Video de-Abstraction
or
How to save money on your wedding video

Aya Aner-Wolf *
Department of Math and Computer Science
Weizmann Institute of Science
Rehovot, Israel
E-mail: aya@wisdom.weizmann.ac.il

Lior Wolf
School of Computer Science and Eng.
The Hebrew University
Jerusalem, Israel
E-mail: lwolf@cs.huji.ac.il

Abstract

There exist an increasing body of work dealing with video still abstraction, the extraction of representative still images from a video sequence. This work focuses in the other direction: given a video abstract and raw unedited video data, we produce an edited video.

We focus on the application of generating wedding videos. We use the existing wedding photo album as an abstract, and produce an edited wedding video from it. The photo album serves us in determining importance of raw shots, as well as style and order.

1 Introduction

Almost all wedding coupled hire a photographer who also designs the wedding album. However, an unofficial survey conducted among wedding photographers revealed that around 30% of the couples do not hire a video crew.

There are several common reasons why many couples wave the video documentation of their wedding. First, the cost of a professional edited video begins at several hundred dollars for a very basic video. Second, in many cases the person who edits the wedding video was not present at the wedding, and therefore might cut out some shots which are important to the wedded couple. This is in contrast to the wedding album where the photographer himself picks the best stills, and where stills could be relatively easily added or removed according to the couple's preferences. A third reason for not hiring a professional camera crew is that many believe that their presence would influence the guests to behave less naturally and freely.

A new trend is not to give up on video documentation altogether, but instead have friends or relatives videotape the wedding using home video cameras. In this work we propose using the raw video data taken this way, and automatically produce an edited video from it. Our method can also be used by couples who did hire a camera crew, but prefer not to use professional video editing.

The basic idea of this work is that the photo album of the wedding is an excellent abstract for a wedding video: Made to capture the highlights of the wedding, it contains all the major events which should appear in the video. Photos in a designed wedding album also appear in special styles which might infer video effects. The order of the photos is probably the desired order of corresponding video shots. From a technical view, it is also helpful that in many cases the viewing angle between the video camera and the still camera varies only slightly.

Using this abstract, we try to construct a video which captures most of its information. We include shots which were taken simultaneously with the photos from the wedding album and translate the styles

*Dr. Aner-Wolf's research is being supported by a Koret Foundation Postdoctoral Fellowship

from the album to the video.

1.1 Related work

Video abstraction has become more important recently as more advanced media services are offered with digital video being the most popular media form. Example services include digital and interactive video and large distributed digital libraries. Video abstracts serve as an efficient representation of the video, enabling the management of large volumes of such data. There are two forms of abstracts: still-image abstracts and moving-image abstracts. Still-image abstracts [4, 10] are formed by a small collection of still images, either extracted or generated from the source video. These abstracts mainly rely on key frames for visual representation. Moving-image abstracts [8, 3] consist of a collection of image sequences, as well as the corresponding audio abstract extracted from the source video. They could be either a short summary sequence or a highlight. In this work we consider a new problem: given a still-image abstract, we want to generate an edited video such that the edited video matches the abstract.

From the multiple view geometry point of view, at the time the still image was taken the video camera and the stills camera were operating as a stereo system. The literature on wide base line stereo [9, 11] deals with how to determine correspondences between two very different images of the same scene. However, these methods have not proved themselves powerful enough or efficient enough to deal with the large number of frames, features and possible matchings present in varying video shots. Moreover, in our application there are more uniform regions and less distinctive features than in the data on which such methods are usually applied to.

Mosaics automatically constructed from the video were suggested in the past to represent shots [7, 6, 12, 5]. Mosaics are not an ideal representation - they are not well defined for complex settings and general camera motion (causing distortions in the mosaics). They are also not invariant to camera viewing point and even for the choice of the reference frame. However, [2] showed that mosaics are flexible enough to serve as a good representation and indexing of video, by using a mosaic alignment and comparison method they develop. Although this alignment algorithm is not an accurate registration algorithm (the transformation between mosaics of the same settings is not given by a limited number of parameters), their comparison method was effective.

2 Overview of our system

The input to our system is raw video taken by one or several unsynchronized cameras, and an abstract of the desired video (the wedding photo album). The system then performs the stages below to produce the output video:

1. Analyzing the still images: The style and type of each still image is recognized.
2. Video enhancement: Stabilization and improving the quality of the video.
3. Segmenting the video into shots.
4. Representing each shot: we use one or several mosaics per shot.
5. Matching shots to stills: we align mosaics to stills and compare them.
6. Editing selected shots: we cut uninteresting parts of the shots and apply style to the video.
7. Generating output video: The shots are ordered and transition effects are added.

2.1 Determining the style of the stills

Given an ordered set of still images, our first task is to determine the style of each still. This will be later on translated to the style of a corresponding video shot. A more technical importance is that some styles produce degraded images compared to the natural colors high resolution photographs we might want to use. Recognizing the correct style would help us deal with this difficulty.

Some styles are determined by the type of film used for capturing the photograph. Color or black & white (B & W) is the most basic distinction and the only film type detection implemented in our current system. Not all color films are the same - many photographers use films intended for slides in order to

get unnatural appealing colors. Analyzing statistical properties of colors in the still could help us determine film type and correct for this. Sometimes a very high sensitivity film is chosen to get a graining effect in the stills. This can be detected by using texture properties of the stills, for example.

The borders of the stills might differ in style, as shown in Figure 1 which summarizes different style variations.

Figure 1: Example of stills from the wedding album. (a) Regular colors with black and border. (b) Grained colors with vanishing borders. (c) Black and white with vanishing border. (d) Grained brown and white with vanishing border.

2.2 Video enhancement

The input video in our application is often taken by an amateur using a home video camera. In our current system we first stabilize the video to reduce jittering. This is done by a simple pyramid based computation of translation in the image planes, and its elimination in cases where it does not generate a smooth motion over time (i.e. when we are convinced that the camera is not supposed to move).

In the future we intend to add a uniform contrast enhancement. Since we later on construct mosaics from the video, applying super resolution is natural in our application and will also be added.

2.3 Segmenting the video into shots

For a video taken using a digital camera, segmenting the video into shots is done simply by considering the time stamp. For videos which were provided by an analog source, an automatic temporal segmentation of the video is required.

In this work we use the method suggested in [1] for detecting shot transitions, and due to lack of space, refer the reader to [1] for more details. Even though it did classify some strong camera motions as shot transitions, these false detections proved beneficial later. They aided us both in the mosaic construction process and in the final editing stage, where shots with large motion are undesirable.

2.4 Representing shots by mosaics

There are some advantages for using mosaics over key-frames in order to represent shots. The mosaic image captures more of the background of the shot, with optionally removing the moving foreground objects [5], resulting in a clear visual representation of the physical scene.

Since most of the raw shots in our input video are long and include severe camera motion, several mosaics are constructed for each shot. We determine the shot segments for which these mosaics are constructed by measuring the type and amount of camera motion. This is done by examining the registration transformations computed between consecutive frames in the shot, during the mosaic construction process. By analyzing these transformations, we classify shot segments as panning, zoomed-in, zoomed-out, or stationary.

2.5 Matching shots to stills

Our method for matching shots to stills is based on the method presented in [2] for matching between physical locations of shots. To demonstrate their ability they used broadcast quality situation comedies and sport events. Here we need to handle images of inferior quality, taken in an sub-optimal non-uniform illumination. Another problem we face here

is possible differences in modalities and quality between the video sequence and the still image data.

Given a mosaic and a still image, we wish to align them and determine a similarity measure between them. The matching is based on aligning regions in the images by comparing their color histograms. Due to lack of space, the reader is referred to [2] for further information.

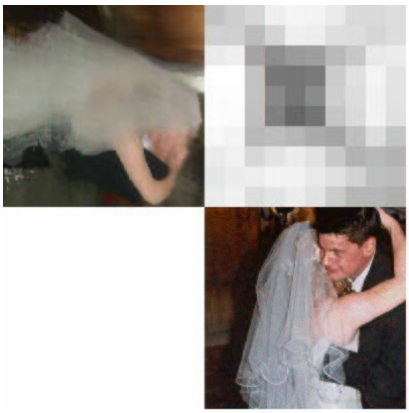

Figure 2: Example of matching a shot's mosaic to a still taken at about the same time. Only the cropped aligned regions are shown here, the cropped mosaic region (on left) was also rotate to illustrate the match. The grey level blocks represent match values between vertical strips from both images, each block is vertically aligned with a strip from the bottom still and horizontally aligned with a strip in the left mosaic.

We do not currently support matchings between B&W stills and mosaics. However, since B&W stills usually appear in the same wedding album page as the color stills taken at approximately the same time, we match the B&W still to the shot which was matched with its neighboring color still.

2.6 Editing selected shots

After we have selected the corresponding shot for each still, we still have to cut out uninteresting parts of the shot. We currently measure the motion within each shot, in the same manner as described in section 2.4. We base this on the assumption that shot segments with large pan, for example, indicate a less interesting event than the following still shot segments. For that reason, we also prefer the still shot segments which follow a zoom-in segment. More sophisticated methods based on sound track analysis and on flow analysis might be added later.

We also apply style modification for some of the shots. A natural modification for a shot associated with a B&W photograph would be to transform the shot's frames into B&W. Likewise, most of the image styles above have a natural video expansion. However, we are not limited by this mapping when choosing the transformation of style from stills to video.

2.7 Generating output video

At the final stage we collect all the edited shots and paste them together. Currently we just select a single shot for each photograph in the wedding album, but we may produce shorter summary videos as well. We can also generate other media types such as an interactive wedding album, where shots are ordered spatially according to the original wedding album.

When pasting the shots together, shot transition effects are added. We base these effects on the style of the shots. For example, B&W images are often associated with more romantic scenes, and we add softer wave like transitions between the associated shots. Currently, these are shots which were directly matched with the color stills that appear in the same wedding album page as the B & W stills.

3 Results

We have analyzed the first 60 minutes of an unedited wedding video sequence. We used the wedding photo album made for that wedding and scanned the first 53 color stills from it. Hand-segmentation of the video into raw shots resulted in 84 very long shots. The automatic segmentation yielded 198 shots, since segments with significant camera motion were split.

An example of the matching results for three pairs of stills and mosaics is shown in Figure 3. The images on the right are the stills and the images on the left are mosaics which were matched to them

using the mosaic alignment technique described in section 2.5. The upper two pairs show successful matches, whereas for the bottom pair a still taken at an earlier time was matched with a mosaic of a shot taken at a later time.

In order to correct such false matches, we apply a temporal constraint. The stills in the wedding album are usually ordered according to the time when they were taken. We therefore use a sliding window over time, both for the mosaics and the stills. This window limits a consecutive group of stills to match a consecutive group of mosaics.

A demo video in Quicktime format is available at *http://www.cs.huji.ac.il/~lwolf/demos/weddingvideo.mov* . A sample page from the wedding photo album is shown together with a segment of the edited video corresponding to that page.

Figure 3: Examples of matching results between mosaics and stills. On the right are the stills and on their left are their matching mosaics. (a),(b) Successful matches between stills and shots taken at about the same time. The still and its corresponding shot in (b) also appear in the attached video. (c) An unsuccessful match: a still image from an earlier time was matched with a mosaic of a shot from a later time. This was corrected by forcing the window time-constraint.

References

[1] A. Aner and J. R. Kender. A unified memory-based approach to cut, dissolve, key frame and scene analysis. In *Proceedings ICIP*, 2001.

[2] A. Aner and J. R. Kender. Video summaries through mosaic-based shot and scene clustering. In *Proceedings ECCV*. Springer-Verlag, 2002.

[3] A. Hanjalic and H. J. Zhang. An integrated scheme for automated video abstration based on unsupervised cluster-validity analysis. In *IEEE Transactions on Circuits and Systems for Video Technology*, volume 9, Dec. 1999.

[4] A. G. Hauptmann. Speech recognition in the informedia digital video library: Uses and limitations. In *ICTAI*, 1995.

[5] M. Irani and P. Anandan. Video indexing based on mosaic representations. In *Proceedings of the IEEE*, volume 86, 1998.

[6] M. Lee, W. Chen, C. Lin, C. Gu, T. Markoc, S. Zabinsky, and R. Szeliski. A layered video object coding system using sprite and affine motion model. In *Proceedings IEEE Transactions on Circuits and Systems for Video Technology*, volume 7, Feb. 1997.

[7] M. Massey and W. Bender. Salient stills: Process and practice. In *IBM Research Journal*, volume 35, 1996.

[8] S. Pfeiffer, R. Lienhart, S. Fischer, and W. Effelsberg. Abstracting digital movies automatically. In *Journal of Visuall Communication and Image Representation*, volume 7, 1996.

[9] P. Pritchett and A. Zisserman. Wide baseline stereo matching. In *Proceedings ICCV*, pages 754–760, January 1998.

[10] M. Smith and T. Kanade. Video skimming and characterization through the combination of image and language understanding. In *Proceedings of the IEEE International Worksop on Content-Based Access of Image and Video Databases (ICCV'98)*, 1998.

[11] T. Tuytelaars and L. Van Gool. Wide baseline stereo based on local, affinely invariant regions. In *British Machine Vision Conference*, pages 412–422, 2000.

[12] J. Wang and E. Adelson. Representing moving images with layers. In *IEEE Transactions on Image Processing*, volume 3, Sep. 1994.