.1 Information Processing

Digital technology offers the potential to radically transform how media information is received and reused. Once distinct forms of distribution, (i.e. radio waves, television, print) can be reduced to one common channel - a bitstream. This bitstream representation allows easy searching, parsing, augmentation, compositing, and combining of different forms of media information. Text articles can be directly embedded into images; radio can become narration track to video; segments of images can be extracted and synthesized into a new scene. These manipulations can be automated and personalized for the human receiver at the point of consumption.

Beyond new arrangements and combinations of signals, the bitstream facilitates signal transformation. Media information can be encoded such that its presentation mode is flexible. Digital data is no longer bound by its original packaging, the audio of a broadcast

can become text; moving images can become information rich stills. This type of transformation between presentation modes is called *cross-media transcoding.*

'Transcoding' is a term commonly used to describe a transformation of the representation of an image. It is becoming used to describe other classes of transformations, as well. An example of transcoding is electronic broadcast signal conversion - NTSC to PAL (a 60 Hz to 50 Hz conversion) or RBG to YUV transformations. This transcoding processes may be lossy in that some degradation of the signal may occur. An occasional frame of a video sequence or some frequency range of an image may be missing. Sometimes context is lost as happens in a 5:3 film to 4:3 video aspect ratio conversion. Ideally, this lossy transformation does not alter the message significantly so that human perception of content of the transcoded signal remains intact.

The goal of creating lossless transformations poses particular problems when the signal's analog manifestation is radically different from its initial encoding. For example, audio delivers certain non-lexical cues which enhance our understanding of the message - intonation, pitch, cadence etc. Video provides movement, a temporal dimension. In the transcoding of this data, every effort must be made to retain this characteristic information whenever it is prominent.

2.1.1 Signal Sense

The simple transformation of electronic signals into their digital counterparts does not guarantee that these media manipulations and transformations are realizable. Computers must be able to sufficiently "know" the signal and "know how" to dismantle it.

While, for example, a computer can search through the bits of words comprising text, and retrieve a particular segment, it has no understanding of the whole document. Bits of the image data can be processed as well, but when, for example, do clumps of white, brown, blue and red pixels become an image of another riot on the streets of NYC? The

arts of machine vision and story understanding are in their infancy, and are not yet up to the task of decoding a binary bitstream as efficiently as humans find meaning in the electromagnetic waves impinging on their sense receivers. Therefore, a signal must be represented in a way that reveals its content to the computer system decoding the information. It must have "a sense of itself" [Lipp91] - some sort of accessible description that transforms these multitudes of bits into a coherent structure.

The 'sense' annotation should provide content information in a manner which allows the selective dismantling of the signal. For example, whole sections of images such as characters should be easily identifiable and retrievable. Additionally, the 'sense' annotation must be easy to update. In this way, any discoveries made by analyzing the signal (on its own or in conjunction with the annotation) can be put in the annotation and the signal rebuilt.

The annotation can be exterior to the signal, in a remote text database, for example, or it can be embedded, as in the closed captioning of television broadcasts. Closed captioning is a complete (or nearly so) audio transcript of a television program that is carried in the vertical blanking interval of the video signal. For those with the special purpose hardware, it is decoded in the home and displayed as scrolling text on the television screen. Though the closed captioning was created solely to provide subtitles for the hearing impaired, it inadvertently provides a structure from which manipulation of the television broadcast can begin. Not only is a full transcript provided, synchronized with the video, but speaker changes are clearly indicated. For some shows, like news broadcasts, story boundaries and changes of venue are clearly marked. Commercials are easily detected. There is also some indication as to where a speaking character is located in the 2-dimensional image. Voice boxes, analogous to speaking bubbles of cartoons, appear near the heads of on-screen speaking characters.

2.1.2 Network Plus

Network Plus [Bend89] was an early example of the use of closed captioning to aid in

decoupling and transcoding a pre-packaged media signal. The system created a personalized augmented summary of the day's news. Although the display environment possessed little video and audio capability, every attempt was made to display features about the data which were considered salient. In an attempt to preserve non-lexical audio cues, the audio track was analyzed for intonation and cadence [Hu88]. This information was used to highlight in the display of the transcript the words that received special acoustic emphasis. Since there was no detailed log of video information, other means were devised to select still images. The results of the audio analysis were used to select images of the newscaster at an emotional moment. Additionally, the video image itself was decoupled from its packaging - the frame. Segments of the image such as the box over a newscaster's shoulder, were extracted from the frame and became images unto themselves. In an attempt to retain contextual coherency and prevent redundancy, the box was only extracted when its contents changed.

These selected images initiated the concept of salient stills. Though they did not attempt to convey temporal information, their creation exemplified the power of a computational agent working on a signal laden with structured content information. In the future, some of the information inferred by the Network Plus system should be highlighted by editors and enclosed with the signal.

2.1.3 Transcoding of the moving image to still

This thesis explores one type of cross-media transcoding - that of the moving image to still form. This transcoding presents several problems. There is a loss of perceived resolution particularly for video footage. The rapid scanning of a television raster and the movement within a displayed image masks the fact that a video image is composed of 525 individual lines, each captured at a different moment of time. Freezing one frame of this data often reveals scanning noise and other transient artifacts such as the checker board pattern of the color subcarrier of an NTSC signal. These artifacts grow more noticeable as the image increases in size. Thus, building a very high resolution still on

the order of 4000 x 3000 pixels becomes a daunting task.

Additionally, film and video are temporal mediums. Characters move in and out from the field of view. The camera moves in space and time. Focus pulls forcefully demand our attention. There is not necessarily a single frame of an entire sequence of images that can capture its intended expression. So, creating a single still becomes a problem of preserving both content and context while condensing data.

2.2 Image Processing Precedents

2.2.1 Model Coding

A system that constructs salient stills must know the camera movement in an image sequence, know the content of the frames, and must have the power to manipulate them. This representation of a visual scene as a series of manipulable objects is the basic model of synthetic images in computer graphics. It has only begun to be explored in the world of 'real' images.

Traditional digital compression of moving images has been accomplished by encoding the entire frame as a complete unit using various encoding schemes. Hybrid video coders [MPEG], subband and pyramid decomposition coders [Burt84], and region coders [Keit88] are all examples. Conversely, model-based coding allows representation and manipulation of separate objects in the scene. The transmitter and decoder agree on the basic model of an image and the transmitter sends parameters to manipulate this model. This can be considered a further extension of the concept of embedded annotation discussed in the previous section. The objects themselves serve as their own descriptions.

2-Dimensional

Two-dimensional versions of model based image coders have been created with the intention of compositing photo-realistic images at the receiving end.

To simulate the activity on a computer network, video objects were composited on a representative background space [Watl89]. As the activities of users on the network changed so did their corresponding video simulations. If they began reading news, they were shown reading a newspaper, if they were idle for a long time, they were shown taking a nap. The images displayed for each individual user were from a stored set of 5 or 6 images taken of them to represent particular actions. No attempt was made to capture real-time images of the individual users. The stored set of images, though only suggesting the activities of the user, were quite effective in communicating information.

A model-based coder was built to reduce the bandwidth needed to transmit a video scene [McLe91]. Rather than transmitting redundant information in every frame, a 2-dimensional background set was extracted from video data and transmitted once. Then only the characters and associated placement and movement instructions were sent to the receiver and the scene animated.

3-Dimensional

Three dimensional model-based coders have been created to alter video footage. A depth-from-focus camera was used to record 2 1/2D images of a scene. From this data, a 3-D database model of the area photographed is created [Bove89]. With this model, he is able to alter the footage as if it had been shot at a different angle or under different lighting conditions or with a different focus setting.

Another application successfully created a 3-D model of a room from its video recording [Holt91]. Using a-priori knowledge about the world, a model is created by hand. Then the actual footage is used to refine the model. Once created, he was able to colorize the footage and create new sequences of the scene shot from different angles and with novel camera traversals.

2.2.2 High Resolution

There are various methods for creating higher resolution images from video. These

include: deinterlacing methods [Doly87], edge enhancement [Xue90],pyramid coding and pattern matching [Holt90], spatio-temporal encoders [Cla88] and fourier sampling [Tsai84]. While all these methods are successful to varying degrees, none has the ability to combine images of varying focal length or images captured with camera motion more complex than pure translation. Also, none addresses the problem of creating larger aspect ratio images while retaining high resolution.

This thesis explores the use of images of variable focal length to create high resolution stills. The images are warped into a common space/time volume. The redundancy in pixels imaged over time and at different resolutions is exploited to create a higher resolution image. Patches of salient stills may have increased resolution due to a camera zooming-in on the corresponding object in the world. This variable resolution, rather than being disconcerting, is a narrative tool, directing the viewer's attention to the parts of the scene that commanded the attention of the camera operator or parts of the scene deemed important in the creation of the salient still.

2.2.3 Computer Manipulated Images

Some other techniques needed to direct a user's attention have already been developed for digital images.

Borrowing from cinematography's use of the rapid changing of a camera's focus point to draw attention to objects in a scene, a technique was developed to simulate focus pull for computer graphics work [Stur89]. Though this technique has been applied to simulated moving images, changes of focus can be used in salient stills to direct a user's attention.

A second approach used focal, transparency and color value gradients in a multi-layered changing image to capture a user's attention [Scho91] [Colb91]. Though the images were transitional, this work exemplified how the above properties of a digital image can be manipulated to direct attention.

2.3 Traditional Images

We turn to inherently still and two-dimensional forms of expression - painting, drawing and photography for a methodology for representing a temporal dimension in a discrete fashion. A detailed historical account of the varieties of methods employed to represent motion, saliency and narrative is beyond the scope of this paper. However, some of the key ideas which have influenced this work are outlined below.

Representing Motion

Four broad categories of two-dimensional images where motion is depicted has been discovered [Frie86]. These categories include single viewpoint - single moment, multiple viewpoints, metaphor and abstract representations.

The single viewpoint - single moment picture shows a momentary frozen environment from one point of view. The movement of the objects in the image is suggested by a particular view of the moving object. Usually the object is drawn in a postural position different from a resting position. For example, walking characters would be shown in a sideways position, one foot up; curtains will be shown extended out from a window pane. Motion paths such as footprints in sand and the wake behind a swimmer are common in this class of images.

The multiple viewpoint image shows objects or parts of objects at successive moments. The change is recorded as multiple images. For example an image of a spinning dancer will show a frontal image superimposed on a sideways and back images. Multiple copies of legs and arms suggest movement. Multiple viewpoints was common in Chinese and Japanese narrative art though it was also used within the confines of one-point linear perspective paintings of the Renaissance. Gozzoli's 'Dance of Salome' while perspectively correct, represents the main character multiple times to show her temporal movement.

Movement can also be suggested through pictorial metaphor. Blurring a character or drawing lines after it will accentuate the sense of movement. It may seem that these

devices have been purloined from the realm of photography which renders moving objects in such a manner. However it has been found that such techniques were used as early as the eleventh century.

Movement can also be represented symbolically without the depiction of any moving object, for example, arrows drawn on a map to indicate troop movement.

Representation of Saliency

To increase or decrease the prominence of an object in an image, Renaissance painters employed classic color perspective and aerial perspective. In their ever-present quest for realism, they developed a color scheme to mimic the perceived changes in color as objects recede into the horizon in an outdoor scene. Distant receding objects were painted in cool colors or grayed-down shades and tints while salient objects were painted with warm, intense or bright colors [Dunn91].

Representation of Multiple Vantage Points

Modern viewers are steeped in a world of one-view point perspective images as delivered by photography and video. Most modern forms of commercial and popular art still adhere to some form of linear perspective as well. These are images where the viewer is grounded in one spot with the eye rigidly focused. There is no room in this type of image for the representation of multiple events and multiple vantage points, something which would be needed in creating an image which has a temporal dimension. But various forms of narrative art successfully have employed non-perspective methods in order to accomplish representing a temporal dimension in a 2-D image. To find some examples of this type of image, we turn to indigenous art, cubism, and eastern art.

Northwest Coast American Indian Art exemplifies image making in which multiple vantage points enable more than one rendition of an object to be seen [Hage86]. Such renderings are often called split-style because the objects being depicted appear as if they have been split down the back and then spread out in the available space. Sometimes the

objects are split apart into sections, which are then distorted and distributed throughout the available space. The depicted object, however, is still identifiable.

Cubism was a short-lived art form which was a reaction to the technological advances of the early 20th century. Photography had rendered futile any attempts to capture realistic images in painting. The concept of relativity was changing perceptions about time, space, and distance. Cubist images fused multiple viewpoints into single forms and multiple images into a single painting creating not a moment but an event. The eye was forced to scan the image for all the pieces and to build the objects through memory, much like what happens as we scan an object over time in the real world [Fry66].

David Hockney was not the first but is the most well known of photographers experimenting with a cubist and narrative form. He created images which he called 'joiners' composited from many still frames captured over a period of time and from different viewpoints. He did not want to capture images of the world as interpreted by the lens on his camera, but rather as he saw it: moving and changing over time. These are images are "at war with established notions of photographic 'realism'" [Joyc88]. Other photographers who experimented with rendering the changes of time in one still image include Jules-Etienne Marley [Lawd75] and Doc Edgerton[Edge87]. However, their multiply exposed images, unlike Hockney's, do not attempt to break the one-point perspective of the photographic lens.

Two of the most successful forms of narrative art are the Japanese yamato-e [Hage86] and Chinese art previous to the 20th century [Silb82] [Cahi82]. Typically, the images are constructed with a vantage point at infinity. Linear perspective is ignored.

2.4 Tools

The success of the above art forms provides convincing evidence that manipulating the space/time features of digital images can accomplish the goals of delivering temporal information in a way that is understandable to human viewers. Coupled with the tenets

of image processing, a series of tools was created to build salient stills including images with enhanced resolution.

For continuous camera movement scenes such as pans and zooms, a 3-dimensional space/time volume of the image sequence is constructed. From this intermediate representation, pan and tilt shots can be unwrapped, zoom shots expanded out, and truck shots peeled out from the center. Additionally, the redundancy in the pixels over time can be exploited to create higher resolution images or high resolution patches in images.

A single background can be abstracted and used to identify moving objects in a scene and their spatial location at discrete times determined. The compositing methodology uses color, blur, size, shear, placement, blending, resolution and repetition of objects to describe some temporal phenomenon.

3 Implementation

A camera visually samples the world in front of it. This sampling provides redundancy in the information captured. That is, even if the camera is moving, spatial locations in the world will be imaged a multitude of times over a given temporal duration. This repeated sampling is exploited to create extended high resolution scenes and locate moving objects. Additionally, if the scene is rendered with varying focal length shots, such as occurs in a zoom sequence, then each spatial location in the world will be rendered with a different resolution in the successive frames. These images can be combined to create a high resolution final image which has both the wide field of view captured by the short focal length frames and the detail captured by the long focal length frames.

The series of imaged frames is transformed into a 3-dimensional space/time volume. Spatial location in the world is on the X and Y axes and time is on the Z axis. Therefore, a vector passing through the volume, perpendicular to the first image plane will pierce the same spatial location in the world in each of the images. (See fig. 3.1). Additionally an image sequence captured with variable focal lengths are mapped to a common, fixed focal length.

In order to create the space/time volume, camera movement must be recovered. This is done by an optical flow analysis with movement modeled as an affine function. Once the space/time volume is created, various forms of statistical operators are performed along the temporal dimension to create the high resolution extended panoramic scenes. A comparison of the actual footage with this extended scene locates objects moving

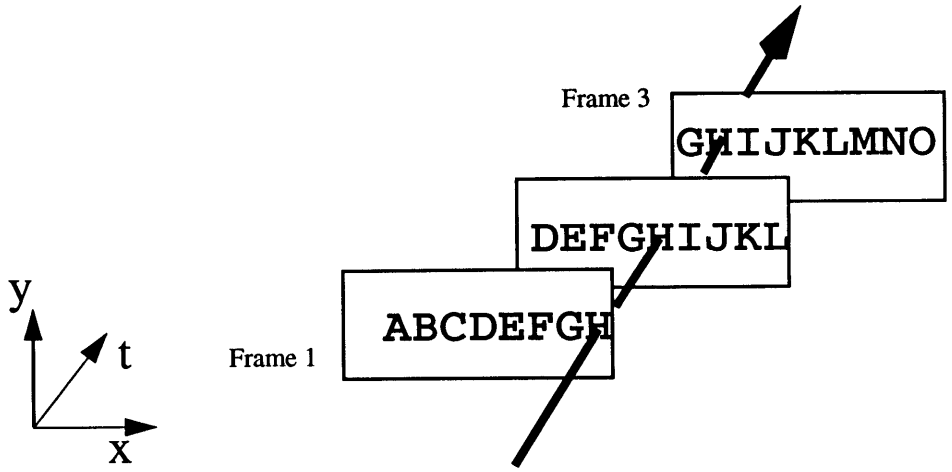


Figure 3.1 Space-time volume for 3 frames from pan shot. A vector passing through the volume lands on the pixels representing the same coordinate in world space.

relative to the camera. These objects then are manipulated and composited back into the extended scene.

A summary of the operations performed on moving image data that results in an extended scene is shown in figure 3.2.

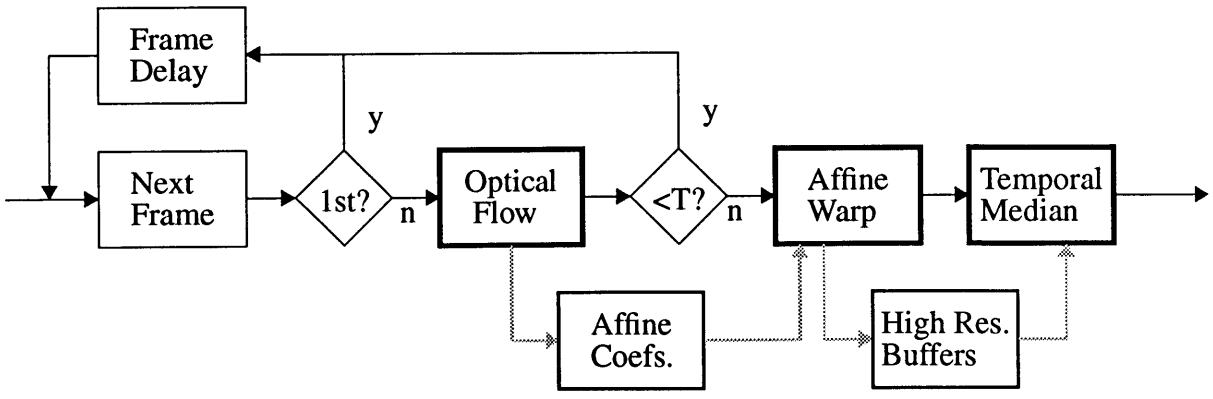


Figure 3.2 System block diagram. There are three steps to the process: 1. The optical flow between successive frame pairs is calculated. 2. Successive affine transformations, calculated from the flow, are applied. These translate, scale and warp each frame into a single high resolution raster. 3. A weighted temporal median filter is applied to the high resolution image data, resulting in the final image.

3.1 Optical Flow

Optical flow is principally used to determine global camera motion.

The apparent motion in brightness fields of an image sequence is called the *optical flow* or *instantaneous velocity field.* It is a two-dimensional projection of three-dimensional motion onto a plane. The result is a series of two-dimensional vectors which describe the translational movement of points in the image as projected onto this plane.

The optical flow, except in special situations, is not too different from the actual motion field of the image [Horn86]. Therefore, motion can be modeled by a continuous variation of image intensity (I) as a function of position and time:

$$I(x, y, t) = I(x + dx, y + dy, t + dt) \qquad \text{(Eq. 1)}$$

Solving the above equation can be done by several techniques including the use of: spatio-temporal gradients [Limb75] [Horn81], energy based models [Ade85] [Heeg87], token matching [Marr79] [Ulma79], block matching [Jain81] [Netr88] or frequency domain correspondence [Giro89].

Two methods were used in this work: block matching and spatio-temporal gradients. The spatio-temporal gradient technique proved more effective in retrieving complex camera motion.

Block Matching

Block matching is a simple way to estimate camera motion. A selected block of an image is compared with blocks in the previous image to find the closest match. It is assumed that the motion is parallel to the plane of the camera, lighting is spatially and temporally uniform and that no areas are revealed or occluded. These assumptions hold well enough over frame-by-frame changes to yield accurate results.

The method employed uses a N-step search with N=4 and block size of 8 [Netr88].

Initially, 8 pixels are searched around the center of the search area in a square of 2N around the center. Once a minimum is found, the next search gets centered on this square. Eight pixels are again searched but with a narrower search space. The process continues until the search width equals zero. The vector passing from the initial center of the block and the final center is chosen as the motion vector.

A histogram of the local block movements is created for the whole image. The distinct peak in the histogram is chosen to be global camera motion. If there is much movement in the image, there will many peaks and perhaps not even a distinct one.

Block matching breaks down when the movement of the blocks is larger than the search space and if two differing motions are in one block. Since all motion is assumed to be parallel to the camera plane, the effects of camera zoom and object rotation are not captured by this method.

3.1.1 Spatio-Temporal Gradient Techniques

Spatio-temporal gradient techniques attempt to measure the local spatial and temporal image intensity gradients and apply these to (Eq. 1). Assuming that the motion field is continuous everywhere, we can expand the right hand side of the equation by using a Taylor series. Dropping the higher order terms gives the standard optical flow equation.

$$\frac{\partial I}{\partial x}\frac{dx}{dt} + \frac{\partial I}{\partial xy}\frac{dy}{dt} = -\frac{\partial I}{\partial t} \qquad \text{(Eq. 2)}$$

This equation is usually solved by modeling image motion as translational in the x and y directions. This works well if the movement of the camera is a pan or tilt but it does not model zoom. If motion is modeled as an affine transformation, however, zooms in combination with other camera movement can be recovered. The specific algorithm and implementation comes from Bergen [Berg90]. The paper discusses a method for accurately deriving optical flow for two differently moving objects in a moving image

sequence. The implementation uncovers both single gross motion and two differing motions in a scene.

The truncated Taylor series is an accurate estimate, as long as the higher order terms would not have been significant. Therefore, the displacement given by (Eq. 2) is accurate only when the frame to frame displacements are small, ideally subpixel. By employing tracking within a multiresolution structure, Bergen relaxes this requirement and expands the estimation procedure to cases of larger displacements. The multi-resolution structure used in this implementation is a gaussian pyramid.

3.1.2 Motion Estimator

Following the notations given in Bergen et al., if a pattern is moving with a velocity *p(x,y)* then modeling only translational changes:

$$p(x, y) = p_x(x, y), p_y(x, y) \qquad \text{(Eq. 3)}$$

Replacing the above motion model into (Eq. 1) gives:

$$Err = \sum_{x, y \in R} (I(x, y, t) - I(x - p_x, y - p_y, t - 1) \qquad \text{(Eq. 4)}$$

Expanding using a Taylor series and dropping the higher order terms gives:

$$(x - p_x, y - p_y, t - 1) \approx I(x, y, t) - p_x I_x(x, y, t) - p_y I_y(x, y, t) - I_t(x, y, t) \qquad \text{(Eq. 5)}$$

where

$$I_x = \frac{\partial I(x, y, t)}{\partial x}, I_y = \frac{\partial (x, y, t)}{\partial y}, I_t = \frac{\partial (x, y, t)}{\partial t} \qquad \text{(Eq. 6)}$$

Consequently

$$Err = \sum_{x,y \in R} (I_t + p_x I_x + p_y I_y)^2 \qquad \text{(Eq. 7)}$$

An equation for image motion is obtained by setting the derivatives of (Eq. 7) with respect to each of the parameters of the velocity components to zero and solving the resulting system of equations. If the motion is modeled as simple translation: $p = (a_x, a_y)$ where a_x and a_y are constants, in units of pixels, then the optical flow equations are as follows:

$$\left[\sum I_x^2\right] a_x + \left[\sum I_x I_y\right] a_y = -\sum I_x I_y \qquad \text{(Eq. 8)}$$

$$\left[\sum I_x I_y\right] a_x + \left[\sum I_y^2\right] a_y = -\sum I_y I_t \qquad \text{(Eq. 9)}$$

Affine Modeling

The above equation works well with certain types of rigid body and camera motion. However, this modeling does not facilitate recovery or simulation of movement in the z plane, typically the kind of motion as occurs in a zoom camera operation (change in focal length of the capturing lens). This motion is more accurately modeled instead as an affine transformation, i.e. a transformation that transforms straight lines into straight lines, parallel lines into parallel lines, but may alter the distance between the points and angles of lines. The method outlined in Bergen is perfectly suited to both uncovering typical video camera motion, and describing that motion succinctly for later warping.

So, if velocity p is described by six parameters, $a_x, b_x, c_x, a_y, b_y, c_y$, the velocity equations become

$$p_x(x,y) = a_x + b_x x + c_x y \qquad \text{(Eq. 10)}$$

$$p_y(x,y) = a_y + b_y x + c_y y \qquad \text{(Eq. 11)}$$

where a_x and a_y are pure translation terms in the x and y directions in units of pixels, b_x is a percentage scaling factor for x in the x direction, c_x is a percentage rotation factor for x in the y direction, b_y is a percentage rotation factor of y in the x direction, c_y is a scaling factor for y in the y direction. In the above equations, zooming of a camera lens is described by constantly changing b_x and $c_y$terms.

Taking the above equations and placing them in (Eq. 7) and differentiating the error with respect to each of these six parameters results in a linear system of thirty-six equations (Eq. 12).

$$\begin{bmatrix} \sum I_x^2 & \sum xI_x^2 & \sum yI_x^2 & \sum I_xI_y & \sum xI_xI_y & \sum yI_xI_y \\ \sum xI_x^2 & \sum x^2I_x^2 & \sum xyI_x^2 & \sum xI_xI_y & \sum x^2I_xI_y & \sum xyI_xI_y \\ \sum yI_y^2 & \sum xyI_x^2 & \sum y^2I_x^2 & \sum yI_xI_y & \sum xyI_xI_y & \sum y^2I_xI_y \\ \sum I_xI_y & \sum xI_xI_y & \sum yI_xI_y & \sum I_y^2 & \sum xI_y^2 & \sum yI_y^2 \\ \sum xI_xI_y & \sum x^2I_xI_y & \sum xyI_xI_y & \sum xI_y^2 & \sum x^2I_y^2 & \sum xyI_y^2 \\ \sum yI_xI_y & \sum xyI_xI_y & \sum y^2I_xI_y & \sum yI_y^2 & \sum xyI_y^2 & \sum y^2I_y^2 \end{bmatrix} \begin{bmatrix} a_x \\ b_x \\ c_x \\ a_y \\ b_y \\ c_y \end{bmatrix} = -\begin{bmatrix} \sum I_xI_t \\ \sum xI_xI_t \\ \sum yI_xI_t \\ \sum I_yI_t \\ \sum xI_yI_t \\ \sum yI_yI_t \end{bmatrix} \qquad \text{(Eq. 12)}$$

This system of equations is solvable in that *Ix, Iy, It* can be determined from analyzing the intensity values of any two successive images. The values for x and y are the physical locations of the pixel from the image center.

Local spatial and temporal brightness gradients are computed for the image. $\frac{\partial I}{\partial t}$ can be

approximated from taking the difference of the signal between frames and convolving with a summation filter. $\frac{\partial I}{\partial x}$ and $\frac{\partial I}{\partial y}$ can be computed by summing two consecutive images then convolving the result with a separable first derivative filter, once in the x direction and once in the y direction. This filter gives the changes in intensity of the signal.

To determine the affine parameters, first a Gaussian pyramid is constructed for each frame pair of images (t and t+1). Computation begins at the lowest level and is refined at each successive level by computing the residual motion (See Fig. 3.3). The number of pyramid levels can be determined at run time but on average a 3 or 4 level pyramid worked well for a 525 line original image.

Starting at the bottom of the pyramid for each pair of frames, affine parameters are computed between the lowest levels. Sometimes, a seed is chosen to approximate known aspects of image movement. If there is confidence of steady camera movement, either by monitoring of previous estimates or some other a priori knowledge, the seed parameters can be the estimates from the previous pair of frames.

These parameters are then used to warp one of the images of the pair into the same space as the other image. This process attempts to undo the movement between frames. Since the affine estimate is not exact, the images will still be slightly different. Residual motion between the two images must be computed.

This residual estimate is then linearly added to the previous estimate. This refined motion estimate is used to warp the first image of the pyramid level again and the process of refinement continues on this level for a specified number of times. After a few iterations at one pyramid level, the final estimate is passed up to the next pyramid level and the process continues.

The above method was calculated for the whole image with the intention of capturing

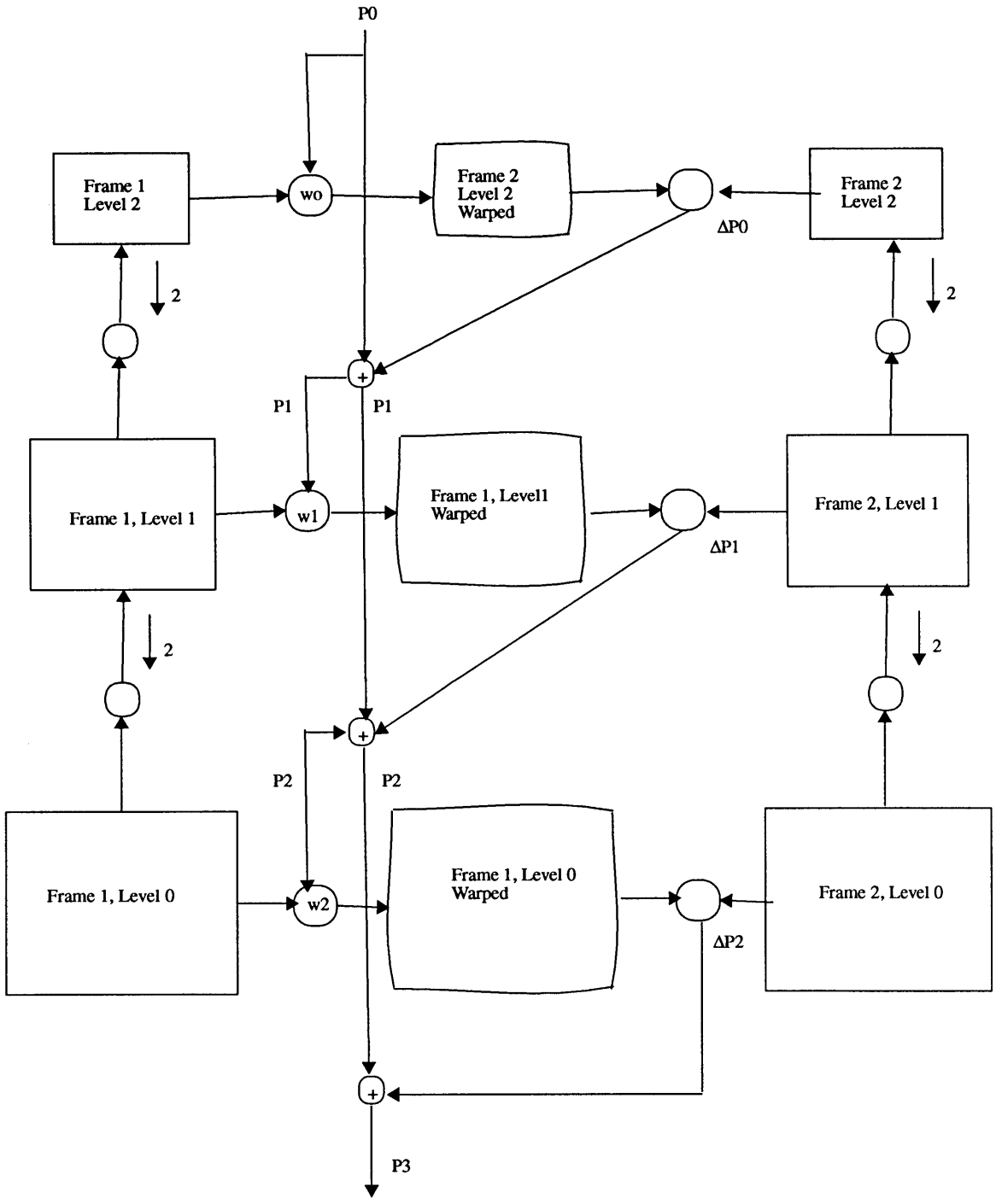


Figure 3.3 Coarse-fine tracking algorithm. Each frame is convolved with a low pass filter than subsampled. At the lowest level, frame 1 is warped by seed parameters P0. Motion estimation is computed between the warped frame and frame 2. The result ΔP0 is added to P0, resulting in P1 which is applied to frame 1, level 1. Motion analysis from the warped frame to frame 2, level 1 is now computed. The result is added to ΔP1 and so on up the pyramid.

gross camera movement. This works well for static scenes where there is only camera movement. Unfortunately, the real world of video is not filled with many shots of that type. There are often objects moving at drastically differing velocities in the image all of which may be moving at different rates than the camera. Therefore some measure of reliability must be built into the estimates.

3.1.3 Cleaning Up The Estimates

It is important to note that during the estimation procedure, the affine parameters determined at the base level (Level 2 of Fig. 3.3) highly characterize the movement for those frame pairs. Estimates at succeeding levels of the pyramid are only refinements of the estimate returned by its lowest neighbor. If noisy estimates are used in warping images into the space/time volume, they will introduce image artifacts.

A number of methods were applied in an attempt to retrieve an accurate set of affine parameters. The methods fall into three categories: masking input into the estimator, coaxing within the estimator, and cleaning up afterwards. Since the solving of the optical flow equations is a least squares problem, it is better to adjust the parameters as they are being formed rather than changing them afterwards. This ensures that a least squares solution will be found.

Masking

It is convenient to have the capability of masking out significant regions of the image which are moving differently than camera motion. This is a dilemma however, since we are attempting to analyze the image to recover camera motion and moving characters. But in order to mask an object, we must already know where the object is.

There are some ways to get an indication that there might be an object in the image sequence moving relative to the camera. A two-motion affine estimator was run first to see if there was a chance there could be a significantly moving second object in the scene. Mixed results were obtained with the motion estimator and more experimentation needs

to be done to find where its limitations lay. For this reason, masks were created by hand. The two motion algorithm is again from Bergen et al; a detailed explanation of it is given in the Appendix.

Brute Force Cleaning Up

If there is a high confidence in the camera motion, for example images produced by a camera rotating on a pivot at a constant rate, or a zoom shot with no translational component, the parameters are adjusted after they are returned from the estimator. If a zoom is expected, then the two scale factors, b_x and c_y, are made to be exactly equal and any spurious shear terms are set to zero. For a panning camera, the y translation terms are set to zero.

Spline

Sometimes in the image sequence, there may be sharp camera movement or other artifact in the video which causes the picture to jump. An ongoing analysis of the estimates as they are returned determines when such an anomaly occurs. The global camera motion between any two pair of frames of a contiguous segment should not change radically. Therefore a window of estimates before and after a given frame is compared. By assigning a threshold for change, any frames which have been created by unintended camera movement or by video artifacts are flagged. These frames are then removed from the sequence and the affine estimates between the remaining frames recomputed.

Covariance Matrix

The matrix equation solving the least square problem is shown in (Eq. 12). The 6 x 6 matrix on the left is the inverse of the covariance matrix. The main diagonal contains the variances σ_i^2 which are the average of (error in b_i)^2. The larger the number along the diagonals, the smaller the variances, and therefore the smaller the error. A quick way to limit the values of a particular number of affine parameters is to adjust the values along the diagonal, increasing in value the parameters which you would like to see allowed to

grow larger. Typically, a simple multiplicative term applied along the diagonal improves the estimates.

Spherical Projection

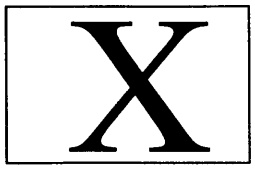
The light rays imaged on a film plane have been transformed by the lens through which they have traveled [Hech79]. There will be some change in perspective in the parts of the resulting image that lie far from the center of view. Therefore as an object moves from the center of an image towards its edges, it will be rendered differently on film. This amount of distortion depends on the focal length of the lens. If the distortion is significant, it presents a distorted image to the affine estimator. One way to counter-act this distortion is to warp the image so as to undo the lens effect by transforming the image from its linear projection into a spherical projection.

3.2 Warping Into Common Raster

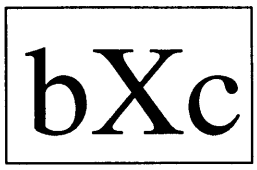
Once the motion between a pair of frames is determined, the second image can be warped back into the space occupied by its predecessor. This is done by inverting the motion between a pair of frames. For example, the second image of a pan left sequence will be adjusted right so the two frames line up; the second image of a zoom sequence will be scaled so it appears the two images were captured with the same focal length lens. (See fig 3.4)

Any selected start frame (Nstart) can be warped into the space of any target frame (Ntarget) by successively applying the inverse of all the affine motion estimates between the two frames. Successive application, however, is not optimal since it is computationally very costly. For each frame added, the computation time increases by N. Since each warp also causes a decrease in resolution due to interpolation of values and round off error, this method leads to image degradation.

Instead of warping each frame (|Nstart - Ntarget|) times, each frame should be warped

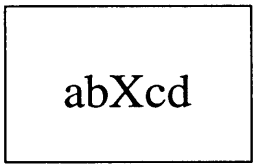


Frame 1



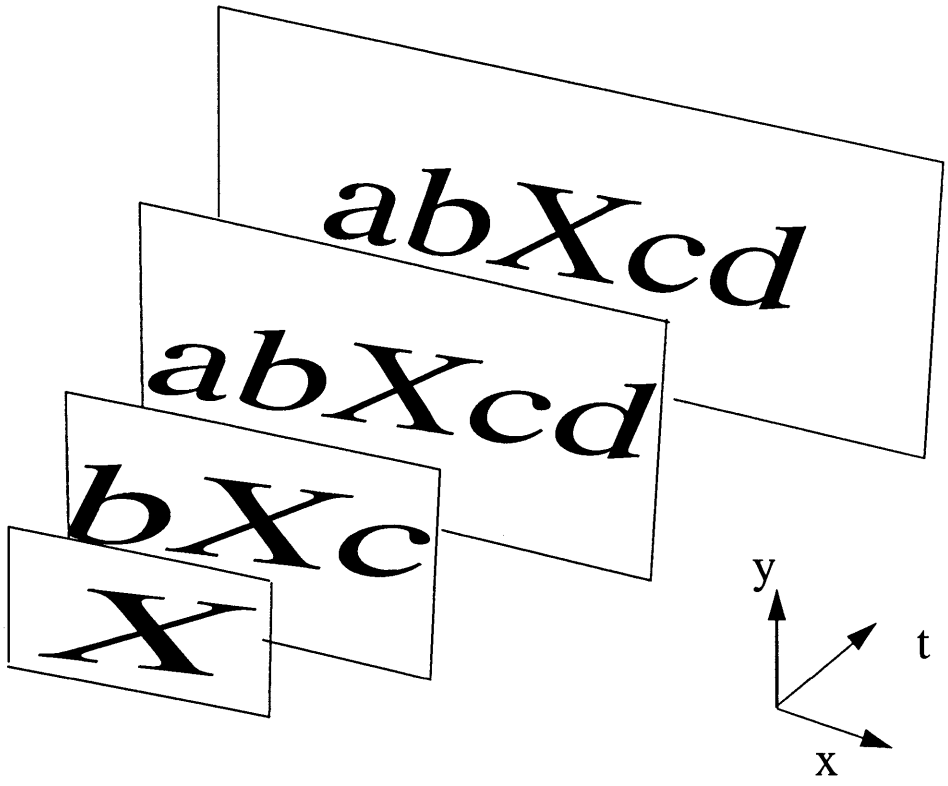
Frame 2

Frame 3



Frame 4

Fig. 3.4. The motion between the images at left was recovered and used to warp all the images into a common raster space as shown below. From this configuration, a high resolution image can be created by applying a temporal median over each of the pixels.



only once. One set of affine parameters was recovered by linearly summing all the intervening affine parameters together and then applying the inverse. This set of parameters completely describes the change from frame Nstart to frame Ntarget.

Affine Warping

The image warping used in this work both in the estimator and for later rendering is a bicubic spline interpolator operating on a 4x4 neighborhood. Both 'to' and 'from' warps were used at various points but in the end most of work was done with 'from' warps. 'To' warps model motion as a change from one start frame to the frame after it. So in the second frame, $I_2(x') = I_1(x+\Delta x)$. 'From' warps model the change in the first frame as a change from the second frame so $I_1(x) = I_2(x'-\Delta x')$. Accordingly, all the following equations in this document which describe the optical flow and rendering procedures assume a 'from' warp.

Summing and inversion of the two sets of affine parameters is shown below. The derivation of the summation and inversion equations is shown in the Appendix. Assuming a "from" warp between image pairs:

Summation:

$$
\begin{aligned}
ax_{new} &= ax_1 + ax_2 - cx_1 \times ay_2 - bx_1 \times ax_2 \\
ay_{new} &= ay_1 + ay_2 - cy_1 \times ay_2 - by_1 \times ax_2 \\
bx_{new} &= bx_1 + bx_2 - cx_1 \times by_2 - bx_1 \times bx_2 \\
by_{new} &= by_1 + by_2 - cy_1 \times by_2 - by_1 \times bx_2 \\
cx_{new} &= cx_1 + cx_2 - cx_1 \times cy_2 - bx_1 \times cx_2 \\
cy_{new} &= cy_1 + cy_2 - cy_1 \times cy_2 - by_1 \times cx_2
\end{aligned}
\qquad \text{(Eq. 13)}
$$

in matrix notation, the above is derived from:

$$\begin{bmatrix}\Delta x\\ \Delta y\end{bmatrix} = \begin{bmatrix}a_x\\ a_y\end{bmatrix} + \begin{bmatrix}b_x & c_x\\ b_y & c_y\end{bmatrix}\begin{bmatrix}x\\ y\end{bmatrix} \quad \text{(Eq. 14)}$$

$$\Delta p = A + Bp \quad \text{(Eq. 15)}$$

therefore

for the ax and ay terms:

$$A_{new} = A_1 + A_2 - B_1 A_2 \quad \text{(Eq. 16)}$$

and for the bx, by, cx, cy terms:

$$B_{new} = B_1 + B_2 - B_1 B_2 \quad \text{(Eq. 17)}$$

Inversion:

$$A_2 = A_1^{-1} = (B_1 - I)^{-1} A_1 \quad \text{(Eq. 18)}$$

$$B_2 = B_1^{-1} = (B_1 - I)^{-1} B_1 \quad \text{(Eq. 19)}$$

where, I is the identity matrix:

$$I = \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} \quad \text{(Eq. 20)}$$

3.3 Constructing Altered Scene

There are various approaches which can be taken in order to create the high resolution panoramic still from this space/time volume of data. The still could be pieced together bit by bit by using the pixels unaltered, or statistical operations could be performed over pixel regions in the space/time volume to over to arrive at a particular pixel value for the

resulting altered image.

3.3.1 Reconstruction Bit-by-Bit

The first approach is to piece bits together an altered image from each temporal sample of the space/time volume. In pan shots, the first imaged occurrence of each point in the world is used for the value of the corresponding point in the final altered image. For example, if the data in Fig. 3.1 (a pan-right shot) were used to create an extended scene, data for the region 'ABCDEFGH' would come only from frame 1, data for the region 'IJKL' would come only from frame 2 and data for the region 'MNO' would come only from frame 3. This is a simple and easy method for extracting an extended scene from mostly static data. Imagine, however, that the letters in the images represented a background in front of which a character was walking. (See Fig. 3.5). If the above method were used to render an extended scene, the random body pieces and their placement would create a chaotic final image. If the intention of the final salient still is to show the path of the moving character, it would be better to extract a background scene and then composite whole body pieces in a more structured and methodical manner.

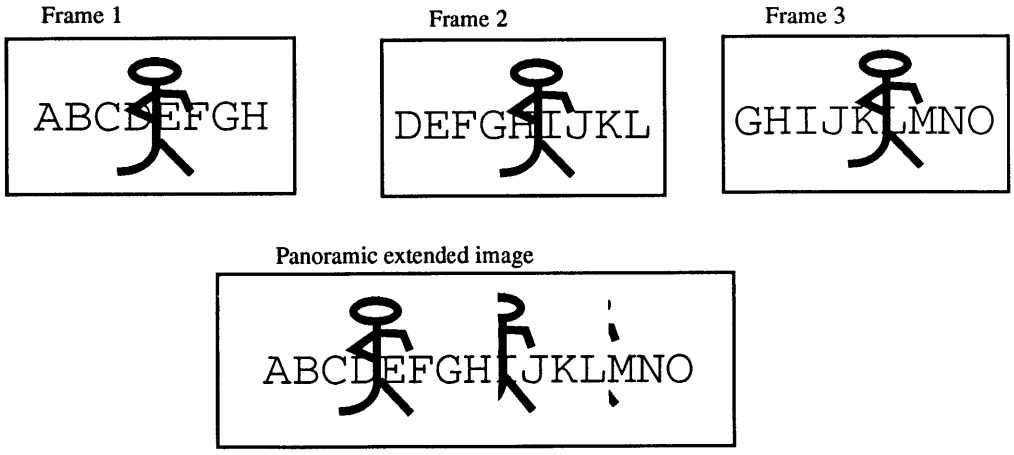


Fig 3.5 Three individual images from a pan-right sequence are shown on top. Warping the images into a common space/time volume and taking the first occurrence of each point in the world results in the extended scene shown on the bottom. A better apporach would to extract the backgournd scene then composite characters back in as desired.

This method of image synthesis often creates distinct artifacts where the image pieces abut. This is due to imperfect alignment of the segments or from actual luminance or chrominance changes in the image. For example, as a camera pans the room, altered lighting conditions or a changed lens setting may result in the same point in the world being imaged with different luminance values.

3.3.2 Temporal Weighting

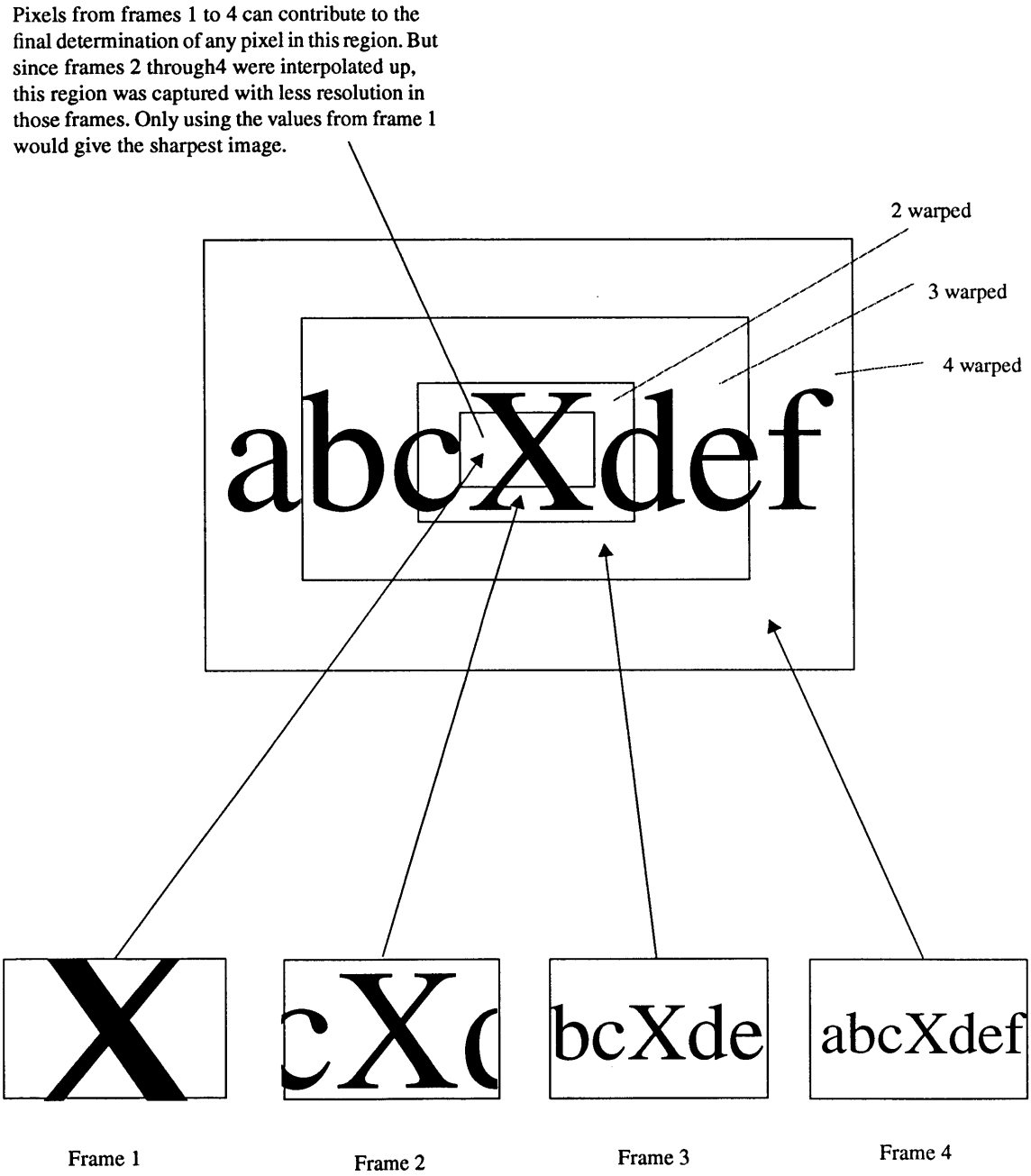
A potential remedy for the above problems is to apply an operator a region of pixels as they change over time to determine the corresponding pixel value in the final image. Ideally, this operator will preserve all of the detail found through out the sequence, while eliminating noise and extraneous motion in front of the camera. A temporal mean operation is not optimal since any noise in the image will be averaged into the result. Therefore, several temporal median operators were tried.

A conventional median over the temporal duration of the sequence successfully removes objects moving relative to the camera for more than 50% of the frames. A simple median is not optimal, however, when zoom sequences are being manipulated, since it gives equal weight to both pixels interpolated between coarse samples and pixels originating from fine samples. For each point in the world captured by a zoom, there is at least one frame where the point is rendered with the highest resolution. The other frames may contain interpolated values for that point. (See Fig. 3.6) Moving away from this frame in the temporal dimension causes the value of that point to be known with less certainty. These pixel values which are most temporally distant from the highest resolution frame should be weighted less (or not included) when computing the median.

Weighted Median

A linear weighting based on the position of each frame in the frame sequence depreciates

Figure 3.6 Synthetic image created from zoom sequence. The four boxed regions in the final image have been taken from the warped frames shown below.



interpolated pixels. Assuming the direction of the zoom is outward, the first occurrence of the pixel is weighted more that the last occurrence of the pixel. (See fig. 3.7).

This method still gives too much weight to the distant pixels, so a inverse weighting method can be applied. (See Fig. 3.8).Two methods of inverse weighting were performed. One was based on the position of each frame in the frame sequenece. A refinement, which accounts for variations in the pace and direction of the zoom assigns a weight according to the scale terms, b_x and c_y, from the affine estimates.

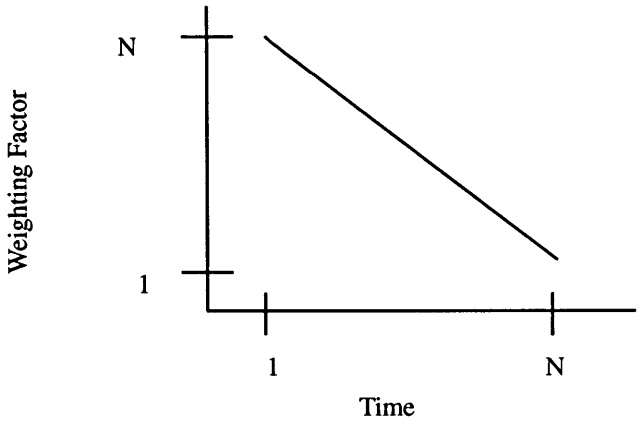


Figure 3.7 Linear weighting factor applied to a series of pixels which have multiple values in the space/time volume. The first occurrence of the pixel gets weighted more that the last occurrence of the pixel.

Envelopes

For some images, in addition of running a temporal median, an additional constraint was placed upon the computation of final pixel value. The number of frames which could contribute to the final pixel was limited by placing a window around these values. Moving this window allowed different temporal values to be used in the computation.

.

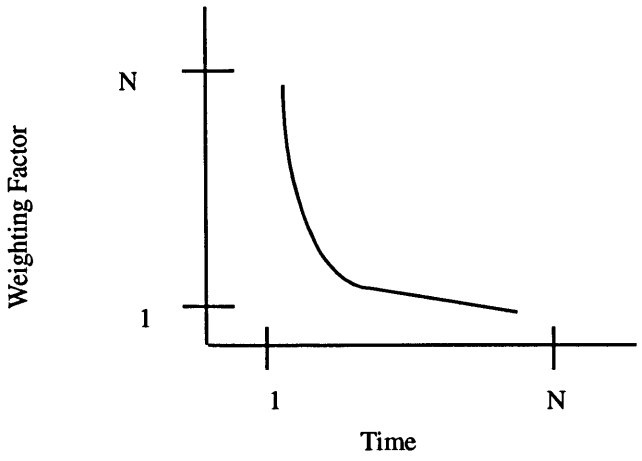


Figure 3.8 Inverse weighting factor applied to a series of pixels which have multiple values in the space/time volume.

3.4 Character Extraction

The task of character extraction was carried out by comparing scenes in the sequence with the actorless background scene that was created from the temporal median process. The regions that differ can be considered foreground objects and extracted. There are a few problems with this method. First, there is temporal variation in any pixel so that it may not match its corresponding pixel in the extracted background. These pixel variations are due to changes in the environment such as shadows cast from moving actors or from the methods of recording.

In most cases, the comparison with the background set results in a noisy segmentation. Therefore, a correlation with the background set was first computed and new and different objects detected. The user was then asked to assist in the segmentation.

3.5 Compositing

Once a background scene is extracted, characters or objects considered salient can be composited back into the set. These regions can be composited at higher resolution or

with changed color or luminance.

3.5.1 Blending Regions

These composited objects, especially if they are culled from other image sequences, which will have very different frequency characteristics than the background. In these cases some form of blending must occur so the images appear to be somewhat seamless.

The simplest approach to merging two images is to take a weighted average of pixel values over a transition region centered on the join line (a ramp function). The width of this transition region is a crucial factor. If it is too small, a blurred region will result from a mismatch in the low frequency components of the image. If the transition region is too big, features with the two images will appear doubly exposed. This is due to a mismatch in high frequency components of the images. If the images being joined have similar frequency characteristics, these effects will be less noticeable. For example, if the two images to be joined are one section of the same wall at different points in time and there has been no drastic lighting changes, the join region will be hard to detect. If the two images to be merged are as different as an apple and orange however, another approach is needed [Adel84].

Adelson et al. developed a laplacian pyramid blending method to overcome the frequency mismatch problem. Laplacian pyramid bands, each of which span one octave in spatial frequency, are created for the two images. The images are merged at each of the band levels over a constant transition range. Since the low frequency images are at a lower resolution, the low spatial frequencies get blended with a larger transition range and the high spatial frequencies with a smaller range.

3.6 Color

Humans are more sensitive to motion of achromatic information than chromatic information. Therefore, motion measurements are traditionally done on the channel of

information. Therefore, motion measurements are traditionally done on the channel of a digital image which represents the most luminance information [Netr88].

Since the channels carry different information, the estimates could be very different for each of them. In order to assure a unique estimate between frame pairs, motion estimation is only done for the luminance component.

In all the examples discussed in this thesis, the motion estimates, warping and median operations are done on the Y channel of a YIQ representation of the image sequences. The color channels are warped with these same affine parameters. The temporal median process is run on the Y channel as well. The corresponding pixels in the IQ channels are chosen as the median pixels for those representations.

4 Applications

4.1 Why salient stills?

Salient stills and their methodology are useful in the areas of compression and visualization of temporal data. The examples given in this thesis are examples of salient stills as applied to a particular type of moving data: video, but the concepts can be applied to the visualization of other forms of changing data as well. Below are some examples and scenarios of salient stills and their intermediate technology.

4.1.1 Video to Print

Sometimes it is desirable to have the ability to print an still image from video. The quality of video is below that of print medium. Another problem is that the image collector is capturing for a temporal medium, following movent often takes precedence over frame composition. There might not be one frame composed well enough to stand on its own as a still. Other times there is information in the moving image which would be distracting or misleading in a still frame.

Salient stills can be built which attempt to retain the content (detail) and context (spatial extent) of the moving image material. If there is a zoom sequence in the moving image, there might be regions of the still with enhanced resolution. This variaiton in resolution rather than being disconcerting draws the viewer's attention to the parts of the scene which commanded the attention of the camera operator or the creator of the salient still.

Figs. 5.6 - 5.9 of chapter 5 are examples of stills which have enhanced resolution and attempt to capture the salient features of a moving image sequence.

4.1.2 Limitations of Display Technology

Unfortunately moving visual information can not be displayed on laptop computers (and even some workstations - thought this is becoming less true) due to memory and display architecture limitations. Yet temporal data such as traditional video mass media should be available in some form to users of portable computers.

For example, in a hand held computer news retrieval system, salient stills could be augmentation to a text story, much like an information-rich photograph of conventual newspapers. Or they could represent a complete story in your news system, an icon of things to come.

In either case a representation must exist so that this data can be extracted and automatic composition can take place. Many of the frames of a traditional television newscast are redundant and offer little visual information: for example, thousands of similar frames of the tight close-up head shot of the evening's anchor. One frame would suffice to set the context of the night's broadcast. An example of a still created for the portable news system is shown in Fig. 5.10. A more detailed description of it is given in chapter 5.

4.1.3 Iconic Depictions of Moving Images

Retrieving a piece of video from an image database requires description, searching, and viewing. Depictions of moving image data fall into two classes, lexical and visual.

The classic example of lexical depictions is the "logging sheet" - reams of paper (sometimes computerized) which describe the contents of moving video in chunks or segments of semantic significance. This representation of a series of frames is brittle since it allows no ability to search for components within a frame. In addition, lexical

descriptions can be ambiguous. Three different people logging a piece of footage will describe it in three different ways. New methods which describe the temporal dimension of significant elements within frames and begin to address the ambiguity of text depictions are being explored[Agui91][Dave91]. Even though lexical descriptions are ambiguous, they are necessary for some computer searches of data. They could be close to useless, however, for someone who has not previously viewed the image data. Some representation closer to the image form is necessary to visualize the image content.

Moving data can also be depicted visually - either through some graphical icon or some abstraction of the video itself. Logging video footage using icons [Davi91] is a powerful and economical way to enter a multitude of descriptive attributes into a video database. Unfortunately, icons are an inconvenience when the data must finally be retrieved. Two levels of abstraction must be decoded - first the icon itself and then its relation to the footage. In the end, the actual footage must be seen for retrieval, but often there are hundreds of hours of footage to search. It could be inconvenient to view every minute of footage that matches a particular query. Is this there some middle ground between depictions that are too abstract (text and cartoon-like icons) and viewing hours of footage?

Perhaps some abstraction of the video signal itself is needed. One possibility is to use every shot boundary - a place where continuous recording of some contiguous action in time and space has stopped. Shot boundaries are fairly easy to detect [Sasn86] [Bend89] [Ueda91]. However, in certain types of filming, such as cinema verité, continuous shots can last as long as 1/2 hour. The end points of these shots would not be significant to any viewer of the footage.

Small segments of moving images, moving icons (MICONS), also have been used to depict a larger segment of video. Since they are very effective in drawing a user's attention, they have been used to assist in navigation through an interactive multimedia presentation [Bron89]. Conceivably, they can be used to depict the contents of a large

image database. The segments selected must be chosen with care so that a naive viewer of the footage can infer the contents being depicted.

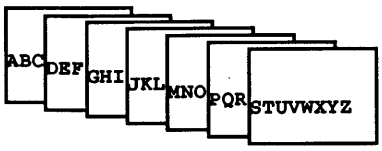
But the problem still remains, that micons can only show information sequentially where salient stills can condense the temporal dimension into one view. If for example, there had been a significant camera pan in a section of footage, the salient still image would be wide enough to display the whole space unwrapped, or if significant zooms were involved, the image would have greatly increased dimensions. Color, altered resolution, translucency, size, location, shear, blending and multiple images could be used to convey salient or changing characters in the image. A study comparing the effectiveness of both of these methods is worth considering.

4.1.4 New Views and Vision

Some researchers, such as visual anthropologists use video as a means of collecting data for their work. Often the context of the whole space they shoot provides vital clues to their investigation; for example, the distance between objects in a house or the proximity of people at a gathering. Sometimes, the physical aspects of the space make it impossible to capture all of this data at once. In order to get one single image of even half of a room, a very small focal length (very wide angle) lens would be needed. But an image captured with this lens would have large aberrations around the edges of the frame (and even deep into the center of the image for some lenses). If the entire 360 degrees of the room is desired, this becomes an impossible task with conventional photographic equipment.

One approach to wide-angle capture would be to use a Glubuscope or other specialized camera to capture the whole space of the room. This technique poses limitations in that the camera only moves in the x direction (or in the y direction, if turned on its side). Because sites are not always known thoroughly before the shooting starts, it would be ideal to be able to create an image of the whole space directly from the data recorded by a conventional video camera.

The image data of the space could be unwrapped into a single frame which retains the contextual information of the imaged area. (See fig 4.1).

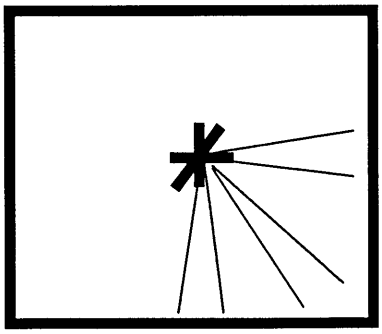


frames from a pan

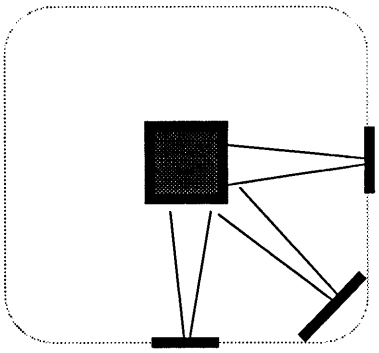
panoramic still

Figure 4.1 Salient still from pan. Repeated image transformations and temporal median filtering is used to create the panoramic still.

One could also imagine that spaces could be imaged not only 360 degrees out from the focal center of a lens but also that the camera could image 360 degrees pointing in to some focal center, for example, around a building as shown in figure 4.2.



Imaging 360° out from a focal center.



Imaging 360° around a building or similar object.

Fig 4.2 Panoramic images can be created from both imaging techniques.

4.1.5 New Sequences and Visualization

For some sequences, once the images are warped into a 3-dimensional representation of the image data, one can attempt to visualize how the same space would have looked if the camera had traversed another path. The methods this thesis provides are immediate, though limited to the 2-D world. While some of these salient stills will not be perspectively perfect, they certainly would give a sense of what the image could have looked like if shot differently.

4.1.6 Display of Moving/Changing Information

The tools and ideas of this thesis were applied to creating single frame representations of moving images but certainly some of the ideas can be applied to representing other forms of changing information in a visual manner. For example, changes in the location or actions of a person could be represented by compositing and manipulating facsimiles of them on a representative background. This positioning information can come from many different sources, not just visual. For example, information from computer systems (like the 'finger' or 'who' command of Unix) or systems which track a user's physical presence, could be visualized [Watl89]. The renderings do not have to be realistic, but can be metaphoric as the compositing methodology allows.

One could imagine also a sort of video radar, where a camera will continuously film an entire room by panning around 360 degrees. Whenever data is encountered which is not a permanent fixture in the room, this data will be realistically imaged onto a static background image of the space much like the output of conventional radar. The tools of this thesis can be used to initially create the extended background scene. Using this, any new activity in the room could be tracked.

The compositing methodology could be used to highlight characteristics about the activity - time person entered the room, the urgency with which they should be noticed.

4.1.7 Scene Retrieval by Content and Self-Similarity

Sometimes, in video retrieval the goal is to find a set of frames that are like some other set of frames. They may be shot in the same space or have similar camera movements. A structured representation can be very effective in this endeavor. Traditional logging is already successful at getting us to footage of the "White House Blue Room", but can it tell us the exact frame that a particular stultifying cabinet member has stepped up to the microphone? This is where a more structured representation helps. If one "truly" know the background scene then one can search the image itself for when the member has reached his destination.

A more elusive property for which to search in a text description is camera movement. For example, during hundreds hours of theater performance footage, there is one shot an editor would like to find immediately. He remembers that he zoomed in on some brunette beehive on the right side of the audience, panned across to the wailing child in the front row and then tilted straight up to the villain, center-stage. The video doesn't even have to be logged for this search to begin. Those camera properties are described by the affine description of motion used here. One could trace out the probable path and then have the computer search for closest matches. The optical flow estimator can also categorize shots if someone was interested in doing some camera style determination or analysis of a group of films.

4.1.8 Photomontage -- Ease of Manipulation

Photomontage has existed for hundreds of years. Traditionally, one would point a camera at a scene, click, move the camera over and click again. The results would then be pasted together. Whatever was captured on film is what the creator of the montage had to use. Of course the space could be photographed again if the results were unsatisfactory. Rather than be limited to the results of conventional photographic images, the tools of this thesis allow a space to captured with traditional video equipment. Warping the captured images into 3-space, allows a user to traverse the space and select sections of different

images to be placed the final synthesized image. For some portions of the final image whole frames might be used, or for some sections the user would rather a temporal average of the data be used, thereby removing moving objects from that part of the scene.

4.1.9 Human Manipulation of Video Space

A 3-dimensional representation of a continuous scene once created, becomes a space a user can move through. The raster size for the individual frames is arbitrary and depends on the traversal of the camera. At any give point in moving through this space, the user can choose to go forward in time or move in some spatial dimension.

5 Images

5.1 Stadium Space

Space and perspective through time are depicted in a salient still created from 25 seconds of a pan sequence around the Hubert H. Humphrey Metrodome in Minnesota during the 1985 Baseball All-star Game[1].

Physically located directly above and behind home plate, the camera pans around the entire stadium. Since there is no tracking, the camera's axis of rotation stays constant which reduces the perspective distortion from frame to frame. While the image transformation between frames is not entirely affine, for the closely sampled frames of this sequence, the transformation can be approximated as affine.

The goal of the image is to allow the viewer to experience the entire space at once. This still is an image with the temporal information retained. Looked at in one glance, it is a stadium unwrapped from the inside out, but it forces your eye to move over it just as the camera did There is an inverted vanishing point one-third of the way in from the left of the frame. This is initially where the eye focuses and then moves in either direction.

The affine estimation of motion was computed on a small central stripe of the image. This was done for 3 reasons - to minimize effects of perspective changes, to minimize the effects of non-spherical projection and to save resources.

1. The National League defeated the American League 6-1.

Fig 5.1 Unwrapped stadium pan

If it obvious from the image that camera panning through space has occurred. The direction of this movement could be indicated by a higher resolution stripe or changes in color and intensity.

Changing Perspective / 3-D Model

In the image shown, the unwrapped projection did not undergo any transformations. Such changes could be made only if a 3-D model of the space could be recovered.

One way to approach the shape recovery of the stadium from this image is to assume some constant size of certain features of the stadium. For example, if the distance from the first mezzanine to the top of the stadium is a constant height, then that height in the image, α in fig. 5.2., is inversely proportional to the distance of the camera from that surface. This assumes constant camera motion during the image capture.

From analyzing this stripe in the image, the shape of the stadium could be plotted in polar coordinates as shown in figure 5.2.

Once a shape is recovered, the image can be warped into 'non-depth' space where the

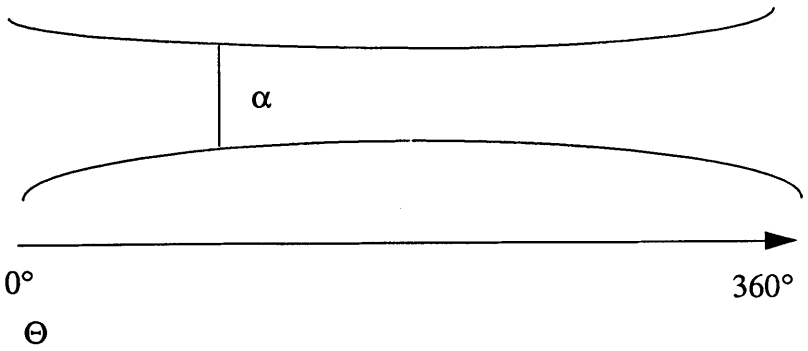


Fig. 5.2 For a 360° pan around a stadium starting at 0°, α is the height of a constant size object as rendered in the extended image.

size of the assumed constant size object is indeed a constant size. This image now can be warped into a new perspective, say for example instead of the camera being behind homeplate, it can be at second base. Then it becomes a matter of sampling the correct pixels horizontally and warping them the correct amount in the y direction. Depending on the viewpoint, some data may have to be interpolated. The amount to warp the scan lines can be found by computing new 1/α for each Θ by moving the origin of figure 5.2. to the new perspective point and mapping polar coordinates to polar coordinates.

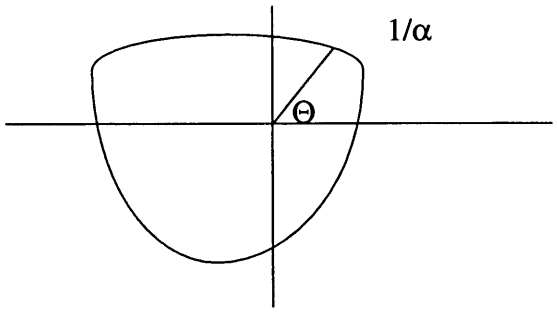


Fig. 5.3 Stadium shape recovered from extended image.

5.2 Yoyo and the Disappearing Man

Four frames from a 12 second zoom sequence of the performer Yoyo Ma on stage during a recent performance at Tanglewood[1] are shown in figure 5.4. The task at hand was to create a still of the performance which would have enhanced resolution and preserve the salient elements of the image sequence.

One may initially be inclined to chose the close-up frame of Ma since this is rendered with the most detail per pixel per area of "interest". But the image does not provide the contextual information of the ambiance of Tanglewood nor does it relay that there is a live audience. In contrast, the far shot retains the ambiance but does not provide enough spatial detail of the performer to recognize that it is Ma. A frame selected from the middle of the zoom contains the assistant walking across stage, detracting from Ma as the center of focus.

The solution chosen to create an image with both enhanced visual quality and semantic saliency is not to use any single frame, but to use all of them. The close up frame renders Ma with the needed detail, the far shot gives us context, and the middle shots provide enough redundancy to assist in the removal of the non-salient musical assistant. If this character walking across stage were an overcome fan intent on displaying his affection to Ma, we could easily select his position in one frame and render him into the scene. More interestingly, we could show his traversals over the stage by displaying him in multiple times in different sizes, shapes or shades. Figure 5.5 compares the resolution of the far shot with the shot used in the still.

1. Yoyo Ma performing *Begin Again Again...* a piece composed for hypercello by Tod Machover of the MIT Media Lab, August 15, 1991, Tangelwood, MA.

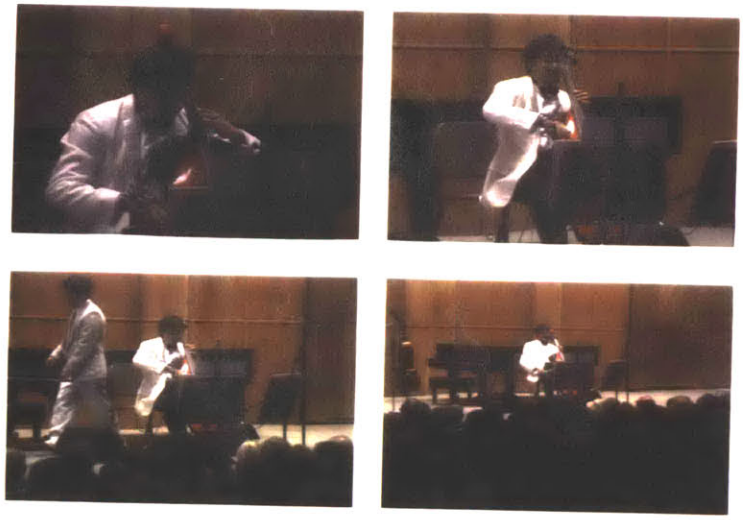


Fig. 5.4 Four frames from zoom sequence.

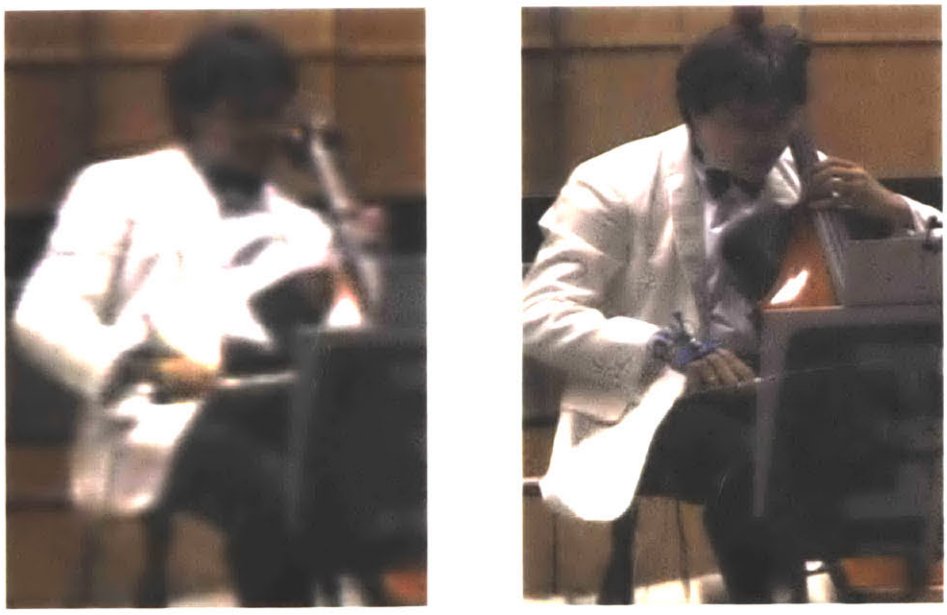


Fig. 5.5 The left image is the resolution of the short focal length (far shot), the right image is the resolution achieved by using the long focal length (tight shot) at the center of the salient still.

Method

The first task was to recover the camera motion of the sequence. The affine motion estimator worked accurately until the musical assistant entered the frame. Since the estimator was working over the entire frame, his differing movement was computed in the global camera estimate. This resulted in erroneous estimates. In order to compensate for his motion, a mask was created to eliminate him from the computations. A comparison of the zoom components of the affine parameters for the masked and unmasked sequence is shown in figure 5.6.

Once a set of estimates was retrieved, the images were warped into a common large raster. At this point, two approaches were taken to creating the still. The first method was to hand pick 3 frames which would be compose the final synthesized image. This prevented the problem of having to remove the unfortunately unwanted musical assistant but it presented many more problems in blending. The other approach was to synthesize the still by using 70 frames of the sequence (11 seconds with a frameskip of 10). The optical flow estimate is most accurate if every frame of a sequence is used in the motion estimation. However, if the change between frames is small, a coarser sampling of the sequence is adequate to accurately characterize the camera motion. Thus the sequence sampling rate could be made a function of the affine parameters, though this was not implemented in this work.

3 Frames Approach - Color Correction

The 3 frames were warped into common raster space as described in chapter 3 and then were overlaid on each other. Great discontinuities were noticeable in the overlap regions from luminance and resolution mismatch. For this particular image sequence, the close-up frames rendered the wall behind Ma darker than in the far shots. This may have been due to filming in automatic aperture mode where tight shots of Ma's white suit caused the lens aperture to close down. Whatever the cause, it was remedied by measuring pixel points around the join regions of the frames and creating a spline which would

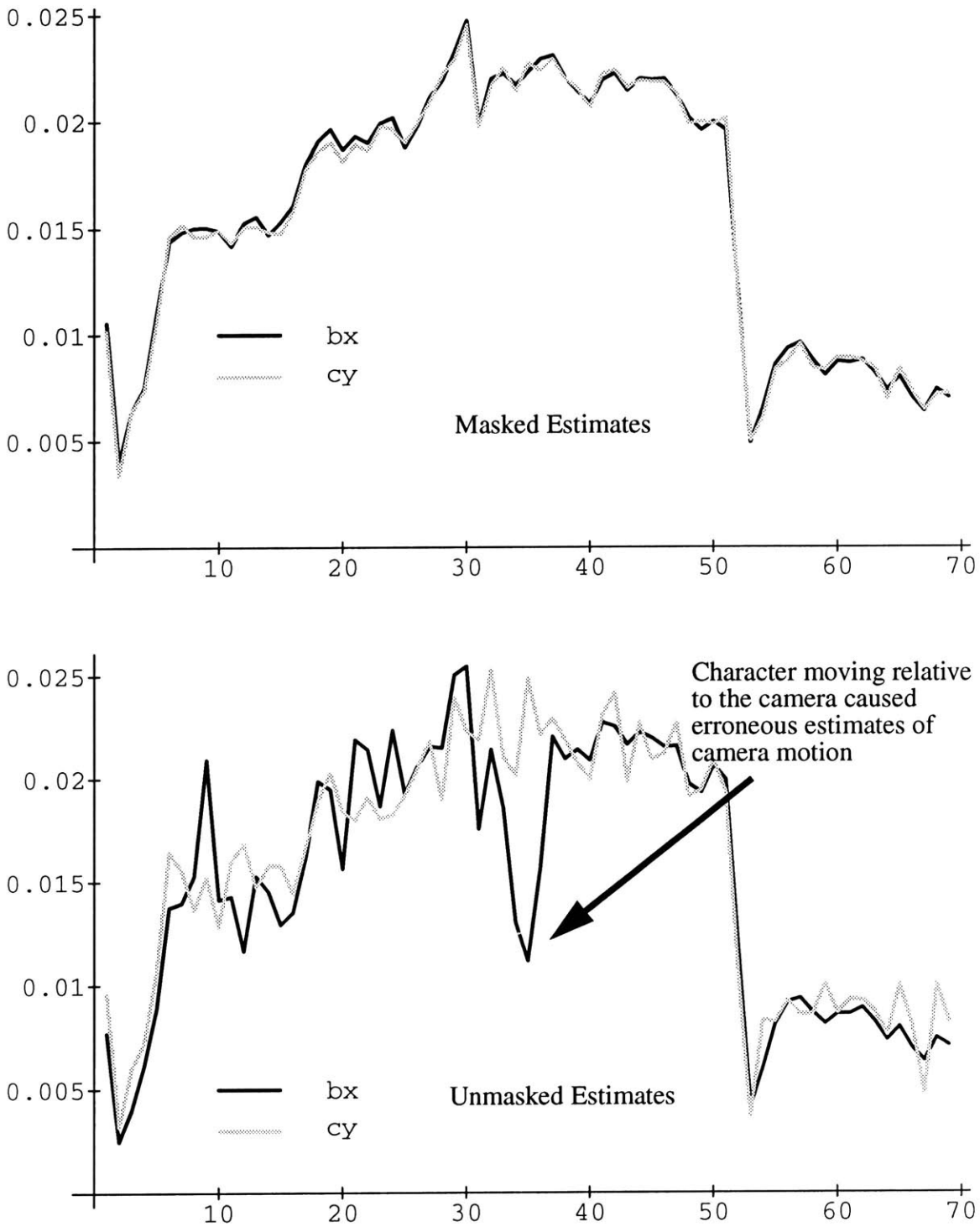


Fig 5.6 Comparison of zoom components, b_x and c_y, of the affine parameters from masked and unmasked Yoyo sequence.

interpolate the luminance values. In this image, the two larger frames were adjusted so the luminance values matching corresponding regions in the close-up frame of Yoyo Ma. In this way, the detail of the region of interest had not been changed. (See Fig. 5.7).

The spline was fit to the luminance channel only and the color channels were simply used untouched. This was successful since the luminance channel carries most of the spatial detail for which the eye is sensitive.

In order to merge the three images, both a ramp function and laplacian pyramid blend were applied to the abutting regions as discussed in chapter 4.

Fig 5.7 Salient still from 3 frames.

Temporal Weightings

The next set of stills were created by using 70 frames from the zoom sequence. This allowed an even greater gain in resolution and the ability to automatically remove

Fig 5.8 Salient still using linear weighted temporal median.

Fig 5.9 Salient still using inverse weighted temporal median.

moving characters. Four images were made from the space/time volume of this image sequence. One was made with a conventional median, the other 3 with forms of temporal medians.

Figure 5.8 shows the image created with a conventional linear temporal median, and the focus region inset. The moving characters were successfully removed but the median caused a considerable lightening of the background. The focus region placed in at the end is clearly distinct. A luminance transformation could be performed on this center region but it is undesirable to decrease the light value of focus regions. Additionally, since a pure linear transformation was computed, an unnecessary decrease in the resolution of pixels closer to the center of the image occurred as described in fig. 4.4.

Figure 5.9 shows the image created from weighting using a inverse method. The result is an imperfect segmentation of the musical assistant but the image

Figure 5.10 shows a still of the same sequence where the musical assistant was added back into the image multiple times to show his traversal across the stage. His transparency is a function of his time in the scene. The earlier he appeared the scene, the more transparent he is in the composite.

Fig 5.10 Salient still with musical assistant composited back in

5.3 Hussein and the King

The image in figure 5.11 is an attempt to create a salient still from a short documentary clip. It may be used as an iconic representation of video database material or it may be displayed on a portable news information server as a visual summary of a news story.

This image is from a video sequence of two middle east experts discussing the similarities of Sadaam Hussein and a past king. At one point, the discussion turned to a painting Mr. Hussein had commissioned showing the two of them together. The camera panned across the image while the experts discussed the psyche of Mr. Hussein.

First the image sequence was spliced into scene units by analyzing the digital video signal. The composite would consist of a extended background of the painting retrieved by unwrapping the pan. Head shots of the middle east experts would be superimposed on that background.

The movement of the pan sequence was recovered and the extended image create. For each of the head shots of the middle east experts, the text captioning describing who they were was extracted. The head shots were then resized to make them proportional and the text caption was laid back into the image. These images were then placed into the extended image of the painting and a laplacian blend was performed.

The images for this still were selected externally, but one could imagine a situation where a computational agent could perform this manipulation automatically.

Fig 5.11 Salient still for portable news system.

6 Conclusion

This work addresses some of the problems of transforming moving images into still. A new type of image is created which combines a multiplicity of frames and image objects. Views not available with conventional lenses are possible such as panoramas around and into a space and images with high resolution patches.

Studies should be conducted to see how users react to such images. What is the best way to simulate movement and importance? Do enhanced resolution sections of an image draw a users attention? Is it possible to create a salient still that is at least or more effective than a moving icon? Does warped perspective make any difference in comprehension of the image? Are their cultural difference in understanding the methods used to describe motion and saliency? Is there a certain level of visual literacy one must have in order to understand these images.

Work should be done in exploring the navigation potential of both the intermediate video space/time volume and the salient stills, particularly the panoramas. That is, these images offer ways for a user to have random access into the larger body of video data from which the salient stills were derived. The space/time volume gives indications to the camera and object movement over the temporal duration of the frames and the panoramas provides spatial context.

Appendix

6.1 Derivation of Linear Sum of Affine Parameters

For affine parameter modelling of motion:

$$\begin{bmatrix}\Delta x\\ \Delta y\end{bmatrix} = \begin{bmatrix}a_x\\ a_y\end{bmatrix} + \begin{bmatrix}b_x & c_x\\ b_y & c_y\end{bmatrix}\begin{bmatrix}x\\ y\end{bmatrix} \qquad \text{(Eq. 21)}$$

$$\Delta x = A + Bx \qquad \text{(Eq. 22)}$$

Summation:

For 2 consecutive warps, $\Delta x1 = A1 + B1x'$ and $\Delta x2 = A2 + B2x$.

$I_2(x') = I_1(x' - \Delta x1)$

$I_3(x) = I_2(x - \Delta x2) = I_1(x - \Delta x2 - \Delta x1)$

$= I_1(x - \Delta x2 - A1 - B1x')$

$= I_1(x - A2 - B2x - A1 - B1(x - A2 - B2x))$

$= I_1(x\,[- A2 - A1 + B1A2] + [-B2 - B1 + B1B2]x)$

The terms in brackets are the combined terms: A1,2 and B1,2

Inverse:

Set A1,2 and B1,2 to zero.

$IA2 + A1 = B1A2$

$[B1 - I]\ A2 = A1$

$A1^{-1} = A2 = (B1 - I)^{-1}\ A1$

$B1^{-1} = B2 = (B1 - I)^{-1}\ B1$

where I is the identity matrix: $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

6.2 Bergen et al. two motion estimator

From [Berg91].

In this method the image I(x,y.t) is modeled as a combination of two distinct patterns, P and Q, having independent motions of p and q. The relationship between I and P and Q may be stated as follows:

$$I(x, y, 0) = P(x, y) \oplus Q(x, y) \qquad \text{(Eq. 23)}$$

and

$$I(x, y, t) = P^{tp} \oplus Q^{tq} \qquad \text{(Eq. 24)}$$

where the operator $\oplus$ represents an operation to combine the two motions such as addition or multiplication and Ptp represents pattern P transformed by motion p through time t. Bergen et al. show that if one of the motion components and the combination rule $\oplus$ are known, it is possible to compute the other motion using the single-component motion technique discussed above, without making any assumptions about the nature of the patterns P and Q. If the motion p is known, only the motion q must be determined and vice versa. The components of the pattern P moving at velocity p can be removed from the image sequence by shifting each image frame by p and subtracting the shifted frame values from the following frame. The resulting difference sequence contains only patterns with velocity q.

Typically, the combination operator ⊕ is addition. Considering three frames, I(1), I(2), I(3) and assigning the variables D_1 and D_2 to the difference frame3s generated between those pairs of frames respectively, (Eq. 12) and (Eq. 23) become.

$$D_1 \equiv I(x, y, 2) - I^P(x, y, 1) \quad \text{(Eq. 25)}$$

$$D_1 \equiv I(x, y, 2) - I^P(x, y, 1) \quad \text{(Eq. 26)}$$

Difference image D_1 is computed by warping I(1) to transform pattern P through one step, followed by a subtraction of I(2) to remove the effect of the motion of pattern P. D_2 is formed similarly by subtracting I(2) from I(3).

The modified sequence now consists of a new pattern Q^q - Q^p, moving with a single motion q.

$$D_n = (Q^q - Q^p)^{nq} \quad \text{(Eq. 27)}$$

Thus, the motion can be computed between the two difference images D1 and D2 using the single motion estimation technique described above. Analogously, motion q can be recovered when q is known. The observed images I(x,y,t) are shifted by q, and a new difference sequence is formed

$$:D_n = I(x, y, n+1) - I^q(x, y, n) \quad \text{(Eq. 28)}$$

This sequence is the pattern P^p- P^q moving with velocity p:

$$D_n = (P^p - P^q)^{np} \quad \text{(Eq. 29)}$$

so p can be recovered using the single motion estimation.

This shift and subtract procedure removes one moving pattern from the image sequence

without regard to, or determining what that pattern is. In practice neither p or q is known at the outset. However, both can be recovered through the above techniques if a rough estimate is given initially for either p or q. Iteratively, the single motion estimator is applied first to compute p and then q.

Bibliography

[Adel84] Adelson, E.H., Anderson, C.H., Bergen, J.R., Burt, P.J. Ogden, J.M., 'Pyramid methods in image processing', *RCA Engineer*, Nov/Dec 1984.

[Adel85] Adelson, E.H., Bergen, J.R., 'Spatiotemporal energy models for the perception of motion', *Journal of the Optical Society of America*, Vol. 2, No. 2, February 1985.

[Agui91] Aguierre Smith, Thomas, 'Stratification: Toward A Computer Representation of the Moving Image', *Working Paper from the Interactive Cinema Group*, MIT Media Laboratory, 1991.

[Bend89] Bender, W., Chesnais, P., 'Network Plus', *Proceedings, SPIE Electronic Imaging Devices and Systems Symposium*, Vol. 900, January 1988.

[Berg90] Bergen, J.R., Burt, P.J., Hingorani, R., Peleg, S., 'Computing Two Motions from Three Frames', David Sarnoff Research.

[Bove89] Bove, V.M., 'Synthetic Movies Derived from Multi-Dimensional Image Sensors', PhD Thesis, Massachusetts Institute of Technology, June 1989.

[Brondmo89] Brondmo, H.P., Davenport, G., 'Creating and Viewing the Elastic Charles - a Hypermedia Journal', *Hypertext II Conference Proceedings*, York, England, July 1989.

[Burt84] Burt, P.J., Adelson, E.H., 'The Laplacian Pyramid as a compact Image Code', *IEEE TransCom*, COM-31(4) (1983).

[Cahi82] Cahill, James, *The Compelling Image: Nature and Style in Seventeenth Century Painting*, Harvard University Press, Cambridge MA, 1982.

[Cla88] Claman, Lawrence N., 'A Two Channel Spatio-Temporal Encoder', Master's Thesis, Massachusetts Institute of Technology, May 1988.

[Colb91] Colby, G., Scholl, L., 'Transparency and Blur as Selective Cues for Complex Visual Information', *Proceedings of SPIE Conference on Image Communications and*

Workstations, Santa Clara, CA, February 1990.

[Dave91] Davenport, G., Aguierre Smith, T., Pincever, N., 'Cinematic Primitives for Multimedia', *IEEE Computer Graphics and Applications*, July 1991.

[Davi91] Davis, M.E., 'Director's Workshop: Semantic Video Logging with Intelligent Icons', *Proceedings of the AAAI-91 Workshop on Intelligent Multimedia Interfaces*, July 1991.

[Doyl86] Doyle, T., Frencken, P., 'Median Filtering of Television Images', *Technical Papers Digest, International Conference on Consumer Electronics*, June 1986.

[Dunn91] Dunning, William, *Changing Images of Pictorial Space: A History of Spatial Illusion in Painting*, Syracuse University Press, 1991.

[Edge87] Edgerton, H.E., *Stopping Time: the photographs of Harold Edgerton*, N.H. Abrams, New York, NY, 1987.

[Frie80] Friedman, S., Stevenson, M., 'Perception of Motion in Pictures' in *The Perception of Pictures Vol. I*, M. Hagen. ed., Academic Press, New York, 1980.

[Fry66] Fry, E.F., *Cubism*, Oxford University Press, New York, 1966.

[Giro89] Girod, B., Kuo, D., 'Direct Estimation of Displacement Histograms', *Optical Society of America Meeting on Understanding and Machine Vision Proceedings*, Cape Cod, MA, June 1989.

[Hage91] Hagen, M., *Varieties of Realism: Geometries of Representational Art*, Cambridge University Press, New York, 1986.

[Hech79] Hecht, E., Zajac, A., *Optics*, Addison-Wesly Publishing Co, Reading, MA, 1979.

[Heeg87] Heeger, D.J., 'Optical Flow Using Spatiotemporal Filters', International *Journal of Computer Vision*, 279-302 (1988).

[Holt90] Holtzman, Henry, 'Increasing Resolution Using Pyramid Coding and Pattern Matching', unpublished technical memorandum, MIT Media Laboratory, 1990.

[Holt91] Holtzman, Henry Neil, 'Three-Dimensional Representations of Video using Knowledge Based Estimation', Master's Thesis, Massachusetts Institute of Technology, September 1991.

[Horn81] Horn, B., Schunk, B., 'Determining Optical Flow', *Artificial Intelligence*, 17, 1981.

[Horn86] Horn, B., *Robot Vision*, The MIT Press, Cambridge, MA, 1986.

[Hu87] Hu, A., 'Automatic Emphasis Detection in Fluent Speech with Transcription',

Journal of the Acoustical Society of America, Vol. 58, No. 4, (1975).

[Jain81] Jain,J., Jain, A., 'Displacement measurement and its application in interframe image coding', *IEEE Transactions on Communications*, COM-29, 1981.

[Joyc88] Joyce, P., *Hockney on Photography*, Jonathan Cape Ltd., London.

[Keit88] Keith, J.M. et al, Digital video compression system, U.S. Patent 4,785,349, Nov. 15, 1988.

[Lawd75] Lawdler, S.D., *The Cubist Cinema*, New York University Press, New York, 1975.

[Limb75] Limb, J., Murphy, J., 'Estimating the velocity of moving images in television signals', *Computer Graphics and Image Processing*, 4, 1975.

[Lipp91] Lippman, Andrew, 'Feature Sets for Interactive Images', *Communications of the ACM*, April 1991.

[Mack86] 'Automating the Design of Graphical Presentations of Relational Information', *ACM Transactions on Computer Graphics*, Vol 5, No 2, April 1986.

[Marr79] Marr, D., Poggio, T., 'A computational theory of human stereo vision', *Proceedings of the Royal Society of London*, B-112, 1981.

[MPEG] MPEG Draft Proposal (Motion Pictures Experts Group), ISO/IEC JTC1/SC2/ WG11, September 1990.

[McLl91] McLean, Patrick Campell, 'Structured Video Coding', Master's Thesis, Massachusetts Institute of Technology, June 1991.

[Netr88] Netravali, A., Haskell, B. *Digital Pictures, Representation and Compression*, Plenum Press, New York, 1988.

[Sasn86] Sasnet, Russell, 'Reconfigurable Video', Master's Thesis, Massachusetts Institute of Technology, February 1986.

[Scho91] Scholl, L., 'The Transitional Image', MIT Master's Thesis, Massachusetts Institute of Technology, September 1991.

[Silb82] Silberbergeld, Jerome, *Chinese Painting Style: Media, Methods, and Principles of Form*, University of Washington Press, Seattle, 1982.

[Stur89] Sturman, D.J., 'Motion Picture Cameras and Computer Graphics', *MIT Media Laboratory Memo, Computer Graphics and Animation Group*, June 1989.

[Turk90] Turk, Matthew, Pentland, Alex, 'Eigenfaces for Recognition', *Journal of Cognitive Neuroscience*, September 1990.

[Ueda91] Ueda, H., Miyatake, Y., Yoshizawa, S., 'IMPACT: An Interactive Natural-Motion-Picture Dedicated Multimedia Authoring System', *CHI Conference Proceedings*, 1991.

[Ullm79] Ullman, S., *The Interpretation of Visual Motion*, MIT Press, Cambridge, MA, 1979.

[Watl87] Watlington, John A., 'Synthetic Movies', Master's Thesis, Massachusetts Institute of Technology, September 1989.

[Xue90] Xue, K., Winans, A. Zhang, C., 'Spatial domain algorithm for video-to-printing image resolution conversion', *Proceedings of SPIE Conference on Image Communications and Workstations*, Santa Clara, CA, February 1990.

Acknowledgments

I'd like to thank some people for making this thesis possible and fun:

Walter Bender, my advisor, for a multitude of good ideas, 2 years worth of patience and the Shadow Masks, our Kentucky Fry League softball team.

John Wang and Ted Adelson of the Vision and Modeling Group for the Bergen algorithm and more good ideas.

My family for unconditional support which came in many media forms - moral, monetary and various types of edible products.

Barbara Woloch, for the well-timed phone call which always delivered academic empathy and humor from hundreds of miles away.

Janet Cahn, for the best.plan files in the world and an incendiary wit (sometimes directed at me) which kept me laughing all the way through this thesis.

Thomas Aguierre Smith, for lending me his good mind when mine became stale and for rooting around Boston with me in search of adventure.

Paddy McLean, for sharing an office, a home, and many a hottub party with me, all the while putting up with my attempts to affect a British accent.

Judith Donath, Stephan Fitch and Shahrok Yadegari for the multitude of dinner parties, drinks and conversation.

Foof Joe Stampleman, for being my thesis comrade in arms.

Henry Holtzman, for being the font of all knowledge.

Mike Halle and Shawn Becker for answering many a pesky question, usually after 2am.

Nathan Abramson, my current officemate who put up with my hogging the office Dec station all in the name of thesis writing.

My memorable friends who left here before I did: Michelle Fineblum, Mok and Hakon Wium Lie.