Film Comic Generation with Eye Tracking

Tomoya Sawada Masahiro Toyoura Xiaoyang Mao
University of Yamanashi

Abstract. Automatic generation of film comic requires solving several challenging problems such as selecting important frames well conveying the whole story, trimming the frames to fit the shape of panels without corrupting the composition of original image and arranging visually pleasing speech balloons without hiding important objects in the panel. We propose a novel approach to the automatic generation of film comic. The key idea is to aggregate eye-tracking data and image features into a computational map, called iMap, for quantitatively measuring the importance of frames in terms of story content and user attention. The transition of iMap in time sequences provides the solution to frame selection. Word balloon arrangement and image trimming are realized as the results of optimizing the energy functions derived from the iMap.

Keywords: Visual attention, eye-tracking, film comic, video processing, frame selection, image trimming, balloon arrangement.

1 Introduction

Film comic, also called cine-manga, is a kind of art medium created by editing the frames of a movie into a book in comic style. It is loved by a wide range of readers from little babies to comic manias around the world. Traditionally, film comics are manually created by professional editors who are familiar with comic styles. The frames of movies are manually selected, trimmed and arranged into the comic-like layouts. Verbal information of the movie, such as dialogues and narrations, are inserted into the correspondent panels as word balloons. Since a movie usually consists of a huge number of frames, those editing tasks can be very tedious and time consuming.

Recently, several research works have been conducted trying to provide computational support to the generation of film comic [1-4]. The work by R. Hong *et al*. [3] first succeeded in automating the whole procedure of film comic by employing modern computer vision technologies for extracting key frames and for allocating the word balloons. However, their method replies on existing human face and speaking lip detecting technologies for identifying speakers, and are not applicable to cartoon animations, which is the main target of film comics, because the appearance of characters in cartoon animation is usually very different from that of human face. Actually, many issues of automatic film comic generation cannot be solved with naive image/video processing or even by combining text analysis. For example, a subtle change of character's posture which causes no remarkable change to image features

may be very critical in conveying the story. When placing a word balloon, only avoiding occluding the speaker is usually not enough since an object about which the speaker is talking may also be very important. All these problems are story or even user dependent.

In this paper, we present a new approach of using eye-tracking data to automate the whole process of film comic creation. It is known that eye movement and gaze information provides good cues to the interpretation of scenes by viewers [5]. Existing studies [6-8] showed the effectiveness of applying visual attention model for measuring the significance of images in creating video summarization. Those techniques use the computation 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 on the content of a story as well as the personal background of the viewers, such as their age, sex and cultural background. By using the eye-tracking data, we can provide a better prediction to the attention of viewers. Furthermore, by using subjects from the assumed population of readers, e.g. using the eye-tracking data of children for creating a film comic mainly targeting at children, we can even create film comics adapted to a particular population.

The major contribution of this paper can be summarized as follows:

1. Proposal of a new framework for automatically converting a movie into a comic based on eye-tracking data.
2. Proposal of iMap, a new computational map which combines eye-tracking data and image feature for quantitatively measuring the local informativeness in a movie frame.
3. Proposal of a new algorithm for detecting critical frames based on iMap .
4. Proposal of the optimization scheme for achieving desirable image trimming and balloon arrangement.

The remainder of the paper is organized as follows: Section 2 reviews the related work. Section 3 gives the overview of the proposed method and Section 4 introduces iMap together with its construction algorithm. The detailed algorithms of frame selection, trimming, balloon allocation are described in Section 5. Section 6 is about implementation and experiment and Section 7 concludes paper.

2 Related works

A pilot work on film comic generation was done by W.I. Hwang et al. [1]. Assuming the correspondence between a balloon and a speaker is known, their method can automatically allocate balloons following the rules of comic, but the major tasks of film comic generation such as the selection of frame images were done manually. Preuß et al. [2] proposed an automatic system for converting movies into comics, but their method assumes that the screenplay of the movie is available and hence limits its application to a very special case. More recently, R. Hong et al. [3] presented a full automatic system which employs face detector, lip motion analysis, and motion analysis to realize the automatic script-face mapping and key-scene extraction. But as

mentioned in the previous section, their technique cannot be used for cartoon animations. M. Toyoura *et al.*[4] proposed a technique to automatically detect typical camera works, such as zooming, panning, fade-in/fade-out, and represent them in a particular comic style. The same function is supported in our system.

Video summarization [9] is the field most closely related to film comic generation. Both video summarization and film comic generation share the two major technique issues: how to select appropriate images from a video/movie and how to arrange the images to effectively depicting the contents of the video/movie. Video summarization has been well studied in the past decade and researchers have challenged the problem through various approaches ranging from computer vision [9-10], text mining [11], to fMRI [12]. Several researchers proposed to use human visual attention model for video segmentation and skimming [6-7]. Those methods were further enhanced through combining multiple visual cues [8] as well as high level semantic and cognitive information [13]. W. Peng et al. proposed an approach using both eye-tracking and facial expression detection for selecting important frames in home video summarization [14]. Comic style was also explored in the field of video summarization for achieving an effective representation of video summaries [15]. 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, many technologies of video summarization cannot be applied to film comic generation directly. For example, W. Peng et al. selected the frame sequence with long gaze fixation as the important frames to be included in the summary [14]. But from the viewpoint of story development, a frame causing the change of attention should be more important. We succeeded in selecting such frames by detecting the significant transition of iMap.

3 Framework

Figure 1 depicts the framework of the proposed technique. Given a movie including subtitle information, our technique converts it into a comic in following 10 steps. First, sample viewers are invited to watch the movie and their eye positions are recorded with an eye-tracker. At the same time SURF image feature detection is performed for each frame of the movie (Step 2). Then an iMap is generated from the eye-tracking data and SURF feature (Step 3). Based on iMap, the frame sequence of the movie is segmented into groups of frames called *shots* (Step 4). Those shorts are subjected to camera-works detection (Step 5). Then for each shot, the first frame, the frames with speech line and the frames for representing camera works are selected (Step 6). Based on the number of selected frames and the information on camera works, panel layout of the comic is generated (Step 7). The information required for trimming the frames and for placing the balloons are computed by referring to iMap (Step 8&9). Finally a comic book is created by arranging the trimmed frames into the comic layout and balloons with speech lines are inserted to the frames (Step 10). We use existing technologies for Step 5 and 7, and the other steps are realized with newly developed techniques, whose details are described in the next section.

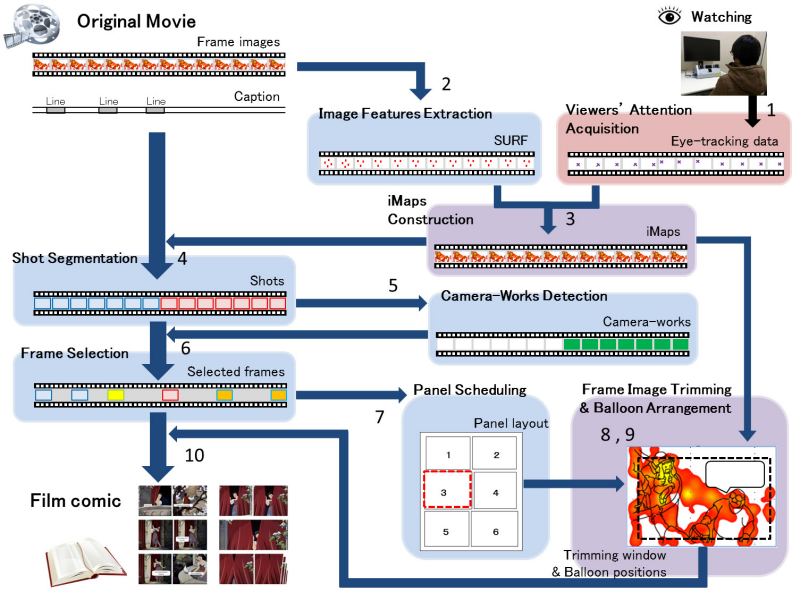


Figure 1 The framework of film comic generation with proposed technique.

4 Film comic generation with iMap

4.1 Construction of iMap

A good film comic should contain the frames that are important for understanding the story and each of the frames should be represented in a way well preserving the original information while being visually pleasing as a comic. Those requirements give rise to the issue of measuring how informative a frame or a region in a frame is in conveying the story. We address the issue by combing eye-tracking data and image features into iMap, a probability distribution map providing a quantitative measure to the local informativeness in each frame. It is known that eye movement plays an important role in event perception [5]. Our eyes are likely to be drawn to a region which is most relevant to the story development and a big eye movement usually provides a cue to the change of events. However, eye-tracking data usually consists of noise. It is also difficult to distinguish an unintentional or meaningless fixation from a deliberate or purposeful one. To solve the problem, we first convert the raw eye-tracking data of multiple viewers into an *attention map,* a probability distribution of attention estimated with Gaussian kernel. At the same time we generate a *feature map* which is a probability distribution estimated from SURF (Speeded Up Robust Feature) [16]. We adopt SURF for taking its advantage in speed and robustness in tracking features across frames. Then we construct iMap as the aggregation of the attention map and the feature map. The significance of combing eye-tracking data and image feature lies in the fact that a fixation at a position with little image feature such

as edges and corners is likely to be a meaningless one and the resulting iMap gives a more reliable prediction to how informative the position is.

A feature map M^t_f representing the probability distribution of image features in t^{th} frame is calculated as follows:

$$M^t_f = f^t_{SURF}(x,y) \otimes G(0,\sigma^2_f) \qquad f^t_{SURF}(x,y) = \begin{cases} 1 & if\ (x,y)\ is\ SURF \\ 0 & if\ (x,y)\ is\ not\ SURF \end{cases} \tag{1}$$

where $G(0,\sigma^2_f)$ is a 2D Gaussian kernel with average 0 and variance σ^2_f. $\otimes$ is convolution. If pixel (x,y) is SURF, its contribution to the probability distribution of image feature is estimated with the Gaussian kernel.

Similarly, an attention map M^t representing the distribution of n sample viewer's attention is calculated as follows:

$$M^t = \sum_{i=1}^{n} f^t_{iEYE}(x,ty) \otimes G(0,\sigma^2_a) \qquad f^t_{iEYE}(x,y) = \begin{cases} 1 & if\ (x,y)\ is\ an\ \ eye\ position \\ 0 & if\ (x,y)\ is\ not\ an\ eye\ position \end{cases} \tag{2}$$

iMap M^t of t^{th} frame is constructed from maps M^t_f and M^t_a as follows:

$$M^t = M^t_f * M^t_a . \tag{3}$$

Since M^t is the product of M^t_f and M^t_a, a high value of M^t indicates an informative region in terms of both image features and viewers' attention. When there is a significant change of M^t in time sequences, there should be a transition of either/both of image features or/and viewers' attention. By detecting such transition, the movie can be segmented into the shots consisting of frames of the same contents.

Although M^t gives a good solution to the detection of content transition among frames, it does not contain enough information for performing frame image trimming and balloon arrangement, which requires measuring the local informativeness in the context given by neighboring frames also. We further extend iMap to encounter the probability distribution in time dimension. An iMap representing the spatial-temporal informativeness is calculated from the neighboring frames of t^{th} frame as follows:

$$M^{Nt} = \sum_{s \in N(t)} G(t,\sigma^2_t)(s) * M^s , \tag{4}$$

where $N(t)$ is the set of neighboring frames of t^{th} frame and the contribution of each neighboring frames is given by Gaussian function $G(t,\sigma^2_t)$ (s) with highest value at t^{th} frame.

4.2 Frame selection

To select frames which are important for conveying the story of the movie, we first segment the movie into shots by detecting significant change of iMaps across adjacent frames. The set of transition frames $\Omega(t)$ are detected as

$$\Omega(t) = \left\{ t \,|\, \| M^{Nt+1} - M^{Nt} \| > T \right\} \qquad T = \iint_M \| G(x,y) - G(x - x_d, y - y_d) \| \tag{5}$$

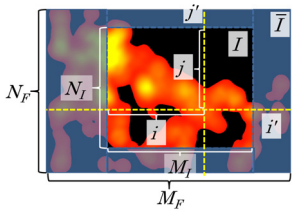
Where $G(x,y)$ is the Gaussian filter used for computing iMap, with $\|*\|$ denotes L_2 norm, $d = \sqrt{x_d + y_d}$ is the possible distance between the eye positions at adjacent frames when a saccade occurs. A saccade is a fast movement of an eye between fixations. The central part of our retina, known as the fovea, plays a critical role in resolving objects. We move our eyes quickly toward a target so as to sense it with

greater resolution. Therefore, if a saccade is detected, we can say that the region of interest has changed, which indicates a possible transition of story contents.

After transition frames are detected, the movie is segmented into the group of frames, say shots, at the transition frames. Then the first frame of each shot, which is the transition frame, is selected for being included in the film comic. All the frames with speech lines are also selected. If a shot contains special camera-works, multiple frames are selected depending on the type of camera-works.

4.3 Image trimming

If the shape and size of a selected frame image are different from those of the corresponding panel, the frame image is trimmed and resized to fit into the panel. As our current implementation uses rectangular panels only, trimming is necessary only when the frame image and its corresponding panel have different aspect ratios. A good trimming should meet two requirements: First, the area being cropped should be less informative. Second, the composition of original frame should be preserved as possible. We achieve the best trimming by minimizing the following cost computed from iMap M^{Nt}:



$$E_{trim} = \frac{\sum_{(i',j')\in\bar{I}} m_{ij}}{S(\bar{I})} + \frac{\sum_{(i',j')\in I} m_{ij}\left(\left|\frac{i}{M_F - i} - \frac{i'}{M_I - i'}\right| + \left|\frac{j}{N_F - j} - \frac{j'}{N_I - j'}\right|\right)}{2Max(M_F, N_F)S(I)}. \quad (6)$$

Figure 2 Image trimming by referring to iMap M^{Nt}.

As shown in Figure 2, F denotes the original frame image, I the trimmed one and $\bar{I}$ the set of the pixels being cropped away with the trimming. (i',j') is the position in original frame image F for the pixel in I or $\bar{I}$. $S(I)$ and $S(\bar{I})$ are the total number of pixels (i,j) in I and $\bar{I}$, M_F, M_I, N_F, N_I are the width and height of the frame image F and trimmed image I, respectively. The first term in Eq (6) measures how less informative the region being cropped is and the second term evaluates whether the relative location of a pixel in original frame is preserved in the trimmed image. By multiplying the cost with the value in M^{Nt}, we give those informative pixels a higher priority for preserving their relative positions.

4.4 Arranging balloons

A balloon is placed into the panel of a selected frame with corresponding speech lines. Balloons should be placed in a way to meet 3 requirements: First, it should not occlude an important region. Second, it should be in a visually pleasing shape. Third, it should be spatially close to the speaker. For the first requirement, iMap provides the information about the importance of regions. In our current implementation, we use oval shaped balloons, which is one of the most popular shapes found in comics. To avoid generating long narrow shaped balloons, we use the ratio of the two axes of the ellipse as the measure to the aesthetic quality of the balloon. To measure the distance from the balloon to the speaker, the position of the speaker should be given. We have

tried out existing face detector and speaking lip detector, but found all of them failed to produce good results. The main reason is that the appearance of characters in cartoon animation are usually very unique, which makes it difficult to apply the existing face detector. Another reason is that a cartoon animation usually contains exaggerated motions throughout the movie and this makes it difficult to detect small motions such as a speaking lip. Put all together, a desirable balloon arrangement can be obtained by minimizing the following energy function:

$$E_{ballon} = a\frac{\sum_{(i,j)\in I_{balloon}} m_{ij}}{S(I_{balloon})} + b\frac{\left\| P_{balloon} - P_{I_{speaker}} \right\|}{M^2 + N^2} + c\frac{L_{short}}{L_{long}}, \quad (7)$$

where $I_{balloon}$ is the region occluded by a balloon, $P_{balloom}$, $P_{Ispeeker}$ are the gravity center of balloon and speaker region, L_{short}, L_{long} are the short and long axes of balloon, respectively. a, b, c are user given parameters for controlling the weight of each term. The values of L_{short}, L_{long} are dependent to the length of speech lines. L_{short} and L_{long} take discrete values to avoid breaking a line at the middle of a word. Given a speech line, we first build a look up table containing all the possible pair of L_{short}, L_{long} and loop through all the entries of the table during optimization.

After the shape and location of a balloon is decided, a tail is attached to the balloon and headed toward the speaker.

5 Experiments

We invited sample viewers (3 male university students in their 20s) to watch the beginning part of “Pinocchio” and “Snow White and the Seven Dwarfs” by Disney Inc. and a part of “Roman Holiday”. The movies were displayed on a 1920x1200 monitor and EMR-AT VOXER of Nac Image Technology was used for tracking subject’s eye positions at 60Hz. Frame rate of the movies was 29.97fps. Figure 8 are some example pages of the resulting film comics and Table 1 shows some statistics of the results for the two cartoon movies.

Figure 3 and Figure 4 compare the results of frame selection by the existing color histogram based video segmentation [10] and by our method. Figure 3 is the scene the cricket sneaks into the house in the freezing night trying to warm his hip up with the burnt stone from the fireplace. Since the color histogram does not change much through the scene, the result by existing work as shown in Figure 3(a) failed to detect enough frames required for representing such mise-en-scene. Figure 4 is a part of “Roman Holiday” where Mr. Joe and Princess Ann try to insert their hands into the mouth of truth. Princess Ann gets frightened and moves her hand toward the mouth timidly. Such process could not be visualized with the result of existing technique in Figure 4(a) because of the little change of color histogram. In Figure 4(b), our proposed method succeeded in catching such subtle but important transition of contents.

Figure 5(a) shows an example of image trimming. In the scene, the Prince exchanges bows with Snow White. Both of the Prince and Snow White are considered to be informative and should be included in the result of trimmed image. Figure 5(b) is the

iMap of the frame, and the lines indicated the trimmed area by energy optimization. The yellow regions indicate high values in the iMap. We can recognize that they are the regions corresponding to the Prince and Snow White. Moreover, the highest position is Prince's face. It demonstrates that our proposed method could estimate the important positions with the use of multiple viewers' eye movement data and SURF feature tracking.

Table 1 Experimental Results

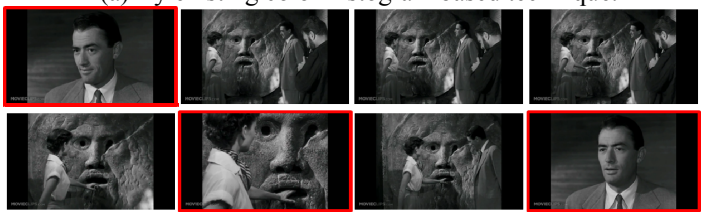
	Length	Panels	Ballons	Position error	Tail error
Pinocchio	20m 00s	352	254	10 (3.9%)	32 (12.5%)
Snow White	11m 10s	200	94	6 (6.3%)	12 (12.7%)
Roman Holiday	1m 24s	38	19	3 (15.7%)	1 (5.2%)

(a) By existing color histogram based technique.

(b) By proposed technique. Images with red frames are corresponding to the ones of (a).

Figure 3 Comparison of frame selection results for "Pinocchio."

(a) By existing color histogram based technique.



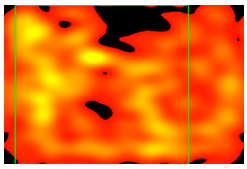
(b) By proposed technique. Images with red frames are corresponding to the ones of (a).

Figure 4 Comparison of frame selection results for "Roman Holiday."

Figure 6(a) shows a result of balloon arrangement. The blue box in Figure 6(b) is the resulting balloon position given by energy minimization. By placing the balloon at this position and directing the tail to the highest area of the iMap, the important area in the frame remains not occluded and the tail is directed to the speaker.

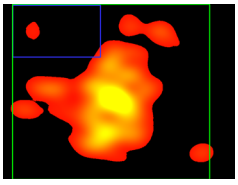
The number of position error in Table 1 is the number of balloons that partially occlude a character or an object which is related to the story development. The

number of tail error is the number of tails not heading to the speaker. Figure 7(a) and 7(b) are the examples of position error and tail error, respectively. In Figure 7(a), the balloon has been placed on the body of Wicked Queen since the surrounding areas have more SURF features than the body area. We expect to solve this kind of problem by using other higher level image features such as those for measuring objectiveness. In Figure 7(b), the tail of the balloon has a wrong direction since the speaker is not included in the scene. We expect to reduce such errors by combining eye-tracking data with some advanced motion detection techniques.

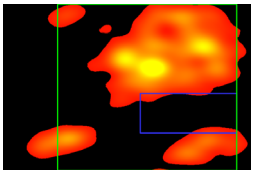
(a) Original frame image. (b) iMap corresponding to (a).

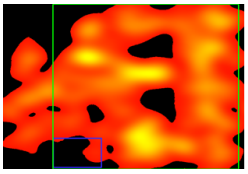
Figure 5 An example of image trimming by referring to iMap.



(a) Composited balloon. (b) Optimizing the position of a balloon on iMap.

Figure 6 Arranging a balloon by referring to iMap.



(a) Balloon arrangement failure. (b) Tail attaching failure.

Figure 7 Example of balloon arrangement and tail attaching failure.

6 Conclusions and feature works

This paper presented a novel technique for generating a film comic from a movie automatically based on the eye-tracking data of sample viewers. We believe that the concept of iMap, together with the optimization scheme for obtaining desirable image trimming and balloon arrangement, has a very high potential for being used in other image/video processing applications. Also we believe that with the advancement of eye-tracking technology as well as the cost down of eye tracking devices in the future, the proposed technique will enable casual users to easily enjoy movie in the style of comic.

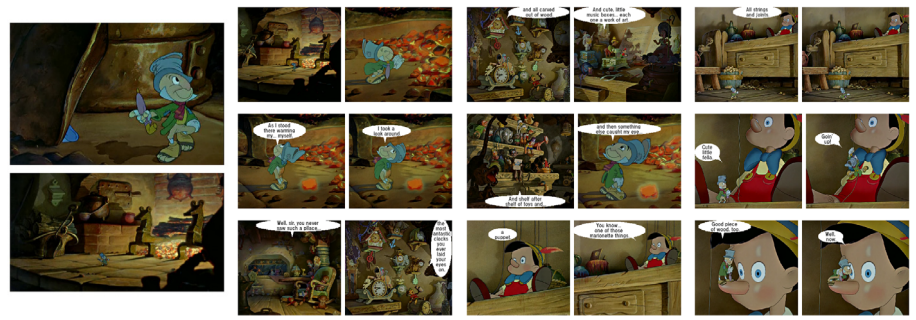
Acknowledgement

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

References

1. Hwang, W.I., Lee, P.J., Chun, B.K., Ryu, D.S., Cho, H.G.: Cinema comics: Cartoon generation from video stream. In: International Conference on Computer Graphics Theory and Applications. (2006) 299–304
2. Preuß, J., Loviscach, J.: From movie to comics, informed by the screenplay. ACM SIGGRAPH (Poster) (2007)
3. Hong, R., Yuan, X.T., Xu, M., Wang, M., Yan, S., Chua, T.S.: Movie2comics: a feast of multimedia artwork. In: Proceedings of the International Conference on Multimedia. (2010) 611–614
4. Toyoura, M., Kunihiro, M., Mao, X.: Film comic reflecting camera-works. In: International Conference on MultiMedia Modeling. (2012) 406–417
5. Smith, T.J., Whitwell, M., Lee, J.: Eye movements and pupil dilation during event perception. In: Proceedings of the Eye Tracking Research and Applications conference. (2006)
6. Ma, Y., Hua, X., Lu, L., Zhang, H.: A generic framework of user attention model and its application in video summarization. IEEE Trans. on Multimedia **7**(5) (2005) 907–919
7. Li, K., Guo, L., Faraco, C., Zhu, D., Deng, F., Zhang, T., Jiang, X., Zhang, D., Chen, H., Hu, X., Miller, S., Liuothers, T.: Human-centered attention models for video summarization. In: International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction, ACM (2010) 27
8. You, J., Liu, G., Sun, L., Li, H.: A multiple visual models based perceptive analysis framework for multilevel video summarization. IEEE Transactions on Circuits and Systems for Video Technology **17**(3) (2007) 335–342
9. Money, A.G., Agius, H.: Video summarisation: A conceptual framework and survey of the state of the art. Journal of Visual Communication and Image Representation **19**(2) (2008) 121–143
10. Porter, S.V.: Video Segmentation and Indexing using Motion Estimation. PhD thesis, University of Bristol (2004)
11. Chen, B., Wang, J., Wang, J.: A novel video summarization based on mining the story structure and semantic relations among concept entities. IEEE Transactions on Multimedia **11**(2) (2009) 295–312
12. Hu, X., Deng, F., Li, K., Zhang, T., Chen, H., Jiang, X., Lv, J., Zhu, D., Faraco, C., Zhang, D., et al.: Bridging low-level features and high-level semantics via fmri brain imaging for video classification. In: Proceedings of the international conference on Multimedia, ACM (2010) 451–460
13. Liu, A., Yang, Z.: Watching, thinking, reacting: A human-centered framework for movie content analysis. International Journal of Digital Content Technology and its Applications **4**(5) (2010) 23–37

14. Peng, W.T., Huang, W.J., Chu, W.T., Chou, C.N., Chang, W.Y., Chang, C.H., Hung, Y.P.: A user experience model for home video summarization. In: International Conference on Multimedia Modeling. (2009) 484–495
15. Boreczky, J.S., Girgensohn, A., Golovchinsky, G., Uchihashi, S.: An interactive comic book presentation for exploring video. In: ACM Conference on Computer-Human Interaction (CHI). (2000) 185–192
16. Bay, H., Ess, A., Tuytelaars, T., Gool, L.V.: Speeded-up robust features (surf). Computer Vision and Image Understanding **110**(3) (2008)



(a) Pinocchio.

(b) Show White and the Seven Dwarfs.

(c) Roman Holidays.

Figure 8 Comic pages generated with Proposed technique.