Dynamic Video Narratives

Carlos D. Correa
University of California, Davis

Kwan-Liu Ma
University of California, Davis

Figure 1: *A dynamic video narrative of a dance sequence from the movie* Happy Go Lovely*(1951) is constructed as a composition of five mosaics. Each mosaic comprises several frames stitched together to highlight different stages of a performance.*

Abstract

This paper presents a system for generating dynamic narratives from videos. These narratives are characterized for being compact, coherent and interactive, as inspired by principles of sequential art. Narratives depict the motion of one or several actors over time. Creating compact narratives is challenging as it is desired to combine the video frames in a way that reuses redundant backgrounds and depicts the stages of a motion. In addition, previous approaches focus on the generation of static summaries and can afford expensive image composition techniques. A dynamic narrative, on the other hand, must be played and skimmed in real-time, which imposes certain cost limitations in the video processing. In this paper, we define a novel process to compose foreground and background regions of video frames in a single interactive image using a series of spatio-temporal masks. These masks are created to improve the output of automatic video processing techniques such as image stitching and foreground segmentation. Unlike hand-drawn narratives, often limited to static representations, the proposed system allows users to explore the narrative dynamically and produce different representations of motion. We have built an authoring system that incorporates these methods and demonstrated successful results on a number of video clips. The authoring system can be used to create interactive posters of video clips, browse video in a compact manner or highlight a motion sequence in a movie.

Keywords: Video exploration, Interactive Editing, Image Compositing, Motion Extraction, Graph-cut Optimization

1 Introduction

The purpose of a visual timeline or a narrative is to display the passage of time by means of a sequence of images. These narratives are intrinsically linear and they are meant to tell a story. Schmandt-Besserat, in her book *When Writing Met Art* [Schmandt-Besserat 2007], argues that the invention of writing coincided with the adoption of linear art compositions to tell a story. The structure borrowed from writing made these compositions linear and direction became time. The linear flow of images, continuity and selective repetition of characters are some of the principles that survived through the middle ages, as seen in the Bayeux tapestry (Fig.2(b)), to modern days in the form of comics [McCloud 1994]. Scientific painting and illustrations have also borrowed these design elements to present unique compositions of extreme scale time lines, such as the evolution of life and geologic time. An example is *The Age of Reptiles* Mural, by Rudolph Zallinger, which depicts the evolution of reptiles from the Devonian period to the age of dinosaurs (Fig.2(a)). Despite the changes in perspective and scale, the mural gives the impression of a single coherent scene.

In today's era of data explosion, videos and animations are becoming ubiquitous and the ability to display long video sequences in a single narrative becomes useful. As static representations, these narratives summarize sports events, help elucidate the plot of a short movie and contextualize the evolution of a location captured by a video camera. But as dynamic representations, these timelines also help understand individual actions within the appropriate context. The essential characteristics of static timelines have been adopted as the de facto standard for displaying thumbnails of video clips in editing software such as iMovie [Apple Corporation 2009]. Because they are based on individual frames, they lack the compactness and coherence that are characteristic of hand-drawn illustrations. Recent image and video collages [Rother et al. 2006; Mei et al. 2009] aid to compactness, but do not convey the flow of time. Static representations of motion [Cutting 2002; Assa et al. 2005; Goldman et al. 2006] summarize a short action, but do not provide the means to explore the sequence dynamically. Other dynamic compositions, such as panoramic video textures and photomontages [Agarwala et al. 2004; Agarwala et al. 2005] are limited to moving backgrounds, where there is no need to track individual actions. This paper presents an interactive system for creating compact representations of long video sequences in order to produce a *dynamic narrative*. In this sense, a video narrative is a summarization of a long video sequence generated as a composition of individual frames in such a way that it indicates motion and flow of time.

We aim to generate *compact*, *coherent* and *interactive* video narratives. The first two principles, compactness and coherence, have been selected based on a careful examination of the principles of

(a) Age of reptiles mural

(b) Portion of the Bayeux Tapestry

Figure 2: *Examples of visual narratives. (a) Age of reptiles mural, as an example of a linear narrative (The Age of Reptiles, a mural by Rudolph F. Zallinger. Copyright 1966, 1975, 1985, 1989, Peabody Museum of Natural History, Yale University, New Haven, Connecticut, USA. All rights reserved. Reproduced with permission.). A coherent background gives the illusion of gradual change. Despite the differences in scale and perspective, the scene appears coherent. (b) Seamless composition of sea lines and characters gives the illusion of time flow.*

sequential art and visual narratives, as suggested by ancient and contemporary art forms [Anderson 1961; Eisner 1985; McCloud 1994]. We adopt the idea of seamless transitions to convey continuity. Two different scenes can be composited together by exploiting natural edges in the images that serve as boundaries. In the Reptiles mural, for example (Fig.2(a)), trees provide natural boundaries between geologic eras. In other cases, where the scene backgrounds are similar, a seamless transition makes them appear continuous. The interactive requirement is a new component that arises with the possibilities that interactive media offer. Unlike traditional narratives, we are not limited by a static representation.

Our system allows users to construct a narrative by composing dynamic mosaics and combining them in a linear manner. A dynamic mosaic is a hybrid between a video panorama and a video storyboard. With a set of spatio-temporal masks, our system selects portions of video frames corresponding to different moving objects and places them within a single panorama. By modifying these masks in real-time, the user can compose narratives that convey motion and flow, and perform in-place playback of the video.

2 Related Work

Creating visual summaries of video sequences has been extensively surveyed by Li et al. [2001]. Following their taxonomy, we can identify two lines of research, often interwoven, one dedicated to the decomposition of a video and extraction of salient shots and another dedicated to the assembly and representation of the video summary. This work is concerned with the latter. The most common approach to represent video is through the use of individual frames, arranged in a meaningful manner. Several layouts have been proposed, such as structure-depicting icons [Ueda et al. 1993], video posters [Yeung and Yeo 1997], comic-book presentations [Boreczky et al. 2000], stained-glass visualizations [Chiu et al. 2004] and the ever-ubiquitous thumbnail sequence in software such as iMovie [Apple Corporation 2009]. Simakov et al. [2008] address the problem of video summarization as retargeting, where a video is resized into a compact summary without image cropping or scaling. Recent approaches attempt at constructing a more compact summary using collages. Inspired by image collages such as Digital Tapestry [Rother et al. 2005] and AutoCollage [Rother et al. 2006], video collages have the additional requirement of maintaining temporal structure. Free-Shaped Video Collages seamlessly assemble multiple frames in a variety of shapes without disrupting their sequence in time [Yang et al. 2008; Mei et al. 2009]. These approaches are intended to represent the story line in the video, but they do not satisfy certain desired properties of visual narratives, such as coherence and continuity. Because these approaches handle individual frames, they seldom convey action and flow.

The representation of motion in static images is a complex task with roots in art and science [Cutting 2002]. Cutting describes five ways in which motion can be represented, including broken symmetry, stroboscopic images, motion blur, forward lean and action lines. Common in comic books, forward lean and actions lines are simple mechanisms to make a static image appear in motion [McCloud 1994], and have inspired techniques for video [Kim and Essa 2005]. Action lines often do not convey a wide range of motion. When applied to video, stroboscopic images seem to be more effective, and can be obtained optically in the form of long exposure shots. As an alternative, it is possible to obtain computational time-lapse images [Bennett and McMillan 2007] by assembling the frames in a manner that simulates a virtual camera shutter. For general video, this issue is more complex, as the camera may move in addition to the moving objects. To obtain such motion representations, a video sequence is assembled in a single panorama using a motion estimation technique, such as optical flow [Shum and Szeliski 1998] or feature-based stitching [Brown and Lowe 2003]. Panoramas, often depicting a static background, have been used for cel animation [Wood et al. 1997] and to photograph long scenes [Agarwala et al. 2006]. Useful for browsing, video panoramas such as PanoramaExcerpts [Taniguchi et al. 1997] seldom tell a story. More effective panoramas can be obtained with foreground extraction, as shown in Digital Photomontage [Agarwala et al. 2004] . Video panoramas also convey motion with action lines [Irani and Anandan 1998], storyboarding metaphors [Goldman et al. 2006], and action pose estimation [Assa et al. 2005]. These video synopses are inherently static. Panoramic video textures [Agarwala et al. 2005], Dynamosaicing [Rav-Acha et al. 2007], dynamic stills [Caspi et al. 2006] and non-chronological video synopses [Pritch et al. 2008] combine the compactness of panoramas with dynamic browsing. Dynamosaics create dynamic panoramas using 4D min-cuts, but are oblivious to the composition of objects and background. This makes the approach applicable to many types of video clips, but are not intended to convey action or tell a story [Rav-Acha et al. 2007]. Aner et al. [2002] use mosaics for video browsing, while Forlines presents a system for skimming through video frames on the recovered background, similar to our skimming technique [Forlines 2008]. However, these assume single mosaics. In our work, we construct narratives that combine several mosaics in a single composition. Instead of single frames, our visual narrative is an assembly of mosaics, or a *summary of summaries*. We believe that the result is a considerably more compact representation of video that manages to show motion and time flow. Concurrently, Barnes et al. [2010] also draw inspiration from tapestries and linear narratives to summarize a long video sequence. Using a continuous temporal zoom, users can quickly browse the video at multiple scales. Similar to our narratives, seamless blending between frames results in a compact representation that is both aesthetically pleasing and space-efficient. In our work, we use narratives to convey an action of shorter video sequences and tell a story.

Although we focus on the interactive assembly of narratives, video summarization techniques are relevant. Li et al. [2001] survey the

most important methods, which extract different properties of individual video frames, shots or segments, such as saliency [Teodosio and Bender 1993], visual attention [Ma et al. 2002] and motion [Sawhney and Ayer 1996]. In our work, we extract metrics from the individual frames and panoramic scenes to convey desired properties of narratives as identified by studies of sequential art [Anderson 1961; Eisner 1985; McCloud 1994; Tufte 1990]. Unlike previous research, which derive information theoretic descriptors of individual frames, we use motion. By ensuring that motion is uninterrupted from scene to scene, we attain the illusion of flow, one of the key properties of visual narratives.

3 Design Principles of Dynamic Narratives

At the core of visual narratives is the fact that art and writing are interwoven [Eisner 1985]. Hence, many principles are shared with writing styles. Here, we extract some of the properties that help us discern visual narratives from other types of compositions such as collages and thumbnails.

Continuity. Continuity refers to the re-use of backgrounds to convey the idea of time. The *Bayeux tapestry*, an 11^{th} century tapestry detailing the Norman conquest of England (Fig.2b), was created to convey, not a single instance of time, but rather two stages of a journey: a sea voyage and their landing on the coast. The sea line remains continuous throughout the tapestry to remind us of the continuity of time. *The Age of Reptiles* mural (Fig. 2a) comprises millions of years in a single panorama to convey the idea of gradual change. Notice the use of trees to naturally break the scene into different periods. Comic books often use panels to enforce a change and break continuity. This seems more an artistic choice than a necessary condition. In this paper, we focus on the creation of continuous narratives.

Linear flow. Schmandt notes that visual narratives seem to have appear contemporary to writing [Schmandt-Besserat 2007]. This explains why the flow of time often follows a conventional reading direction. Although videos and films often depict time as a complex network, and moving back and forth in time is a common narrative device, it is not our intent to depict the chronological time within a film, but rather the linear flow of the video.

Indication of motion. Unlike static or moving panoramas, narratives tell a story, which are collections of interconnected actions. In a video, actions can be understood as motion. For the sake of compactness, showing every single moment of an action is not possible. Instead, narratives use several strategies to convey motion in a static manner, such as broken symmetry, stroboscopic images, affine shear, blur and action lines [Cutting 2002]. Some of them, such as broken symmetry, are inherently static, but stroboscopic images and action lines, on the other hand, can also be very powerful when one allows them to become dynamic. In our work, we apply this idea for interactive playback of video narratives.

Based on these principles, we focus on narratives as seamless dynamic compositions. We explore the use of natural boundaries to provide sharper transitions where possible and smooth blending where it is not. Aesthetically, video narratives are more compact and coherent than a simple layout of mosaics.

4 Technical Approach

A dynamic narrative can be defined as a linear collection of mosaics, blended together to ensure seamless transitions. A mosaic, in turn, is a panoramic summary of a short video sequence occuring over a common background. Therefore, the generation of a narrative can be decomposed into three parts: (1) A pre-processing of the video to stabilize individual frames. Frames that share a background are grouped into the same mosaic. (2) Mosaic generation using spatio-temporal masks, and (3) Narrative composition using graph cut blending.

4.1 Pre-processing

As a first step, we obtain frames that have been stabilized for motion. We accomplished this following the approach by Lowe et al. [2003] for matching frames and computing image panoramas. Simply stitching frames at the seams does not produce a compelling representation of the video, due to the presence of moving foreground objects. We use their approach for finding matches between frames and stabilizing them against the common background. This operation can be expensive and it is therefore computed a priori. For each frame we compute scale invariant features (SIFT) and use a translational and zoom model to find matches between consecutive frames. Two consecutives frames are matched together if the extracted SIFT features can be modeled as a translation and a uniform zoom. The outcome of this step is a set of *registered* images, each having the same size as the corresponding mosaic.

4.2 Mosaic Generation using Spatio-Temporal Masks

A mosaic M can be defined as the composition of a background B and selected parts of individual *registered* frames I_k into a single image. The individual parts should correspond, in principle, to moving foreground objects. To find these moving parts, we followed the approach by [Kaewtrakulpong and Bowden 2001], where a Gaussian Mixture Model is used to tag pixels as either foreground or background. A similar approach is followed by Pal and Jojic [2005] to extract moving objects from security video. The result is a series of Gaussian blobs and a set of foreground masks, binary images which indicate if a pixel is considered as foreground when it has value 1 or background, otherwise. We construct the mosaic from these blobs. Our mosaics are dynamic, therefore this generation stage is a time-dependent process.

Let us consider the output of the foreground estimation stage a set of blobs $Blobs(k)$ for a given registered frame I_k and a foreground mask F_k. A blob j in a frame k can be characterized by a spatial 2D mean μ_{kj} and a spatial standard deviation σ_{kj}.

The process of creating a mosaic can be defined as the application of a spatio-temporal mask to the image I_k. A spatio-temporal mask is a grayscale image representing the alpha or opacity of a region in an image, which depends on both space and time. Here, we define three main masks for every frame I_k.

Blob Mask (G_{kj}). This mask defines the extents around a detected blob $i \in Blobs(k)$, defined as a Gaussian blur with mean μ_{kj} and standard deviation σ_{kj}. This mask blends the parts of the frame that corresponds to blobs with the background.

Rim Mask (R_k). This mask assumes that foreground objects tend to be centered in a frame and therefore defines the mask as a smooth rim of width r. This rim mask is 0 for pixels in the edge of the frame, 1 for pixels at a distance r from the edge of the frame, and smoothly interpolated in between.

Temporal Mask (T_k). This mask specifies the temporal behavior of the composited mosaic. A static mosaic has a constant temporal mask of 1. In other cases, however, it is desired to make foreground objects more transparent when they correspond to older frames, and make clear the direction of motion. The user can perform interactive playback on the mosaic by interacting with the parameters of this mask. Setting the mask transparent for certain frames is useful when only key poses need to be shown, while semi-transparent

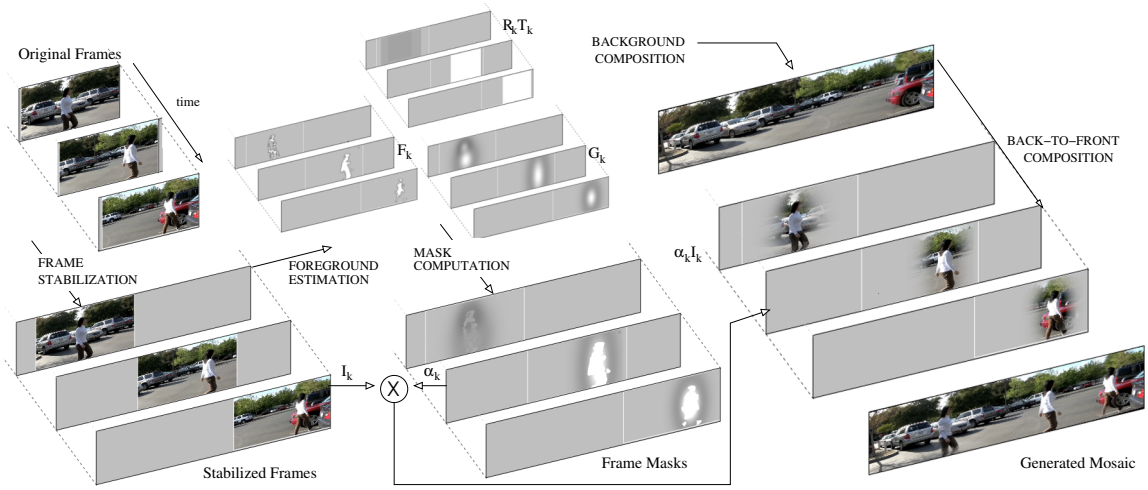


Figure 3: *Using spatio-temporal masks to generate mosaics. (1) Frame stabilization ensures that corresponding pixels from consecutive frames are in close proximity (2) We run a foreground estimation algorithm to tag pixels as foreground or background, the result is the foreground image F_k and a series of blobs per frame. We apply a temporal mask and a Gaussian mask per blob to obtain the frame mask α_k. (4) The mosaic is then the composition of a background and the opacity modulated frames $\alpha_k I_k$.*

masks help us simulate motion blur. Please refer to the accompanying video for examples. Details are given in Section 4.3.

Figure 3 shows the process of generating a mosaic. After stabilizing the frames, we show the foreground masks F_k obtained for a number of frames. These masks correspond to Gaussian blobs, shown as images G_k. Also, note the temporal mask applied to the rim mask as images $R_k T_k$. In this case, there is a falloff that makes older frames more transparent and highlight only the newer ones.

Now let $\mathbf{p}$ define a pixel position in the mosaic M consisting of N registered frames. Since the frames are registered, this pixel position is also defined in frames I_k. For a time value $t \in [1,N]$, we can find the resulting color $M_N(\mathbf{p})$ of the mosaic as a back-to-front blending of the background and the registered images after applying the resulting spatio-temporal mask α_k. This can be expressed as the recursive application of linear blending of the registered frames:

$$M_0(\mathbf{p},t) = B_N(\mathbf{p}) \quad (1)$$

$$M_k(\mathbf{p},t) = \alpha_k(\mathbf{p},t)I_k(\mathbf{p}) + (1-\alpha_k(\mathbf{p},t))M_{k-1}(\mathbf{p},t) \quad (2)$$

where B_N is the extracted background. The suffix N refers to the fact that this generated background is also the product of the blending of the N frames in the mosaic, although with different masks. α_k is the spatio-temporal mask for a given frame k, computed as the product of the rim and temporal masks with the overall foreground mask, defined as the sum of the Gaussian blobs in the image and the foreground image. This sum is useful as the foreground estimation may not always produce a precise mask. When a foreground object remains still for a while, some pixels are often mistaken as background pixels. In this case, we ensure that most of the object will be represented in the mosaic by adding the Gaussian mask:

$$\alpha_k(\mathbf{p},t) = R_k(\mathbf{p})T_k(\mathbf{p},t-k)\left(F_k + \sum_{j=1}^{Blobs(k)} G_{kj}(\mathbf{p})\right) \quad (3)$$

The background B_N can be extracted in numerous ways. Assa et al. [2005] use the temporal median. More sophisticated methods have been proposed [Granados et al. 2008]. Motion segmentation approaches already provide an estimate of the background, but the result may contain artifacts due to mis-registration. Instead, we follow a similar approach to the one above, where the background is the composition of different regions of each frame. Instead of choosing masks corresponding to foreground objects, we choose the inverse mask. Fig. 4 compares our approach for computing the background with the extracted background using the method in [Kaewtrakulpong and Bowden 2001] and the temporal average. On top, we see that the foreground extraction method is not intended to compute a perfect background, but rather track the moving objects. We see some artifacts towards the left part of the mosaic. The temporal average, on the other hand, is blurry. Our approach produces a single background without the blur, by using stabilized portions of the frames that are not considered foreground (after applying the masks). We compute the background using a back-to-front composition scheme. For each frame I_k:

$$B_0(\mathbf{p}) = \hat{B}(\mathbf{p}) \quad (4)$$

$$B_k(\mathbf{p}) = \beta_k(\mathbf{p})I_k(\mathbf{p}) + (1-\beta_k(\mathbf{p}))B_{k-1}(\mathbf{p}) \quad (5)$$

where the background mask is:

$$\beta_k(\mathbf{p}) = R_k(\mathbf{p})\left(1 - \sum_{j=1}^{Blobs(k)} G_{kj}(\mathbf{p})\right) \quad (6)$$

and $\hat{B}$ is an initial guess of the background, which can be the one obtained from the foreground estimation method, the temporal median or the temporal mean. In our case, the temporal mean gives us acceptable results, since the blurry portions are likely to be masked by β_k. Note that the background, in this case, does not depend on time. This is intended to provide a single background for the mosaic and avoid artifacts when exploring the video dynamically.

4.3 Dynamic Exploration of the Mosaic

Unlike previous attempts to produce visual summaries of videos, spatio-temporal masks allow us to produce dynamic explorations of the mosaic. This is achieved by modifying the parameter t, which controls the frame currently displayed in the video, and the temporal mask T_k. Since the mosaic is defined in terms of t, the result varies depending on the shape of function T_k. For example, one can enable in-place playback of the video while leaving a semi-transparent trail of the previous frames as a stroboscopic image,

Figure 4: *Comparison of background generation methods. Left: A few video frames. Right, from top: (1) The output from foreground estimation process may contain errors due to overlapping intensities, (2) The temporal average is blurry, (3) Our results.*

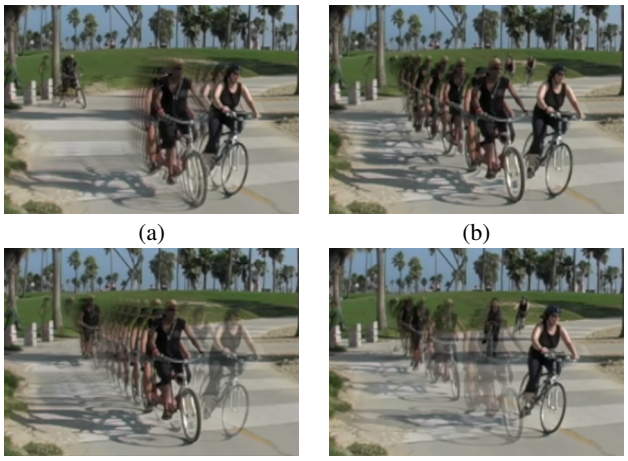


Figure 5: *Different effects can be obtained by manipulating the temporal mask. The temporal mask indicates an opacity for each time t. (a) A mask is used to simulate motion blur and make explicit the difference in velocity of the two moving actors. (b) A mask to highlight the trajectory of motion (c) Using a temporal falloff helps us identify who moved first (the more transparent bicycle appears earlier) (d). We can invert the falloff to change the temporal relationship. Now, the other bicycle appears to have moved first.*

using a smooth falloff function T_k. We use an exponential falloff:

$$T_k(\mathbf{p}, t-k) = e^{-(t-k)^2/2\sigma^2} \tag{7}$$

where the parameter σ controls the falloff. In this case, the falloff produces stroboscopic images of the frames preceding and succeeding the frame at time t. By modifying the parameter t, the user can produce a stroboscopic playback of the video. In many cases, excessive use of this effect introduces clutter and the motion is no longer visible. Instead, one can introduce temporal filtering to sample sparse frames, as shown in Figure 5.

4.4 Alpha Matting

In our approach, we attempt to maximize the probability of segmenting the foreground objects by considering a Gaussian blur that covers most of the foreground pixels, but that also includes some background pixels. One issue with the Gaussian blur added to the foreground mask is the bleeding of background pixels into other foreground replicas when creating stroboscopic images. This is seen in Figure 6-left. Here, the second replica of the walking person (from left to right) is blended with the background. We can see a greenish halo (from the grass) emanating from the first replica. This can be alleviated when the background does not move and the stabilization of the frames is accurate enough. We use an alpha matting approach, where we compute a new mask $\hat{\alpha}_k$ by solving the matting equation for every pixel $\mathbf{p}$:

No matting Matting

Figure 6: *Alpha matting can be used to improve the blending of foreground objects. Left: no matting results in bleeding of background pixels to the other replicas. Right: matting results in crisper foreground objects (Video courtesy of Dan B Goldman).*

$$I_k = \hat{\alpha}_k(\alpha_k I_k) + (1 - \hat{\alpha}_k) B_N \tag{8}$$

This equation states that the new alpha mask should statisfy the matting equation for the foreground image $\alpha_k I_k$, where α_k is the one obtained using the spatio-temporal masks and the background image B_N. The result can be seen in Figure 6-right, where individual foreground replicas can be shown at full opacity without introducing background halos.

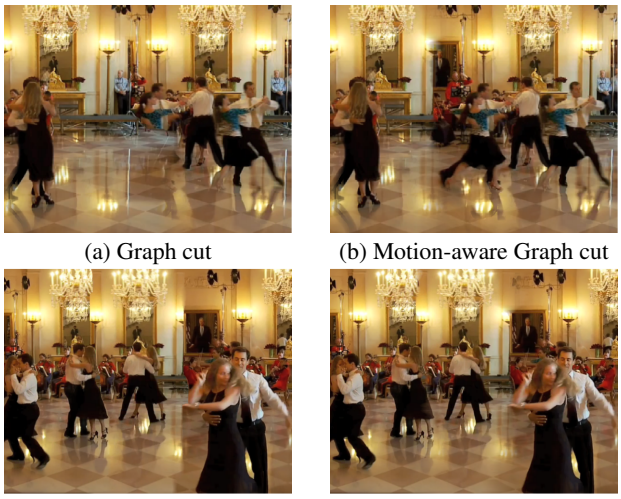
4.5 Narrative Composition

A full narrative of a video is a composition of mosaics in a linear manner. One can try to put all mosaics in sequence to signify the separation of disparate scenes. However, representing this narrative as such a sequence does not produce compact images. In our approach, we allow overlap to create compact and seamless transitions between mosaics. To blend the two mosaics in the intersection region, we use a modified version of the graph-cut seams introduced by [Boykov and Kolmogorov 2004]. This method considers the pixels in the overlap area as a graph, where edges are defined between the horizonal and vertical immediate neighbors of each pixel. The optimal seam (or minimum cut) is one that maximizes flow [Kwatra et al. 2003], which depends on the cost of each edge.

The traditional application of graph cuts does not ensure the continuity of dynamic narratives, since an overlapping mosaic may obscure an important action of the occluded mosaic. On the other hand, performing the graph cut optimization on-the-fly may be costly. Instead, we define a cost function that blends two mosaics M_{src} and M_{dst} and adds a penalty, in terms of the motion masks A_{src} and A_{dst}, that increases the cost in regions with motion. The motion mask image is a grayscale image that composes the aggregate mask of all the frames in a mosaic. That is, for a given mosaic M, the corresponding motion mask A is computed using back-to-front compositing as:

$$\begin{aligned} A_1(\mathbf{p}) &= \alpha_1(\mathbf{p}, 1) \\ A_k(\mathbf{p}) &= \alpha_k(\mathbf{p}, k) + (1 - \alpha_k(\mathbf{p}, k)) A_{k-1}(\mathbf{p}, k) \end{aligned}$$

for $k \in [1, N]$, where N is the number of frames in the mosaic. Notice how we apply the temporal parameter t of α_k as the time k. This is done with the intention of obtaining the highest mask for that



(a) Graph cut (b) Motion-aware Graph cut

(c) Unfeathered graph cut (d) Feathered graph cut

Figure 7: *Motion-based graph cuts. (a) Traditional graph cut may be placed through moving characters. (b) With a motion term, we prevent the seam from being placed along the moving characters. As a faster alternative to gradient-domain blending, we use feathering around the graph cut proportional to the local gradient. (c) with no feathered cuts, we get visible seams, especially when the lighting varies. (d) feathered graph cuts produce acceptable results with little cost and preserves sharp edges (wall paintings). (Video courtesy of the White House, public domain)*

frame and avoiding missing important actions due to the temporal falloff. One can alternatively define A as a temporal motion mask, which changes as we change time. Therefore, graph cuts need to be computed for each frame. This approach, however, might produce popping artifacts. The result of incorporating motion in the graph cut blending is shown in Figure 7(a-b).

The cost function between two neighboring pixels $\mathbf{p}$ and $\mathbf{q}$ is therefore defined as

$$\omega V(\mathbf{p},\mathbf{q},M_{src},M_{dst}) + (1-\omega)(A_{src}+A_{dst}) \quad (9)$$

for two overlapping mosaics M_{src} and M_{dst} and their respective alpha masks A_{src} and A_{dst}, and ω a weighting factor to give more weight to the image features or the motion. When $\omega = 1$, the result is that of traditional graph cut blending, i.e., ignoring the motion. The term V is formulated as in AutoCollage [Rother et al. 2006], known to work better for disparate overlapping images:

$$V(\mathbf{p},\mathbf{q},M_{src},M_{dst}) = min(\frac{||M_{src}(\mathbf{p})-M_{dst}(\mathbf{p})||}{\varepsilon+||M_{src}(\mathbf{p})-M_{src}(\mathbf{q})||}, \frac{||M_{src}(\mathbf{q})-M_{dst}(\mathbf{q})||}{\varepsilon+||M_{dst}(\mathbf{p})-M_{dst}(\mathbf{q})||})$$

where ε is a small number to prevent division by zero. The numerators on this equation corresponds to pixel differences between the two mosaics at any given point for two neighboring pixels $\mathbf{p}$ and $\mathbf{q}$. The denominators define the image gradient in the direction of the neighbor. Therefore, this equation is minimum when either the pixel differences are small, representing a seamless transition, or when the gradient is high, representing a hard edge where the cut can be placed.

Feathered Graph Cuts. In image stitching and panorama generation, seams obtained with graph cuts are usually blended using gradient-domain approaches [Rother et al. 2006]. The same can be used in our approach. As a fast alternative, our authoring system uses feathered graph cuts, which also consider the gradient of the image to define a smooth interpolation between two

Figure 8: *Our authoring system consists of three windows. The top window shows the scenes detected by the stabilization process. On the top right corner we provide the original video clip. The main window, the narrative view, contains all the clips selected by the user. In this example, the user has selected two clips, which the user can explore dynamically.*

overlapping mosaics. This is especially necessary for disparate mosaics or mosaics with slightly different backgrounds (possibly due to light variation). The effects of not using blending is seen in Fig 7(c). To alleviate this, we define an exponential falloff of $e^{-\max(||\nabla M_{src}||,||\nabla M_{dst}||))/\sigma^2}$ for two mosaics M_{src} and M_{dst}, where σ^2 defines the steepness of the falloff, and ∇M_{src} and ∇M_{dst} are the gradients of the source and destination mosaics at a pixel in the seam, respectively. When the gradients are high, the falloff is small, which preserves the sharp graph cut at that point. When the gradients are low, the falloff is larger and the seam is replaced by a smooth transition. An example is shown in Fig. 7(d).

4.6 Interaction

Unlike hand-drawn narratives and image collages, our system provides interactive video playback and skimming. The user can play, pause or rewind different portions of the narrative at any time. A snapshot of the authoring system is shown in Figure 8, and consists of three windows: the thumbnail view, which hosts all the scenes detected by our system, the original video playback window (top right corner) and the narrative view, which is the main canvas where the user places the different clips. In there, we provide the following capabilities for building an effective narrative:

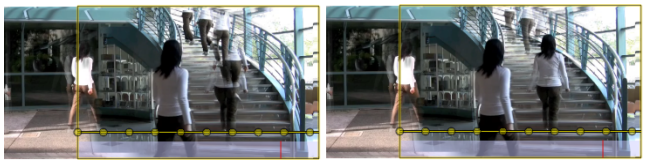
Narrative Assembly. The user creates a narrative by selecting clips from the thumbnail view and placing them on the narrative view. Dragging one clip left or right allows the user to control the size of the narrative. In the figure below, the two clips in Fig. 8 are collapsed in a shorter narrative. Notice how the ticket machine on the left provides the transitional seam between the two scenes.

Playback. Dynamic video narratives can be played back and skimmed in real-time. We provide a playback bar (bottom) that the user can slide right or left to go forward or backward, respectively, in time. Below, the user interactively rewinds the second clip.

Temporal Exploration. The user can explore the temporal aspects of a clip to highlight an action or movement. This is obtained in

our system using the temporal mask. We enhance the playback bar with yellow dots representing the distribution of foreground objects used in the narrative. In the example below, the user chooses five replicas. Each replica has a motion tail, represented as vertical line segments. Dragging the mouse over this section allows the user to change the distribution of replicas or the density of the tails. Here, the user increases the density of the tail.

Temporal Ordering. Since some clips have the moving objects that approach the camera while others go away from it, the temporal order of frames is important. As described before, this ordering can represent different temporal aspects. In this example, the user switches the order to show the correct motion up the stairs.



5 Discussion and Limitations

Our approach offers a variety of possibilities for generating effective visual narratives with little effort. For the purposes of creating narratives, the nature of the video bears some importance. In our case, our approach is more effective for videos that capture an action in both space and time. We believe a wide range of video clips fall into that category, including performance capture (such as dance), sportscasts, action sequences in movies and shots from TV shows. Professional movie clips usually have clear shots of the foreground objects and are centered. Home made movies are often shot from a single person's perspective and may not be suitable for extracting narratives. In contrast, animated shorts often have static backgrounds and color uniformity is the norm.

Examples of narratives created using our approach are shown in Figures 10-12. Figure 10 illustrates the use of narrative to show directions in a video clip that follows a person from one place to another. Figure 11 shows the ability to tell a story from a cartoon short, and the use of temporal effects to highlight speed and motion. Figure 12 shows the capabilities of video narratives to summarize general footage, in this case an equestrian show. Our system is more effective for video shots where the motion has a clear directionality in the 2D plane (panning and zooming) and the foreground objects are clearly visible. There are certain cases where our approach has limited use when composing a narrative:

Motion parallax and fast camera motion. Since we rely on frame stabilization, our system is subject to the limitations of current stabilization algorithms. At this stage, we handle panning and zooming camera motions, and our work can be extended to other more complex camera motions, including rotation. Motion parallax remains an issue, and our approach is limited by the accuracy of the foreground estimation process under such conditions. Other complicated cases are those with fast moving sequences where consecutive frames cannot be stitched together in a single mosaic. In such cases, our approach is left with a collection of disparate frames of low compactness, and the resulting narratives appear more like a collage, similar to those in [Yang et al. 2008]. Other limitations are more unique to our approach:

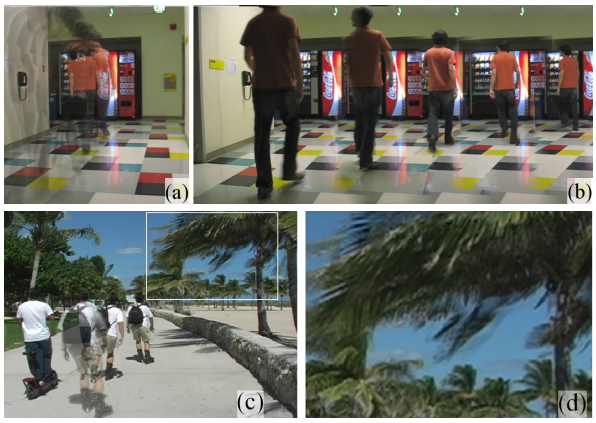


Figure 9: *Limitations. (a-b): Occluding motion may not be conveyed properly using our approach (a). As an alternative, we can split a scene with occluding motion into separate scenes and apply our narrative assembly mechanism to reduce the effects of occlusion (b). (c-d): Moving background. Here, a palm moving in the wind is handled by the system as a collection of individual moving blobs, resulting in a discontinuous motion. Compare to the two moving actors, which are properly segmented.*

Occluding motion. Since we do not modify the relationship between the foreground object and the background, overlapping motion may not help depict an action. An example is shown in Figure 9(a). As an alternative, we can split a scene into smaller subscenes, in which case our approach can exploit the narrative assembly mechanism to deal with occlusion (Figure. 9(b)).

Moving background. When the moving objects cannot be clearly segmented, such as with large moving backgrounds (e.g., waves of the ocean) or large deformable objects (e.g., the palm leaves in Figure 9(c)), the temporal masks do not align with clearly defined objects. The resulting narrative will contain blurry regions representing the parts where the background changes.

6 Conclusions and Future Work

We have successfully addressed several key challenges in producing compact and dynamic narratives from video clips. Narratives created with our approach follow certain key design principles. Coherence is ensured by stitching the background into a single panorama. We use motion estimation algorithms to recover the camera motion from a video clip and foreground estimation to extract foreground pixels. We decompose the generation of narratives as the blending of foreground and background regions in a way that depicts the flow of time. We have shown how to create a series of spatio-temporal masks that can be used to extract a crisp background without the blurriness of temporal averages and to also indicate the foreground regions that are extracted from individual frames. Through interactive manipulation of these masks, we have successfully created a playable and interactive mosaic that also lets users explore temporal effects. This enables interactivity and dynamism and overcomes the limited communicative value of static counterparts. We ensure compactness by allowing mosaics to be blended together in a seamless manner. We use a novel variation of the graph cut algorithm for building seams, which now uses the motion masks to prevent cuts along moving objects.

In the spirit of usability, our system makes a lot of automatic choices about the foreground and background. Our spatio-temporal masks are constructed to alleviate the artifacts that may appear when a static background cannot be retrieved accurately or moving objects cannot be segmented with precision. With manual inter-

Figure 10: *Video narrative of a sequence (7 minutes) depicting directions from a parking lot to a particular office room. Our approach is able to compress the video and exploit hard edges, such as the parking machine and door frames, to provide seamless transitions.*

vention, such as specifying regions as either foreground and background, our results can naturally be improved to obtain even more visually pleasing images. Although the generation of narratives that tell a story remains a craft, our system gets us closer to automatic video summarization and provides a test-bed to conduct studies about the expressive power of visual narratives and the principles of sequential art.

Acknowledgements

This research was supported in part by U.S. National Science Foundation through grants CCF 0811422, CCF 0938114, OCI 0850566, CNS 0716691, and CCF 0808896 and U.S. Department of Energy through the SciDAC program with Agreements No. DE-FC02-06ER25777. Videos courtesy of the White House video galleries, FootageFirm.com and Dan B Goldman. Thanks to archive.org for access to the public domain movies.

References

AGARWALA, A., DONTCHEVA, M., AGRAWALA, M., DRUCKER, S., COLBURN, A., CURLESS, B., SALESIN, D., AND COHEN, M. 2004. Interactive digital photomontage. *ACM Trans. Graph. 23*, 3, 294–302.

AGARWALA, A., ZHENG, K. C., PAL, C., AGRAWALA, M., COHEN, M., CURLESS, B., SALESIN, D., AND SZELISKI, R. 2005. Panoramic video textures. *ACM Trans. Graph. 24*, 3, 821–827.

AGARWALA, A., AGRAWALA, M., COHEN, M., SALESIN, D., AND SZELISKI, R. 2006. Photographing long scenes with multi-viewpoint panoramas. *ACM Trans. Graph. 25*, 3, 853–861.

ANDERSON, D. M. 1961. *Elements of Design*. Holt, Rinehart and Winston.

ANER, A., AND KENDER, J. R. 2002. Video summaries through mosaic-based shot and scene clustering. In *ECCV '02: Proceedings of the 7th European Conference on Computer Vision-Part IV*, 388–402.

APPLE CORPORATION, 2009. iMovie. http://www.apple.com/ilife/imovie .

ASSA, J., CASPI, Y., AND COHEN-OR, D. 2005. Action synopsis: pose selection and illustration. *ACM Trans. Graph. 24*, 3, 667–676.

BARNES, C., GOLDMAN, D. B., SHECHTMAN, E., AND FINKELSTEIN, A. 2010. Video tapestries with continuous temporal zoom. *ACM Transactions on Graphics 29*, 3.

BENNETT, E. P., AND MCMILLAN, L. 2007. Computational time-lapse video. In *SIGGRAPH '07: ACM SIGGRAPH 2007 papers*, 102.

BORECZKY, J., GIRGENSOHN, A., GOLOVCHINSKY, G., AND UCHIHASHI, S. 2000. An interactive comic book presentation for exploring video. In *CHI '00: Proc. SIGCHI conference on Human factors in computing systems*, 185–192.

BOYKOV, Y., AND KOLMOGOROV, V. 2004. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Trans. on Pattern Analysis and Machine Intelligence 26*, 9, 1124–1137.

BROWN, M., AND LOWE, D. G. 2003. Recognising panoramas. In *ICCV '03: Proc. Ninth IEEE International Conference on Computer Vision*, 1218.

CASPI, Y., AXELROD, A., MATSUSHITA, Y., AND GAMLIEL, A. 2006. Dynamic stills and clip trailers. *Vis. Comput. 22*, 9, 642–652.

CHIU, P., GIRGENSOHN, A., AND LIU, Q. 2004. Stained-glass visualization for highly condensed video summaries. *IEEE Conf. on Multimedia and Expo, 2004. ICME '04. 2004 3*, 2059–2062.

CUTTING, J. 2002. Representing motion in a static image: constraints and parallels in art, science, and popular culture. *Perception 31*, 1165–1193.

EISNER, W. 1985. *Comics and Sequential Art*. Poorhouse Press.

FORLINES, C. 2008. Content aware video presentation on high-resolution displays. In *AVI '08: Proceedings of the working conference on Advanced visual interfaces*, 57–64.

GOLDMAN, D. B., CURLESS, B., SALESIN, D., AND SEITZ, S. M. 2006. Schematic storyboarding for video visualization and editing. *ACM Trans. Graph. 25*, 3, 862–871.

GRANADOS, M., SEIDEL, H.-P., AND LENSCH, H. P. A. 2008. Background estimation from non-time sequence images. In *GI '08: Proc. Graphics Interface 2008*, 33–40.

IRANI, M., AND ANANDAN, P. 1998. Video indexing based on mosaic representations. *Proc. of the IEEE 86*, 5 (May), 905–921.

KAEWTRAKULPONG, P., AND BOWDEN, R. 2001. An im-

Figure 11: *Video narrative of a cartoon short (Superman cartoons (1942)). Here, we exploit our capabilities to simulate time-lapse to depict motion and speed. For example, the use of semi-transparent motion tails helps us convey the speed of Superman taking off.*

Figure 12: *Video narrative of a rodeo show. The result is a compact way of summarizing long video clips, which can help in the search of particular scenes. Unlike individual thumbnails, narratives help convey motion to quickly identify an action or a particular moment.*

proved adaptive background mixture model for realtime tracking with shadow detection. In *In Proc. 2nd European Workshop on Advanced Video Based Surveillance Systems, AVBS01*, Kluwer Academic Publishers.

KIM, B., AND ESSA, I. 2005. Video-based nonphotorealistic and expressive illustration of motion. In *CGI '05: Proc. Computer Graphics International 2005*, 32–35.

KWATRA, V., SCHÖDL, A., ESSA, I., TURK, G., AND BOBICK, A. 2003. Graphcut textures: image and video synthesis using graph cuts. *ACM Trans. Graph. 22*, 3, 277–286.

LI, Y., LI, Y., ZHANG, T., ZHANG, T., TRETTER, D., AND TRETTER, D. 2001. An overview of video abstraction techniques. Tech. rep., HP-2001-191, HP Laboratory.

MA, Y.-F., LU, L., ZHANG, H.-J., AND LI, M. 2002. A user attention model for video summarization. In *Proc. tenth ACM international conference on Multimedia*, 533–542.

MCCLOUD, S. 1994. *Understanding Comics*. Perennial Currents.

MEI, T., YANG, B., YANG, S.-Q., AND HUA, X.-S. 2009. Video collage: presenting a video sequence using a single image. *The Visual Computer 25*, 1, 39–51.

PAL, C., AND JOJIC, N. 2005. Interactive montages of sprites for indexing and summarizing security video. In *CVPR '05: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, 1192.

PRITCH, Y., RAV-ACHA, A., AND PELEG, S. 2008. Nonchronological video synopsis and indexing. *IEEE Trans. Pattern Analysis and Machine Intelligence 30*, 11, 1971–1984.

RAV-ACHA, A., PRITCH, Y., LISCHINSKI, D., AND PELEG, S. 2007. Dynamosaicing: Mosaicing of dynamic scenes. *IEEE Trans. Pattern Anal. Mach. Intell. 29*, 10, 1789–1801.

ROTHER, C., KUMAR, S., KOLMOGOROV, V., AND BLAKE, A. 2005. Digital tapestry. In *CVPR '05: Proc. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1*, 589–596.

ROTHER, C., BORDEAUX, L., HAMADI, Y., AND BLAKE, A. 2006. Autocollage. *ACM Trans. Graph. 25*, 3, 847–852.

SAWHNEY, H. S., AND AYER, S. 1996. Compact representations of videos through dominant and multiple motion estimation. *IEEE Trans. Pattern Anal. Mach. Intell. 18*, 8, 814–830.

SCHMANDT-BESSERAT, D. 2007. *When Writing Met Art: From Symbol to Story*. University of Texas Press.

SHUM, H.-Y., AND SZELISKI, R. 1998. Construction and refinement of panoramic mosaics with global and local alignment. In *ICCV '98: Proc. Sixth International Conference on Computer Vision*, 953.

SIMAKOV, D., CASPI, Y., SHECHTMAN, E., AND IRANI, M. 2008. Summarizing visual data using bidirectional similarity. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1–8.

TANIGUCHI, Y., AKUTSU, A., AND TONOMURA, Y. 1997. Panoramaexcerpts: extracting and packing panoramas for video browsing. In *Proc. fifth ACM international conference on Multimedia*, 427–436.

TEODOSIO, L., AND BENDER, W. 1993. Salient video stills: content and context preserved. In *Proc. first ACM international conference on Multimedia*, 39–46.

TUFTE, E. R. 1990. *Envisioning Information*. Graphics Press, Cheshire, Connecticut.

UEDA, H., MIYATAKE, T., SUMINO, S., AND NAGASAKA, A. 1993. Automatic structure visualization for video editing. In *CHI '93: Proc. INTERACT '93 and CHI '93 conference on Human factors in computing systems*, 137–141.

WOOD, D. N., FINKELSTEIN, A., HUGHES, J. F., THAYER, C. E., AND SALESIN, D. H. 1997. Multiperspective panoramas for cel animation. In *SIGGRAPH '97: Proc. 24th annual conference on Computer graphics and interactive techniques*, 243–250.

YANG, B., MEI, T., SUN, L., YANG, S.-Q., AND HUA, X.-S. 2008. Free-shaped video collage. In *Lecture Notes in Computer Science*, vol. 4903, 175–185.

YEUNG, M., AND YEO, B.-L. 1997. Video visualization for compact presentation and fast browsing of pictorial content. *IEEE Trans. on Circuits and Systems for Video Technology 7*, 5 (Oct), 771–785.