Using Eye-Tracking Data for Automatic Film Comic Creation

Masahiro Toyoura* Tomoya Sawada Mamoru Kunihiro Xiaoyang Mao†
University of Yamanashi

Abstract

A film comic is a kind of art work representing a movie story as a comic. It uses the images of the movie as panels. Verbal information such as dialogue and narrations is represented in word balloons. A key issue in creating film comics is how to select images which are significant in conveying the story of the movie. Such significance of images is inherently semantic and context-dependent and hence, technologies purely based on image analysis usually fail to produce good results. On the other hand, the word balloon arrangement requires understanding not only the semantic of images but also the verbal information, which is difficult except for the case the script of the movie is available. This paper describes a new attempt to use eye-tracking data for the automatic creation of a film comic from a movie. Patterns of eye movement are analyzed for detecting the change of scenes and gaze information is used for automatically finding the location for inserting and directing the word balloons. Our experiments showed that the proposed technique can largely improve the selection of significant images compared with the method using image features only and realize the automatic balloon arrangement.

CR Categories: I.2.10 [Artificial Intelligence]: Vision and Scene Understanding—Video analysis; I.3.3 [Computer Graphics]: Graphics Utilities—Picture description languages

Keywords: film comic, image trimming, balloon arrangement, visual attention, region of interest

1 Introduction

Comic-like video summaries generated from cartoon animations are called as film comics. Most of popular cartoon animations are reformed to film comics and we can buy printed film comics at book stores. Traditionally, film comics are manually created by professional editors who are familiar with comic styles. The images of the cartoon animations are manually selected, trimmed and arranged into the comic-like layouts. Verbal information of the movie, such as dialogue and narrations, are inserted into the correspondent panels as word balloons. Since a movie usually consists of huge number of images, such editing task is very tedious and time consuming. Recently, several researches on applying computational approach for film comic creation have been published. One representative work is done by Preuß and Loviscach [Preußand Loviscach 2007], which provided an easy-to-use interface for supporting the whole procedure of film comic creation.

In this paper we present a new idea of using eye-tracking data to automate the whole process of film comic creation. To convert a movie into a film comic, we first invite a subject to watch the movie and record his/her eye positions with an eye-tracker. Then our technique takes the movie together with the eye-tracking data as input, and automatically converts the movie into a film comic by using the eye-tracking data and image features in a mutually supplemental way. It is known that eye movement and gaze information provide good cues to the interpretation of scenes by viewers. Existing studies [Ma et al. 2002; Li et al. 2010] showed the effectiveness of applying visual attention model for defining the significance of images in creating video summarization. Those techniques use the computational model of visual attention which is mainly based on the low level features of images and motions. The real attention of viewers, however, is largely dependent to the semantics and context of videos as well as their age, sex and cultural background. By using the eye-tracking data of the viewers, we can provide a better prediction to the attention of viewers. Furthermore, by using a subject from the assumed population of readers, e.g. inviting a child as the subject for creating a film comic mainly targeting at children, we can even create film comics adapted to a particular population of readers.

Video summarization is the field most closely related to film comic generation. Video summarization and film comic generation share the two major technique issues: how to select appropriate images from a video and how to arrange the images to effectively depicting the contents of the video. Video summarization has been well studied in the past decade and researchers have challenged the problem through various approaches ranging from computer vision [Calic et al. 2007; Girgensohn 2003; Ngo et al. 2003; Boreczky et al. 2000], text mining [Chen et al. 2009], fMRI [Hu et al. 2010], and eye-tracking [Peng et al. 2009; Goldstein et al. 2007]. While a video summary is mainly for providing the overview of a video or serving as the index enabling users to quickly browse the required information in the original video, film comic is an alternative of movie for storytelling and hence should enable readers to easily understand the whole story without refereeing to the original movie. Therefore, selecting frames which are critical in conveying the story and the representation of dialogue and narrations lines are particular important in film comic generation. We propose to identify those frames by analyzing the patterns of eye movement. We use gaze information to identify the ROI (region of interest) and place word balloons in a way avoiding occluding those regions. Moreover, since a user's eyes are usually drawn to the speaking objects, gaze information can be used to easily direct the balloons to the speakers.

2 Framework of the Proposed Film Comic Generation Technique

This section presents the general framework of our film comic generation technique. Before going into details, let us first introduce several technical terms of movie and comic. Frame images are individual ones of a movie. A shot consists of a set of frame images recorded at one time. A cut is the boundary between two shots. A scene is a single or a series of shots describing a single action. Panels are the sub-regions constituting the pages of a film comic.

As shown in Figure 1, our technique takes a video with caption data and the eye-tracking data of a viewer as the input, and generates the

*e-mail: mtoyoura@yamanashi.ac.jp
†e-mail: mao@yamanashi.ac.jp

ETRA 2012, Santa Barbara, CA, March 28 – 30, 2012.
© 2012 ACM 978-1-4503-1225-7/12/0003 $10.00

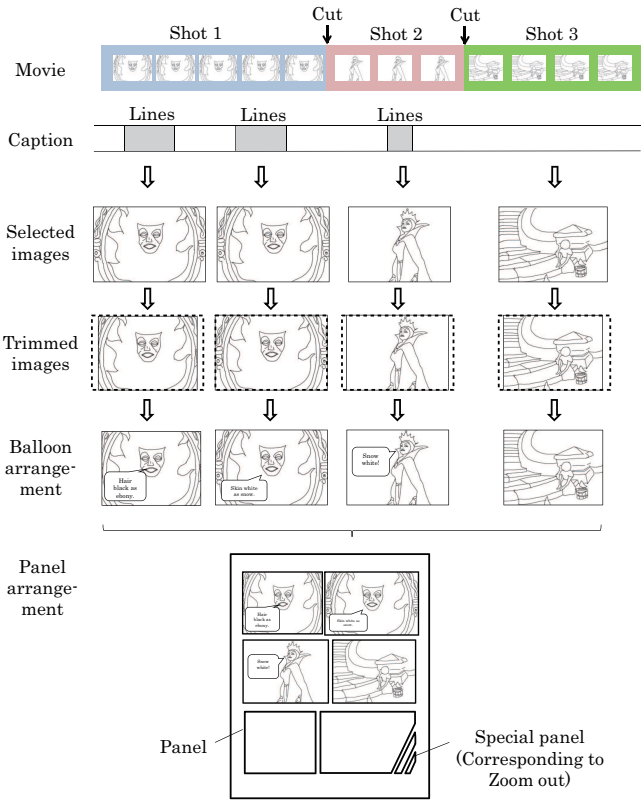

Figure 1: *The framework of the proposed automatic film comic generation technique.*

film comic in the following 6 steps:

1. Divide the movie into shots.
2. Analyze the caption data to establish the correspondence between lines and frame images.
3. Select images from each shot.
4. Trim the images to fit into panels.
5. Insert word balloon to the panel if the image has lines.
6. Arrange the panels into the comic layout.

The first step detects the cut through comparing the color of successive frame images [Porter 2004]. A 6-bits binary number is computed for each pixel by taking the two most significant bits from R, G, B values, respectively. A histogram of the 6-bits numbers is calculated for each frame image and the difference of the histograms in two adjacent frame images indicates the likelihood of cut.

Our technique uses captions for obtaining verbal information. The captions consist the information of lines with their beginning and ending frame numbers. In Step 2, we use such information to establish the correspondence between frames and lines, and frame images with corresponding lines will be selected at Step 3.

At Step 3, the patterns of eye movements are first analyzed to decide whether a shot needs to be further divided into different parts. Then frame images are selected from each shot or part based on the following criteria.

1. Select all frames corresponding to a line.
2. Select the middle frame images from each shot or parts.

The selected images are trimmed to fit into panels at Step 4. Here gaze information is used for detecting ROI and the images are trimmed in away without cutting the ROI away. At Step 5, for the images with corresponding lines, the lines are arranged into balloons. The balloons are inserted into panels in a way avoiding occluding the ROI defined with gaze information. Finally at Step 6, panels with balloons are arranged into the layout of comic. Eye-tracking data are mainly used in step 1, 2 and 5. We describe the details in the following two sections.

3 Image Selection Based on Eye Movement

A major difference between the video summary and the film comic is whether images are selected in consideration of individual shots. In film comics, one or more images should be selected from each shot. Each line of the movie has one or more corresponding panels. The difference is inherently caused by the difference of intended purposes. Grasping the overview of a movie is the main purpose of video summary, meanwhile representing the whole story of a movie is the main purpose of film comic.

As described in the previous section, the cut detection at Step 1 divides the movie into a series of shots. When a shot includes the change of content or focusing object, more frame images should be selected for depicting the development of the story. An extreme example is one-shot for one-scene, with which only one frame image might be selected from a whole scene if we simply select one frame image from one shot.

Let us explain this problem in more detail with the example of a shot including the frame images shown in Figure 2(a). The black crosses in the image represent the tracked eye position. The shot comes from Pinocchio. The old man and Pinocchio become the focus alternately in the shot. Such change cannot be detected with the cut detection algorithm since the color of images does not change much in the shot. However, representing the interaction between the old man and the Pinocchio is important in helping the readers to understand the story. To solve the problem, we propose to use eye tracking data to identify the change of content within a shot. As the viewer tends to gaze at focused objects, a large movement of eye position usually indicates the change of focused object. We divide the shot into two parts when a large eye movement is detected and select one representative image from each part. The shot in Figure 2(a) is divided into 3 parts and as the result, 2 more frame images are selected with the using of eye-tracking data from the shot. Compared with the result generated without using eye-tracking data, those additionally selected images can help the readers better understanding the context of story.

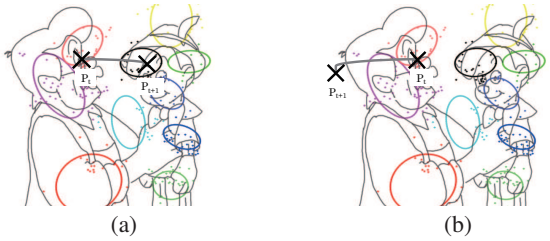

Figure 2: *Detecting the change of focus with the pattern of eye movement. (a) An eye movement corresponding to the change of focus; (b) A meaningless eye movement considered to be a noise.*

Note that not all large eye movements are meaningful for identifying the change of content or focus in the shot. We assume that

a meaningful eye movement is the one moving from one "informative area" to another "informative area." To define the informative areas, we extract KLT (Kanade-Lucas-Tomasi Feature Tracker) [Shi and Tomasi 1994] in the frame images and cluster the feature points with ellipse shaped k-means method. In Figure 2, the clusters are represented with the ellipses. If the eyes move from one cluster to another as shown in Figure 2(a), we assume the movement is caused by the switching of focused objects, and we divide the shot into two parts at the frame where the movement occurred. If the eyes moved to a position not enclosed in any clusters as shown in Figure 2(b), we assume the movement to be meaningless. The detection is performed in the following way: Denoting the tracked eye position on the t_{th} frame image as P_t, we first detect whether a large eye movement occurred by testing whether the eye position on the $t + 1_{th}$ is within the same cluster ellipse. If not, a large eye movement occurred and we further check if the movement is meaningful by checking whether P_{t+1} is enclosed in any cluster of $t + 1_{th}$ frame and whether the eye position keeps staying in the same cluster for several more frames.

4 Inserting Word Balloons With Gaze Information

The lines are arranged into the word balloons in panels. The balloons should avoid to be placed on important regions in the panels. The tails of the balloon should head to speaking objects. We solve the tasks with eye tracking data.

We first detect the speaking objects. Although face detection technologies are available, they would not work mostly in case of cartoon animation even if the speaking object is human. All existing face detection technologies use the features extracted from the photograph of human face. In case of cartoon animation, objects are usually deformed and represented in a particular exaggerated style, which means that both the shape and shading information can be completely different from that of the photographs. We solve the problem by using gaze information of viewers based on the fact that viewers tend to pay attention to the speaking objects. For all the frame images corresponding to the same lines, we first perform the clustering of eye positions. The cluster with the maximum number of tracked eye positions, which is regarded as the most attended region, is assumed to be the position of the speaking object. The tails of balloons are directed to the barycenter of the cluster.

To decide the positions of balloons, we first divide the panel into 9 sub-regions with the horizontal and vertical lines crossing at the center of the two clusters with the maximum number of eye positions (Figure 3). Then we arrange the balloon at the center of the region with the maximum area, so as to avoid covering the centers of the two largest clusters with the possible maximum area. By keeping a margin to the boundary of the sub-reion, we can expect that the balloon would not cover the centers as well as their surrounding regions of the two largest clusters.

5 Experiments

In experiments, we used a LCD monitor of 23 inches for displaying movies. The eye positions of viewers were tracked in 60Hz with EMR-AT VOXER of nac Image Technology. The movies were "Snow White" and "Pinocchio." The viewer was a male university student in his 20s. He watched the initial 10 minutes of the movies.

With the color histogram based shot detection only, totally 62 and 32 frames were selected for the 10 minutes clips of "Snow White" and "Pinocchio" respectively. Combining the eye tracking data, we obtained 182 and 215 frames. Most of these additional frames are selected from the shots in which objects change their poses largely or interact with each other. The selected frames represent either the poses which are critical in understanding the story or the switching of focus during the interaction.

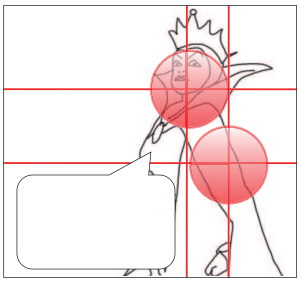

Figure 3: *Balloon arrangement.*

Next let us verify the effectiveness of using eye-tracking data for balloon arrangement. Table 1 shows the results.

Table 1: *Result of balloon arrangement.*

	Lines	Success	Fail
Snow White	99	77	22
Pinocchio	99	75	24

We define the success of balloon arrangement as the case that a balloon does not cover any important regions and its tail heads to the speaking objects. As shown in Table 1, the success rates are 77.8% and 75.8% for Snow White and Pinocchio, respectively. Two kinds of failure were observed; (1) A balloon is directed to another position instead of a speaking object. This happened when the focusing object was different from the speaking object or when the speaking object did not appear in the image; (2) The speaking object was partially occluded by the balloon. This occurred either when the focusing object was different from the speaking object or when the speaking object or the balloon is too large to avoid overlapping. The detection of speaking objects can be refined by analyzing the motion in the maximum cluster of eye positions. When the maximum cluster is not likely the speaking object, we can remove the tail of balloon. The overlapping of a balloon and a speaking object can be avoided by using a more sophisticate algorithm to find the largest available area in the image and adapting the shape of balloon to that of the available area.

6 Conclusions and Future Works

We presented a new idea of using eye-tracking data for automatic film comic creation. Eye tracking data is used for the selection of important images and balloon arrangement. Although the results were not perfect, our experiment result demonstrated the feasibility of using eye-tracking data in solving the problems which cannot be solved with image processing technique only.

The result shown in this paper was generated with the eye-tracking data of one subject. We expect to improve the reliability of the result by using more subjects. Exploring other usage of eye tracking data in film comic generation is one of the major future research directions. A potential topic is to use eye-tracking data to improve the representation of comic, such as using the patterns of eye movements in deciding the layout of panels and the gaze information to estimate the importance of frame images so as to change the size of the corresponding panel. User adaptation is another important

future work. As mentioned in the beginning of the paper, one interesting advantage of using eye-tracking data is that we can create a film comic adapted to a particular population of readers by using the eye-tracking data of the viewer from that population.

Acknowledgements

This work was supported by KAKENHI 21300033 Grant-in-Aid for Scientific Research (B), Japan Society for the Promotion of Science (JSPS).

References

BORECZKY, J. S., GIRGENSOHN, A., GOLOVCHINSKY, G., AND UCHIHASHI, S. 2000. An interactive comic book presentation for exploring video. In *ACM Conference on Computer-Human Interaction (CHI)*, 185–192.

CALIC, J., GIBSON, D. P., AND CAMPBELL, N. W. 2007. Efficient layout of comic-like video summaries. *IEEE Transactions on Circuits and Systems for Video Technology 17*, 7 (July), 931–936.

CHEN, B., WANG, J., AND WANG, J. 2009. A novel video summarization based on mining the story-structure and semantic relations among concept entities. *IEEE Transactions on Multimedia 11*, 2, 295–312.

GIRGENSOHN, A. 2003. A fast layout algorithm for visual video summaries. In *IEEE International Conference on Multimedia and Expo*, vol. 2, 77–80.

GOLDSTEIN, R. B., WOODS, R. L., AND PELI, E. 2007. Where people look when watching movies: Do all viewers look at the same place? In *Computers in Biology and Medicine*, vol. 37, 957–964.

HU, X., DENG, F., LI, K., ZHANG, T., CHEN, H., JIANG, X., LV, J., ZHU, D., FARACO, C., ZHANG, D., ET AL. 2010. Bridging low-level features and high-level semantics via fmri brain imaging for video classification. In *Proceedings of the international conference on Multimedia*, ACM, 451–460.

LI, K., GUO, L., FARACO, C., ZHU, D., DENG, F., ZHANG, T., JIANG, X., ZHANG, D., CHEN, H., HU, X., MILLER, S., AND LIUOTHERS, T. 2010. Human-centered attention models for video summarization. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, ACM, 27.

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

NGO, C.-W., PONG, T.-C., AND ZHANG, H.-J. 2003. Motion analysis and segmentation through spatio-temporal slices processing. *IEEE Transactions on Image Processing 12*, 341–355.

PENG, W.-T., HUANG, W.-J., CHU, W.-T., CHOU, C.-N., CHANG, W.-Y., CHANG, C.-H., AND HUNG, Y.-P. 2009. A user experience model for home video summarization. In *International Conference on Multimedia Modeling*, 484–495.

PORTER, S. V. 2004. *Video Segmentation and Indexing using Motion Estimation*. PhD thesis, University of Bristol.

PREUSS, J., AND LOVISCACH, J. 2007. From movie to comics, informed by the screenplay. *ACM SIGGRAPH (Poster)*.

SHI, J., AND TOMASI, C. 1994. Good features to track. In *IEEE Conference on Computer Vision and Pattern Recognition*, 593–600.