Jiku director 2.0: a mobile video mashup system with zoom and pan using motion maps

Duong Trung Dung Nguyen, Axel Carlier, Wei Tsang Ooi, Vincent Charvillat

Jiku Director 2.0: A Mobile Video Mashup System with Zoom and Pan Using Motion Maps

Duong-Trung-Dung Nguyen
National University of Singapore
nguyend1@comp.nus.edu.sg

Axel Carlier
Université de Toulouse
axel.carlier@enseeiht.fr

Wei Tsang Ooi
National University of Singapore
ooiwt@comp.nus.edu.sg

Vincent Charvillat
Université de Toulouse
vincent.charvillat@enseeiht.fr

ABSTRACT

In this demonstration, we show an automated mobile video mashup system that takes a set of videos filming the same scene as input, and generate an output mashup video consisting of temporally coherent clips selected from these input videos. The key difference of our system over the existing state-of-the-art is that it can generate virtual close-up shots and three camera operations: zooming in, zooming out, and panning, automatically. To achieve this, the system first computes the motion maps of the input videos and then determines a set of rectangles that correspond to highly interesting regions (in terms of motion). The choice of which shot types to use is done heuristically, ensuring diversity and coherency in the content presented in the mashup.

Keywords: Mobile Video; Virtual Director; Video Mashup

1. INTRODUCTION

The proliferation of mobile devices equipped with high-quality cameras and video capturing software has led to a new phenomenon – during public events, it is common to have many people filming the scene at the same time. A single video filmed by a user, however, might not capture the scene completely, due to tired arms, depleted battery, limited viewing angle, etc. Furthermore, mobile videos tend to be of low quality, due to shaking hands, occlusion, and overexposure/underexposure. Even if the above problems are avoided, a video filmed from one single viewpoint is monotonous and could be boring to watch.

To address these issues, researchers have proposed algorithms and systems that automatically stitch the videos filmed by multiple users (of the same event) into a mashup video that is temporally coherent, by switching between the videos. The system operates much like a human director that switches between different cameras shooting an event. Two such prominent systems are the *Virtual Director* system proposed by Shrestha et al. [3] and the *Jiku Director* system developed by us [1].

We have extended Jiku Director with the ability to automatically zoom and pan. This extension is the focus of this demonstration and is motivated by the following. The existing systems produce a mashup video that always shows the scenes as filmed by the users. This limitation has two implications. First, the camera motions in these mashups are limited only to the physical movement of the cameras or zoom control operated by the users when they filmed the scene. In contrast, human directors often adopt zooming in, zooming out, and panning, as a way to add dynamics into the camera shots. Second, the field size[1] of the produced mashup is limited by that decided by the users when they filmed the scene (which are mostly long to medium shots). In contrast, human directors use a mixed of shot types, including close-up shots to affect the narrative of the production.

The new version of Jiku Director (2.0) addresses the lack of camera motions and close-up shots in the following way. The system, besides deciding which input video to use in the output mashup video, also decides a region of interest (ROI) in the selected input videos. The ROI can then be cropped to produce a close-up shot. Virtual camera motions can be produced in the mashup (through cropping) to simulate zooming effects in to, or out from, the ROI. The system could also decide to have two ROIs, and virtually pans the camera from one ROI to another.

The next section describes (i) how the system computes the ROIs, (ii) how it decides which ROI to use, and finally (iii) which shot type/camera motions to generate in the output.

2. JIKU DIRECTOR 2.0

2.1 Background

We begin with a brief description of Jiku Director. Readers are referred to [1] for a description of the system, and to [2] for a detailed description of MoViMash, the underlying mashup algorithm.

Jiku Director is a Web-based application for sharing user-generated videos. A screenshot of the system is shown in Figure 1. A user can upload videos of an event which they have filmed and share it with others. The videos are organized based on events. For each event, Jiku Director can generate a "representative" video of the event, by mashing up all the videos (of that event) that have

[1]Field size refers to how much of the subject and the surrounding area is framed in a shot

been uploaded. At each decision time, the mashup algorithm first filters out bad quality videos (shaky, tilted, occluded) through simple content-based analysis. The remaining videos are then selected based on the view angle and distance to the stage of the event. The algorithm imitates a human director (through learning from professionally edited videos) in deciding which angle and distance to choose and what is the length of the clip to use from the chosen video. If there is more than one video with approximately the same angle and distance, the algorithm picks the clip from the best quality video camera that has not been recently picked before.

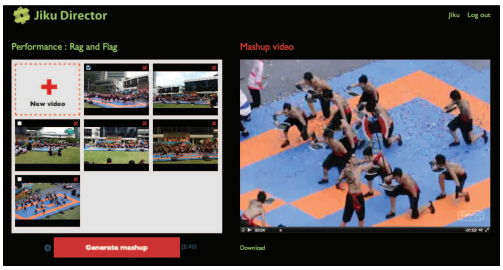

Figure 1: Screenshot of the Web Interface of Jiku Director

2.2 Computing Candidate ROIs

To support zooming and panning in Jiku Director, we first need to compute the "interesting regions" for each video. A region is defined by a 3D ((x, y) coordinates and time) bounding box, parameterized by its position, width, height, starting and ending frame.

To build the bounding boxes, we start by computing the motion maps at each frame, normalizing the difference between the current frame and the previous frame. The motion map obtained is viewed as a probability distribution, following which we sample 500 points for each frame. We therefore build a set of moving points with three coordinates: two spatial coordinates (in pixels), and one temporal coordinate (frame index).

We use Mean-Shift clustering to cluster the moving points. Mean-Shift converges towards the modes of a distribution, which, in our case, are the spatio-temporal regions with the highest quantity of motion. It also associates each point from the data to one mode to build clusters. We ignore transient clusters that exist less than one second or clusters with too few (< 30) moving points in each frame.

For each remaining cluster, we create one corresponding bounding box. The starting frame (resp. ending frame) of the bounding box is the minimum (resp. maximum) of the third coordinate value from all points in the cluster. To determine the remaining parameters of the bounding box, we use the Levenberg-Marquardt algorithm to compute the smaller bounding box that contains at least 95% of all points in the cluster. We use this threshold to limit the size of the bounding boxes, and allow for larger zooms during the video mashup. Figure 2 shows bounding boxes from two sample frames filmed from two different cameras.

Figure 2: Bounding Boxes

2.3 Deciding the ROI

Each object appearing in each frame is associated with a bounding box and a centroid with a weight indicating the average number of clustering points of the associated cluster per frame.

We rank the bounding box of each object in order of their importance in the scene. In each video clip, each object i has an *importance* value, $I(i,t) = \sum_j v_j(t)w_j(t)$, where $v_j(t)$ is 1 if object j appears in the bounding box i at time t and 0 otherwise; $w_j(t)$ is the average number of clustering points of cluster j per frame at time t. The intuition here is that a bounding box with more clustering points is more important. Note that even though each bounding box is associated with an object, other objects may appear in the bounding box. Naturally, a bounding box that contains multiple important objects becomes important.

2.4 Choosing the Shot Type

The output of our heuristic for selecting the next shot will be a sequence of *shots* (t, j, B, l, m), where t is the beginning time of the shot, j is the camera to choose the shot from, B is the set of regions of interest to use in this shot, l is the length of the shot, and m is the virtual camera movement in this shot.[2]

Jiku Director will decide the best camera j in each shot. Once j is fixed, we determine t, B, l, and m as follows. Let L_{max} be the maximum possible shot length (we use $L_{max} = 7s$). To determine the current shot, we look at the most important bounding boxes and objects at each time instance for the next L_{max} seconds in the video shot by camera j. B is set to the most important bounding boxes that frequently appear in the next L_{max}. To decide l (and therefore, the next t), we stop the current shot at the time when these bounding boxes disappear.

It remains to determine m for the current shot. The choice of m can be either: wide-angle shot, close-up shot, move the camera (zoom in, zoom out, or pan). The heuristic follows the following principles: (i) there should not be frequent camera movement, and (ii) the choice of m should observe continuity editing. It should avoid an abrupt illogical shot transition. To do that, a wide-angle shot is used after each close-up, zoom or pan shot as a re-establishing shot.

3. DEMONSTRATION

During the demonstration, we will show the features of Jiku Director in general, with a focus on generation of a mashup video that consists of zoom and pan camera motion and close up shots.

Acknowledgement. This research is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office.

4. REFERENCES

[1] D.-T.-D. Nguyen, M. Saini, V.-T. Nguyen, and W. T. Ooi. Jiku Director: a mobile video mashup system. In *Proceedings of the ACM Multimedia*, MM '13, pages 477–478, Barcelona, Spain, 2013.

[2] M. K. Saini, R. Gadde, S. Yan, and W. T. Ooi. MoViMash: Online mobile video mashup. In *Proceedings of the ACM Multimedia*, MM '12, pages 139–148, Nara, Japan, 2012.

[3] P. Shrestha, H. Weda, M. Barbieri, E. H. Aarts, et al. Automatic mashup generation from multiple-camera concert recordings. In *Proceedings of the ACM Multimedia*, MM '10, pages 541–550, Florence, Italy, 2010.

[2]In contrast, the first version of Jiku Director only decides on (t, j, l).