This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2015.2474263, IEEE Transactions on Multimedia

Content-Aware Video2Comics with Manga-Style Layout

Guangmei Jing, Yongtao Hu, Yanwen Guo, Yizhou Yu, *Member, IEEE,* Wenping Wang, *Member, IEEE*

***Abstract*—We introduce in this paper a new approach that conveniently converts conversational videos into comics with manga-style layout. With our approach, the manga-style layout of a comic page is achieved in a content-driven manner, and the main components, including panels and word balloons, that constitute a visually pleasing comic page are intelligently organized. Our approach extracts key frames on speakers by using a speaker detection technique such that word balloons can be placed near the corresponding speakers. We qualitatively measure the information contained in a comic page. With the initial layout automatically determined, the final comic page is obtained by maximizing such a measure and optimizing the parameters relating to the optimal display of comics. An efficient Markov chain Monte Carlo sampling algorithm is designed for the optimization. Our user study demonstrates that users much prefer our manga-style comics to purely Western style comics. Extensive experiments and comparisons against previous work also verify the effectiveness of our approach.**

***Index Terms*—Comics, layout optimization, video presentation.**

I. Introduction

COMICS, as a popular art form, are graphical media used to concisely express ideas via visual information combined with textual information. Nowadays, comics are widely used in newspapers, magazines, and graphic novels. Figure 1 shows several representative comic pages. Although from different countries, these comic pages still exhibit certain common styles, such as: 1) simple layout structure and spatial arrangement of panels, which make comics easy-to-read; 2) few complex structures, thereby enhancing visual richness; 3) non-casual placement of word balloons, meaning that word balloons should be placed on less important regions so as not to occlude salient objects. Figure 1 also shows an example of Japanese comics (manga, on the right). Different from the traditional Western comics, which are more rigid and grid-based, artists typically stylize the layout of manga by introducing some customized features, including flexible layout, variations in panel size, and irregular panel shapes [1]. These features help to enhance visual richness.

The large quantities of readily available TV series and movies provide abundant sources for the production of comics.

G. Jing, Y. Hu, Y. Yu and W. Wang are with the Department of Computer Science, The University of Hong Kong, Pokfulam Road, Hong Kong. E-mail: gmjing@cs.hku.hk, ythu@cs.hku.hk, yzyu@cs.hku.hk, wenping@cs.hku.hk.

Y. Guo is with the National Key Lab for Novel Software Technology, Nanjing University, Nanjing, Jiangsu 210023, P.R. China. He is also affiliated with the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, and State Key Laboratory of Virtual Reality Technology and Systems, Beihang University. E-mail: ywguo@nju.edu.cn.

Y. Guo was supported by the National Natural Science Foundation of China under Grants 61373059 and 61321491.

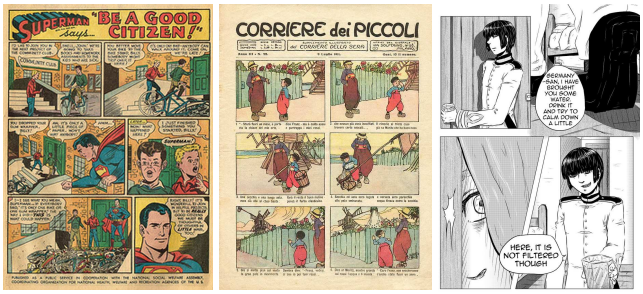


Fig. 1. Sample comic pages from different countries with different styles. From Left to Right: Superman (America), Corriere dei piccoli (Italy) and HetaOni (Japan).

The goal of this paper is to convert a video sequence, for instance an episode of a TV series or a movie, into the comics that inherit the intrinsic styles of comics, while at the same time, have the intriguing features of manga. We call such comics with flexible layout, varying panel sizes, and irregular panel shapes, manga-style comics.

Achieving this goal poses a few challenges. First, readers should be able to mentally construct events from the panels and textural information. In this sense, frames displayed on a comic page need to be carefully selected, to facilitate storytelling by including more key frames containing speakers and using word balloons. Therefore, simply detecting key frames by visual changes of frames is insufficient. Second, layout, as the geometry of the panels, needs to be designed according to video content. Previous methods employ either heuristic rules or pre-defined templates for the generation of comic layouts, thus limiting their ability to produce rich and distinctive styles [2], [3], [4]. Third, and most importantly, considering the limited size of each comic page, optimal selection, from the key frames, of the area to display and placement of word balloons are the key factors affecting the generation of an aesthetically pleasing, easy-to-read comic page. This is intrinsically a complex, combinatorial optimization problem whose solution space is huge, considering the parameters concerning layout geometry, visible content in each panel and word balloon positions.

Our approach. We introduce a framework to conveniently produce comics with manga-style layout for conversational videos, such as an episode of a TV series or a movie (see Figure 10). Our approach optimizes the comic layout and computes all the parameters relating to the display of a comic page. The approach works completely in a content-driven manner and does not rely on any comics database. By

1520-9210 (c) 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2015.2474263, IEEE Transactions on Multimedia

contrast, creation of a manga layout is achieved by the previous approach [1] in a data-driven manner with a large manga database as support. Compared with the above method, another distinct advantage of our approach is that word balloons, which are more helpful for storytelling, are preserved.

Our approach selects key frames on speakers by exploiting the subtitle file associated with the input video, in addition to detecting the key frames by visual changes. We also detect speakers with a speaker detection technique [5]. These promote the speakers, accompanied by word balloons, to be the *main characters* of the still comics, facilitating storytelling. We determine an initial page layout by analyzing conversations among speakers. The core of our system is an optimization process which maximizes the information exhibited in a comic page. Through the optimization, we obtain the key parameters relating to the display of a comic page, which include the final state of layout geometry, the visible content in each panel, and word balloon positions. An efficient Markov chain Monte Carlo sampling algorithm is specifically designed for the refined optimization. Alternatively, we can cartoonize the results using various stylization methods. We demonstrate two ways of cartoonization: color abstraction by [6] and pencil-shading stylization by [7].

Contributions. We develop a new Video2Comics framework to convert the videos that contain conversation between speakers, such as TV series and movies, to comics with a manga-style layout. We are the first to consider a content-aware approach to intelligently organize panels and word balloons together. This is realized by maximizing the information presented in a comic page and optimizing the parameters relating to the optimal display of comic pages. An efficient Markov chain Monte Carlo sampling algorithm is developed for this purpose.

II. Related Work

Video2Comics. Some related work has tackled problems similar to Video2Comics. CORVIS [8] and CINETOON [9] are two semi-automated comics cartooning systems to transform a video (a cinema film) into a comic book. Representative scenes are selected manually and several stylized comic effects, such as the speed line and background effect, are inserted. Movie screenplay is employed in [10] to aid in converting a film to a comic strip. Scene segmentation and dialogue extraction are based on the information from the screenplay. User intervention is often required in word balloon placement since their software cannot determine speaker location.

The work most related to ours is [4], which aims to turn a movie clip into a comic automatically by the integration of a variety of techniques including subshot detection, key frame extraction, face detection, and cartoonization. This method employs eight pre-defined templates of layout and tries to make the panels accommodate the extracted speaker region, without explicitly considering the problem of leaving sufficient space for word balloons. By contrast, our system determines the layout geometry in a content-driven manner, yielding more flexible panel shapes. We will make a thorough comparison with this method in our experiments.

Comic layout. Most previous methods [11], [12], [13], [14] employ either simple heuristic rules or pre-defined templates of layout, thus limiting their ability to produce rich and distinctive styles. In the most recent work, Cao et al. [1] propose a data-driven approach to generate stylistic manga layout by learning from a set of input artworks with user-specified semantics. By contrast, our content-driven method is more flexible, as we do not rely on any comics database. We jointly optimize all the parameters relating to the display of a comic page, including the positions of word balloons, which are not touched by this method.

Word balloon placement. Word balloons are the essential elements for storytelling. Kurlander et al. [2] address four types of word balloons as well as their layout and construction in their comic chat system. Chun et al. [15] propose to position each word balloon relative to its respective actor at first, then refine its position based on a measure that estimates the quality of the balloon layout. Gaze information is used in [16] for automatically finding the location for inserting and directing the word balloons. Cao et al. [17] propose an approach for novices to synthesize a composition of picture subjects and balloons on a page that can guide the reader's attention to convey the story, through a probabilistic graphical model learned from a set of manga pages. They rely on user-specified semantics as input. By comparison, our method optimizes an objective function quantifying the information embedded in a page to position word balloons automatically. Both methods can provide a continuous and fluid reading experience. Recent research has shown that the multiframe personalized content synthesis in the form of comic-strips can be made easier with a Poseshop system [18].

Video stylization. Agarwala et al. [19] describe an approach which creates cartoon animation from an input video by tracking user-specified contours. The system of [20] aims to transform an input video into a highly abstracted, spatio-temporally coherent cartoon animation with a range of styles. Shamir et al. [3] create a sequence of static comic-like images summarizing the main events in a virtual world environment and present them in a coherent, concise and visually pleasing manner. Other methods focus on converting the video into stylistic effects [21], [22].

Video summarization. Video summarization, as an important video content service, produces a condensed and succinct representation of video content, facilitating the browsing, retrieval, and storage of the original videos. There has been a rich literature on summarizing a long video into a concise representation, such as a key-frame sequence [23], a video skim [24], [25], [26], video collage [27] and visual storylines [28], [29]. We refer the readers to [30] for a comprehensive review of video summarization methods. Different from most previous work, we select the middle frame in a speaker tracklet as a key frame with the aid of speaker tracking. By doing this, the word balloon and its corresponding panel are naturally synchronized in our result.

Some recent methods focus on summarizing video in the form of a single static image named schematic storyboard [31], or a multi-scale tapestry [32] for better navigation. By contrast, we aim at producing traditional style comics which tell the

1520-9210 (c) 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2015.2474263, IEEE Transactions on Multimedia

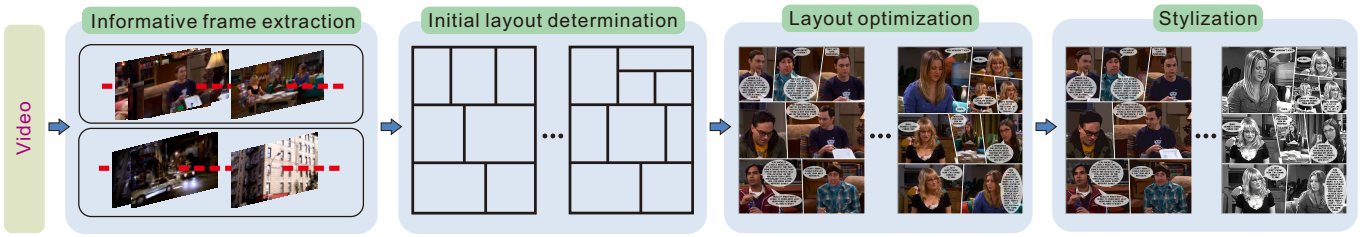


Fig. 2. The overall pipeline of our framework. We first extract the informative frames, then set up an initial layout for each comic page. The final state of a page is determined by layout optimization. The comics can be further processed to provide stylized rendering effects. From "The Big Bang Theory" (©Chuck Lorre Productions, Warner Bros. Television and CBS).

story in a lively and concise manner.

III. Our System

Our system consists of three main components: informative frame extraction, initial layout determination, and layout optimization. The speaker-key-frames and representative frames of the scenes are first extracted as informative frames. Initial layout determination then sets up geometric structure of the layout according to the video content, which is fed into the final layout optimization for determining all the parameters relating to the display of informative frames and word balloons on each comic page (see Figure 2). Although stylization is not the core of this work, our system provides two ways to stylize the comics.

A. Informative Frame Extraction

We consider TV series or movies associated with corresponding subtitle files. Time information from subtitle files always indicates the duration during which the subtitle is continuously superimposed on video frames, which could help extract key frames with speaking characters. We call them speaker-key-frames. However, only speaker-key-frames are insufficient for representing the movies' theme. Scene change caused by shot transition is also vital to the viewer's understanding of the content and comprehension of the story. In our approach, we first extract speaker-key-frames based on speaker detection, and then employ the GIST Descriptor [33] to detect scene changes.

Note that if subtitle files are not available, speech recognition techniques could be employed to generate them.

Speaker detection. To determine the speaker in a given video clip, we use a newly proposed speaker detection algorithm [5]. The procedure can be summarized as follows: 1) face tracking for all the faces by face detection and matching clothing appearances; 2) speaker detection to identify the true speaker based on features, including lip motion, center contribution, length consistency, and audio-visual synchrony. As reported, the speaker detection algorithm has proven to be robust and accurate (over 90% accuracy) for a variety of TV/movie types. Please see [5] for the details. Note that, in order to generate accurate speaker-word mapping for high-quality comic creation, we require the user to check the detection results and correct those incorrect ones manually. In our experiments, the manual effort required for each comic is less than ten percent of video frames.

Informative frame extraction. We extract informative frames by using the following strategies:

- For each subtitle, we perform speaker tracking within the time duration, and the middle frame in the speaker tracklet is chosen as a key frame within this time period. If it fails to detect the speaker, we simply use the middle frame of the duration as a key frame.
- We also extract a key frame once a shot transition is detected.
- Merge similar key frames. With the above two steps, similar key frames may exist which should be merged to reduce redundancy. We employ a two-pass scheme to merge them. In the first pass, neighboring key frames are merged based on scene similarity of the frames, which is measured by the $L2$ distance of GIST descriptor [33]. We set the distance threshold to 0.55 in our experiments. If two adjacent frames have similar visual content and close speaker positions as well, we simply retain either of them and merge their subtitles. For the case of similar visual content but quite different speaker locations, besides keeping either of them, we further map the word balloon of the speaker in the discarded frame to the retained one by associating it to the corresponding speaker's face. The retained frame is called a multi-speaker key frame. Figure 7 left shows an example of the multi-speaker key frame. The second pass merges frames for the same speaker in the case of loop structure and conversation structure as defined in the following section.

B. Initial Layout Determination

People are accustomed to reading articles line-by-line. To keep correct reading order and ease storytelling as well, most panels should be arranged in scan-line order in a page. As perceived from real comic examples, locally variant layout structures can augment visual richness and make the content more engaging. To account for this, we also design more complex local layout structures based on conversations, which help make storytelling more vivid. In our method, we first determine local structure based on video content, and then generate the initial layout of a comic page automatically.

1) Local Structure: **Loop structure.** We observe that speakers may talk in turn in a conversation and that the same speaker may appear several times within a short duration. For instance, in the movie "Les Choristes", shown by Figure 3, the countess talked to the two men again after two shots of

1520-9210 (c) 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2015.2474263, IEEE Transactions on Multimedia

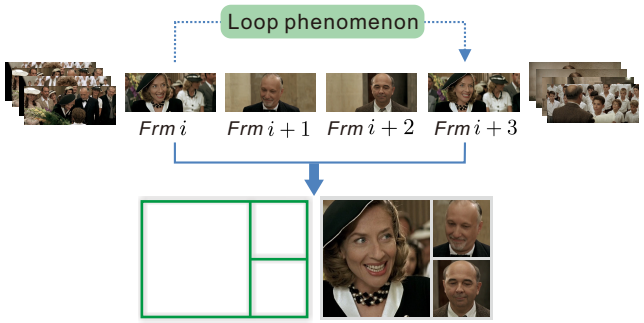


Fig. 3. Loop structure. The key frames i and $i+3$ are similar to each other, so we merge them together by reserving only one of them, and define a loop structure as shown in the bottom. From "Les Choristes" (©Pathé (UK/France) and Miramax Films (USA)).

them. This actually forms a "loop", and we call this the *loop phenomenon* during a conversation. Based on this observation, when such a loop is detected, we merge similar key frames and define local layout structure as shown in Figure 3, which we call loop structure. It would be better to include at most 4 panels in a loop structure to avoid misunderstanding of the reading order.

Conversation structure. When two speakers talk alternately as shown in Figure 4, we develop a simple conversation structure to avoid too many repetitive panels of the two speakers in the same page. Such a structure is composed of two side-by-side panels. For long conversations, we suggest that each conversation structure contains at most 6 key frames as too many word balloons in a single panel may make the reading tedious and boring.

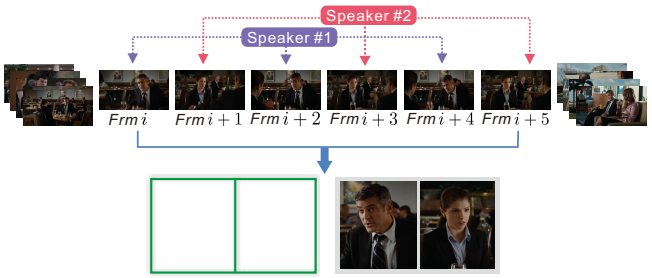


Fig. 4. Conversation structure. The key frames from i to $i+5$ are on two speakers, so we merge them in two juxtaposed panels to represent the conversation as shown in the bottom. From "Up in the Air" (©DW Studios, The Montecito Picture Company, Rickshaw Productions and Paramount Pictures).

2) Initial Layout Generation: To generate an initial layout, we first scan the key frame sequence to detect local *loop* structures. Following the usual styles of comics described earlier, we set the following rules to determine an initial page layout:

- To avoid the visual content in a panel being too small to be seen clearly on an e-book reader like Kindle or a standard $A4$ paper, we set 3 rows per comic page.
- Observed from real comic pages, the local structure will stay in the same row. A multi-speaker key frame will occupy one row, in order to completely present all the subtitles.
- For all other cases, we just randomly put two or three panels in each row.

Obeying the above rules, the initial height of each row is estimated based on the ratio of saliency of the frames to be displayed in this row to the saliency of all the frames in the current page. Similarly, the width of each panel is initially set proportional to the ratio of its saliency to the saliency of all the frames to be displayed in the same row. We generate the initial layout with a rigid and grid-based style.

C. Layout Optimization

Given the initial layout, we need to select the area from each key frame to display in its corresponding panel and to determine word balloon positions in the panel. Different from videos, which not only tell the story in a temporally continuous manner but also normally have very good resolutions, *comics* are a relatively concise medium to express ideas via static pages consisting of reduced-sized, cropped images combined with text. Considering this point, it would be more helpful if a comic page contains more information. To do so, we quantitatively measure the information embedded in a comic page and obtain the parameters relating to the display of a comic page by maximizing the measure.

In the following, we first formulate our objective function for the page layout problem. We define an energy function which quantitatively measures the information contained in the current page in terms of visual saliency values. We then solve the optimization problem by using a specifically designed Markov chain Monte Carlo sampling method.

Fig. 5. Page parameters. $L_{k|k=1,\dots,7}$ represent line parameters and $1\sim8$ denote panels for displaying video frames and word balloons. From "The Man from Earth" (©Anchor Bay Entertainment and Shoreline Entertainment).

1) Page Energy: The initial page layout and the number of frames per page N have been determined by the previous steps (see Figure 5). Given the input frames $\{I_i\}_{i=1}^{N}$ and their word balloons, our goal is to arrange those already selected frames and word balloons on a comic page in an optimal manner, based on the initial page layout.

We have the following variables: the parametric coordinates of line segments of the layout, which partition the comic page into panels, denoted as $\{L_{k|k=1,\dots,K}\}$ with K the number of lines; the scaling factor of each frame s_i when it is mapped

1520-9210 (c) 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2015.2474263, IEEE Transactions on Multimedia

onto the panel i; and the position of word balloons denoted as $\{p_{Wb_i}\}$. For simplicity, we use x to denote the set of state variables: $x = \{\{L_{k|k=1,\ldots,K}\}, \{s_i\}, \{p_{Wb_i}\}\}$.

In our method, visual saliency is used to indicate the pixel-wise importance of each frame. An example of the visual saliency map is shown in Figure 6 (Right) which indicates the saliency value for each image pixel. In this map, white indicates higher saliency values while black indicates lower. The information is thus measured by the sum of saliency values. To maximize the information presented on a comic page, our energy E is defined as follows:

$$E(x) = \sum_{i=0}^{N-1} (f(s_i) \times E_{Frm_i}(x) - E_{B_i}(x) + E_{Wb_i}(x)), \quad (1)$$

where the first term, E_{Frm_i}, represents the information quantified as the sum of visual saliency values of pixels contained in the area to display. The second term, E_{B_i}, is the sum of saliency values of pixels in the area occluded by those word balloons in panel i. It is subtracted because of invisibility due to occlusion. The third term, E_{Wb_i}, denotes the information of word balloons in panel i, which is measured by the sum of given importance values of points in its bounding boxes. $f(s_i)$ is a function of s_i for preventing the area selected from the original frame from shrinking too much when it is mapped onto the panel. Adjusting $f(s_i)$ can affect the scale of visual content presented in the comic page. We simply define $f(s_i) = s_i$ in our experiments.

Saliency representation. For each key frame, we compute its saliency map via a global contrast based saliency detection method [34]. For speaker-key-frames, we set the highest importance value (1.0) to the face region since the face is the most important area in our scenario. Based on the face position, we further estimate the body rectangle, which is fed into GrabCut segmentation [35] for extracting an approximate body area. We believe that the body is of secondary importance, and so it should also be given a relatively high importance value (see Figure 6). We set it to 0.9 in our implementation. As text is also important in conveying the story for a comic page, we assign the highest importance value (1.0) to the bounding box of each word balloon. As a result, E_{Wb_i} is always greater than or equal to E_{B_i}, which guarantees that the total energy E in Eq (1) is always positive.

Recall that we extract informative key frames by using speaker detection and shot transition detection, without resorting to saliency detection. We, however, measure the information contained in a comic page in terms of visual saliency values. This is because we would like to display more important information, rather than the dull background, given the very limited panel size.

2) Optimization: Our goal is to maximize the energy $E(x)$ defined by Eq (1):

$$x^* = \arg\max_x E(x). \quad (2)$$

The objective function is a high dimensional, non-convex combinatorial optimization problem which is difficult to solve analytically. In statistics, Markov chain Monte Carlo (MCMC) methods [36] are generally used for sampling from multi-dimensional distributions, especially when the number of dimensions is high, and are based on constructing a Markov chain that has the desired distribution as its equilibrium distribution. We propose a MCMC sampling algorithm specifically designed for our optimization.

Fig. 6. Our saliency model. Left: the body rectangle loosely around the person. Middle: segmentation result of GrabCut [35]. Right: our saliency map. In our experiments, the face rectangle is assigned with the highest saliency value which is 1.0, while each pixel in the body area is set to 0.9. From "Friends" (©Bright/Kauffman/Crane Productions, Warner Bros. Television, NBC and Warner Bros. Television Distribution (worldwide)).

Markov chain Monte Carlo. Given a distribution $\pi(x)$ of variables x, which represents the set of state variables $\{\{L_{k|k=1,\ldots,K}\}, \{s_i\}, \{p_{Wb_i}\}\}$ determining a comic page in our case, MCMC is an approach to generate samples $\{x^t\}_{t=1}^T$ from the probability distribution by constructing a Markov chain. Most MCMC methods are based on the Metropolis Hastings (MH) algorithm [37]. In MH sampling, the proposal function $Q(x^\star|x^t)$, also called the transition function, can be arbitrary which is used to generate a candidate state $x^\star$ given the current state x^t. The MH algorithm runs iteratively and it essentially works as follows [36]:

- Draw y from the proposal function $Q(y|x^t)$.
- Draw $U \sim$ Uniform $\mathcal{U}(0,1)$ and update

$$x^{(t+1)} = \begin{cases} y & \text{if } U \leq r(x^{(t)}, y) \\ x^{(t)} & \text{otherwise} \end{cases} \quad (3)$$

where $r(x,y)$ is called the acceptance ratio and is defined as

$$r(x,y) = \min\left\{1, \frac{\pi(y)Q(y|x)}{\pi(x)Q(x|y)}\right\}. \quad (4)$$

For a high dimensional, non-convex optimization problem, multiple local optima may exist. To avoid getting stuck at local minima, we design a mixture of proposals to solve our problem: a local proposal Q_l that locally explores the state space and a global proposal Q_g that helps to jump out of the local minimum. *Proposal* here means the suggested parameters given the current distributions of parameters. The local and global proposals are defined separately as:

$$Q(x^\star|x^t) = w_l Q_l(x^\star|x^t) + w_g Q_g(x^\star|x^t) \quad (5)$$

where w_l and w_g are two dynamically adjusted weights with $w_l + w_g = 1$.

a) Local proposal: The local proposal only changes one parameter every time. Since x is used to represent the set of state variables $\{\{L_{k|k=1,\ldots,K}\}, \{s_i\}, \{p_{Wb_i}\}\}$, which correspond to line segments, scales, and positions of word balloons separately, we randomly select one of the line segment, scale, and word balloon proposals.

1520-9210 (c) 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2015.2474263, IEEE Transactions on Multimedia

Line segment proposal: The transition kernel of each end point of a line segment is defined as a Gaussian distribution $N(d_l; 0, \sigma_l)$, where d_l is the distance from its current position and σ_l is set proportional to the width of a page. The ratio is 0.05 in our experiments. At each time step, we update only one line segment of the layout by employing the above MH algorithm to generate a new sample.

Scale proposal: This proposal is designed for controlling the scale factor when a key frame is mapped onto its corresponding panel. If the face of a speaker is relatively small compared with the image size, it would be more reasonable to select a small scale since the frame probably contains more background. This will facilitate the understanding of the context of the conversation.

Denote the area of the comic page, the key frame to display, and the bounding box of a speaker's face as A_P, A_{Fri}, and A_F, respectively. We model the scale, s_i, of panel i as a uniform distribution:

$$\begin{cases} \mathcal{U}(s_{\min}, \dfrac{s_{\min}+s_{\max}}{2}), & \text{if } \dfrac{A_F}{A_{Fri}} < \mathcal{T}_A \\ \mathcal{U}(s_{\min}, s_{\max}), & \text{otherwise} \end{cases} \tag{6}$$

where $\mathcal{T}_A$ is a threshold set to 0.3 in our implementation. $s_{\min}$ and $s_{\max}$ are the minimum and maximum scales of an image. They are set as

$$\begin{cases} s_{\min} = \dfrac{A_P}{5A_{Fri}} \\ s_{\max} = \min\{1, s_{\min} + 0.5\}. \end{cases} \tag{7}$$

Word balloon proposal: Each word balloon should be placed at a position where it occludes the least information, in terms of visual saliency, while being close but not overlapping with the speaker's face. So its state is modeled as a Gaussian distribution $N(d_r; 0, \sigma_r)$ where d_r is the distance between the center of the word balloon and the center of the speaker's face and σ_r is set equal to the width of the rectangle about the speaker's face.

Fig. 7. Voronoi diagram for candidate word balloon positions in a multi-speaker key frame. Left: word balloon positions of the two speakers. Right: Voronoi diagram to constrain balloon positions of the two speakers. From "Les Choristes" (©Pathé (UK/France) and Miramax Films (USA)).

For multi-speaker key frames, word balloon positions should be close to the corresponding speakers. The sampling space of each balloon position is confined to the cell of a Voronoi diagram (see Figure 7) whose seeds are the centers of the rectangle about the speaker's face. Furthermore, to keep a correct reading order, the word balloon position of the preceding speaker should be higher than those balloon positions of the following speakers. This is imposed as a hard sampling constraint.

Reading order. Under normal circumstances, the reader can read a comic page in scan-line order. Special reading

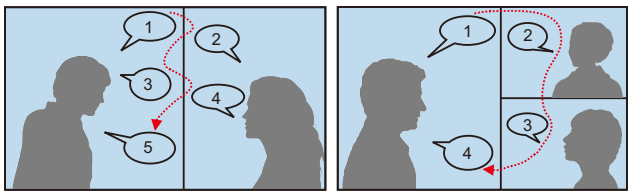


Fig. 8. Reading order of word balloons. Left: reading order in a conversation structure showing the conversation between two speakers. Right: reading order in a loop structure.

rules exist for the cases of local structures as shown by Figure 8. More specifically, for the panel with multiple word balloons, the balloons from top to bottom are consistent with the chronological order of their corresponding subtitles in the movie. For conversations between two speakers, the reader should start reading the balloons from the left, and alternate between the left and right panels. For the loop structure, the reading order forms a loop as indicated by Figure 8. The sampled positions of word balloons should obey the above rules, and they are posed as hard sampling constraints as well. That also means that when sampling word balloon positions, we follow the above rules of reading order.

b) Global proposal: To make a sample jump away from the local minimum, we sample the parameter set independent of the current state. Each line, scale factor, and word balloon position are separately sampled from the corresponding distributions defined in the three local proposals described above.

c) Dynamic weighting: The two weights w_l and w_g represent our expectation of the frequencies of the local and global proposals being utilized. When the local proposal cannot improve the result after a certain number of iterations, the global proposal should have a larger probability of being used. Similar to [38], we set $w_l = \exp(-\frac{n^2}{2\sigma_n^2})$, where n is the iteration number that the local proposal does not improve the result continuously. σ_n controls the probability that the local proposal is used and is set to $5N$, with N being the number of key frames displayed in the current comic page.

In the above MCMC algorithm, given the current state $x^{(t)}$, the proposal function consisting of the local and global proposals is used to generate a candidate state $x^{(t+1)}$ for the next iteration. The optimization process works in an iterative manner, and it terminates when the energy measured by Eq (1) remains stable after a certain number of iterations.

D. Stylization

Stylization of photographs has become a tool for effective visual communication. It is also a vital part of comic generation. Although stylization is not the core of this work, our system provides two ways to stylize the comics we produce. Specifically, we generate the abstraction results with simplified color illustrations by [6] and black-white, pencil-shading effects by [7]. Other stylization techniques still apply to our system.

1520-9210 (c) 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2015.2474263, IEEE Transactions on Multimedia

Fig. 9. From left to right: initialization; comic page after 50 iterations; the final result after 200 iterations; changing curve of the minus log of energy. From "The Big Bang Theory" (©Chuck Lorre Productions, Warner Bros. Television and CBS).

Fig. 10. Several comic pages generated from an episode of "The Big Bang Theory" (©Chuck Lorre Productions, Warner Bros. Television and CBS).

IV. Experiments and Discussion

A. Performance

We first discuss the runtime performance of layout optimization, which is the core of our method. We have implemented our method on an Intel® Core™ i7 CPU @ 3.40GHz computer with 16G RAM. To show the efficiency, we present an example to show the changing curve of the energy (minus log of the information in a comic page we measure using Eq (1) in Figure 9, where the x-axis represents the iteration number and the y-axis is the lowest energy up to the current iteration. In addition to the final comic page, we also show the initialized comic page and the page after 50 iterations. Actually, the reduction in energy implies an improvement of the layout. In this example, the input is a 720p video, meaning that each frame is of size 1280×720. The comic page is of size 800×1066 with a 3:4 aspect ratio, an aspect ratio widely adopted by most prevalent eBooks such as Kindle and iPad. It takes less than 2 minutes to generate this comic page with 7 panels, excluding the time for video pre-processing, i.e., informative frame extraction and speaker detection.

Our current system is designed for offline use now. It normally spends a couple of hours to generate a comic book for an episode. This is acceptable to create a book that can be printed for publication. On the other hand, considering that the optimization of each comic page is separable, we could further parallelize the optimization of all comic pages constituting a comic book for acceleration.

B. Our Results

We conduct experiments on several video clips that are extracted from five movies: "Up in the Air," "Les Choristes," "The Man from Earth," "Titanic," and "The Message," and two TV series: "Friends" and "The Big Bang Theory." The videos are associated with subtitle files. For a normal video, for instance a 22-minute episode of "Friends," about 70 comic pages will be generated.

Figure 10 and Figure 11 show several stylized comic pages generated by our system. Overall, the relatively simple layout together with non-casual placement of word balloons make the comics easy-to-read. Moreover, the irregular panel shapes accompanied with a few complex local structures enhance visual richness of the comic pages. Word balloons, with automatically computed positions, are placed in less important background regions. Even in the worst case, when the character occupies almost the whole panel, the word balloon is positioned at his (her) body without occluding the speaker's face. Figure 10 shows three comic pages generated from the "The Big Bang Theory" with irregular panel shapes. The

1520-9210 (c) 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2015.2474263, IEEE Transactions on Multimedia

Fig. 11. Comic pages generated from "The Man from Earth" (©Anchor Bay Entertainment and Shoreline Entertainment). Up: the original comics. Bottom: the black-white, pencil-shading effects.

second row of the first page contains a local structure, which is determined by the conversations among Sheldon, Howard and Leonard. They were talking about taking vacations. Sheldon first said that "You must take a vacation...." Then Howard responded "I don't think..." and Leonard said "Sheldon, everybody takes vacations." Sheldon continued, "One time...." Note that Sheldon appeared twice during this conversation. We therefore use local structure, as defined previously, to show this. The other two pages of this example also contain similar local layout structures.

It is noted that, for all the results presented in this paper, the reading order of panels in a comic page is from left to right and top to bottom, except for the more complex local structures defined in Figure 8. Please see our accompanying video and supplemental material for additional results.

C. Comparisons

We compare our method against Movie2Comics [4]. All the videos aforementioned are used for comparisons. Figure 12 shows several comic pages produced by our method and Movie2Comics.

Movie2Comics produces the traditional Western comics whose layout is more rigid and grid-based. Our method, by comparison, takes some features from manga in terms of variations in panel size and irregular panel shapes, making the comics more visually appealing and interesting.

In Movie2Comics, the layout of each comic page is chosen from eight pre-defined templates by searching for the best match between the sub-sequence cropped from the keyframe sequence and each of the templates. In each panel, a word balloon is positioned on the right, top, or left of a speaker's face. In some panels, especially when the speaker's face occupies the full panel, word balloons may overlap with the speaker's face, as can be seen from Figure 12(e)(g)(h). Our proposed method, however, effectively avoids this shortcoming as we optimize the parameters of balloon positions and display area for each keyframe in a unified framework.

In addition, Movie2Comics favors the selection of more repetitive frames for neighboring panels, especially when two speakers talk alternatively. This is due to the fact that they select key frames based on extracted subshots. Static subshots appear frequently during a conversation. This, as a result, leads to more repetitive frames, as in Figure 12(e). Furthermore, since their key frames are extracted based on different types of subshots, rather than a speaker tracklet, as we have used, the risk that some word balloons cannot find their corresponding speakers due to unpleasant extracted keyframes is increased. Figure 12(f) is such an example.

It is noted that if speaker detection fails to detect a speaker, both our method and Movie2Comics position the word balloon indicated by a rectangle at the top left corner of a panel.

1520-9210 (c) 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2015.2474263, IEEE Transactions on Multimedia

(a) Up in the Air *(b) Friends* *(c) Titanic* *(d) The Message*

(e) Up in the Air *(f) Friends* *(g) Titanic* *(h) The Message*

Fig. 12. Comparisons against Movie2Comics [4]. Up: our results. Bottom: the results by Movie2Comics. "Up in the Air" (©DW Studios, The Montecito Picture Company, Rickshaw Productions and Paramount Pictures), "Friends" (©Bright/Kauffman/Crane Productions, Warner Bros. Television, NBC and Warner Bros. Television Distribution (worldwide)), "Titanic (1997)" (©20th Century Fox, Paramount Pictures and Lightstorm Entertainment) and "The Message" (©Huayi Brothers).

D. User Study

We also conducted a user study that compares the performance of our approach to Movie2Comics [4]. A web interface was designed for the study. We recruited 60 subjects whose ages range from 20 to 30 for this study. The comics produced by our method and Movie2Comics from the video clips in our experiments are used for the study.

Study details. We evaluate the methods in two aspects. The first is content comprehension, which measures how well the comics convey the story. Another is visual perception in terms of naturalness and enjoyment. Each subject was first asked to fill out a form on our web site about his/her profile, including age, his/her preference on the specific comics type, and whether or not he/she reads comics very often or sometimes. Then for each of the video clips, the subject was asked to watch the original video first, and read the comics generated by the two methods. We randomly change the presenting order of the two comics in our user study. The subject finally responded to the following questions.

1) To what extent do you think the comics convey the story?
2) How easy is it for you to follow the story/conversation?
3) How satisfied are you with the visual content presented in the comics?
4) How satisfied are you with the positions of word balloons?
5) To what extent do you think the presentation style is natural and enjoyable?
6) Which one do you prefer?

All the above questions were required to be assigned a score from 1 to 5, except the last one. Here 1 indicates the worst experience while 5 means the best.

Results and discussion. We first examine the overall performance of both methods. Furthermore, to avoid the results being biased by the cultural background of the users (who tend to prefer manga-style over western-style regardless of the method used to make either one), we further take the users' profiles into consideration and discuss the results accordingly.

1) Overall Performance: Figure 13 shows the results of user study. In general, the majority of subjects showed a significant preference towards our results, as seen from the average score on each of the 6 questions in Figure 13(h).

In terms of content comprehension (Q1-Q2), the comics generated by our approach convey the story better than Movie2Comics. As for visual perception (Q3-Q5), most subjects believed that our comics are easier to read and were more satisfied with the visual content presented and placement of word balloons in our comics than those of Movie2Comics. Besides, most subjects agreed that our style is more natural and enjoyable.

We try to further analyze the results of user study statistically. Using a paired-sample, two-tailed t-test, we found that, for each of Q1-Q5, there is a statistically significant difference in subjects choosing our method over Movie2Comics (all p-value $\ll 0.001$). This is expected, as our method uses a content-

1520-9210 (c) 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2015.2474263, IEEE Transactions on Multimedia

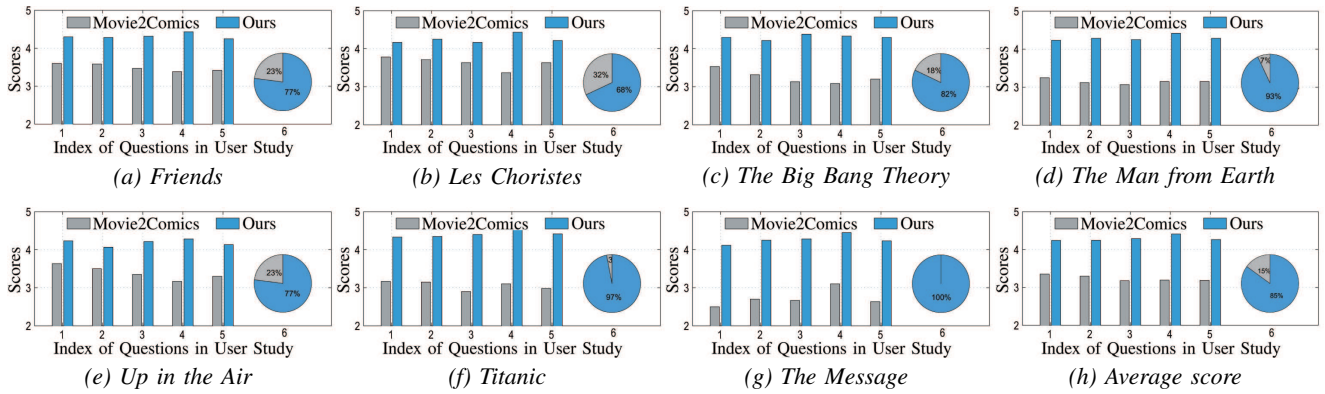


Fig. 13. User study results. Scores of the 6 questions for different movie/TV clips are shown by (a)-(g) and (h) illustrates the average score on the questions.

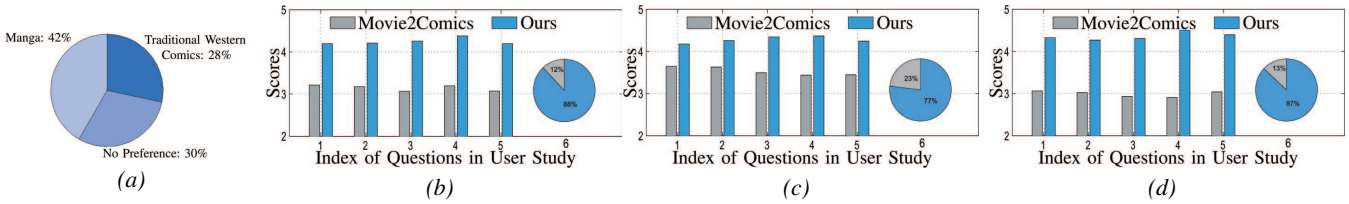


Fig. 14. (a) Information about users' preference; (b) (c) and (d): the average score on each of the 6 questions for the users who prefer manga, traditional western comics and no preference, respectively.

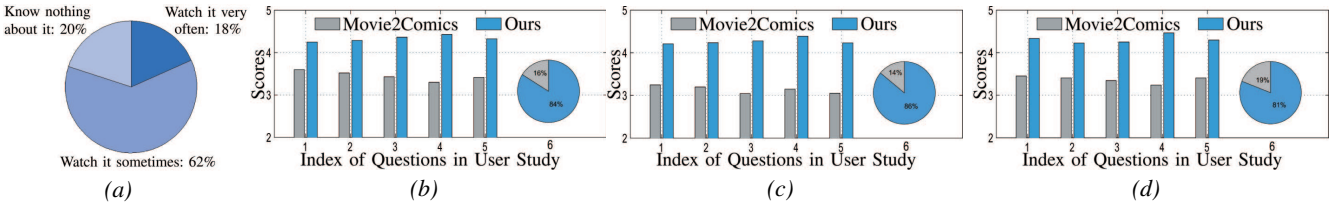


Fig. 15. (a) Information about users' familiarity with comics; (b) (c) and (d): the average score on each of the 6 questions for the users who watch it very often, who watch it sometimes and who know nothing about it, respectively.

aware approach to intelligently organize panels and word balloons together, and thus has the ability to present better visual content (Q3), find more proper positions to place word balloons (Q4), and all together make the comics presentation more natural and enjoyable (Q5).

2) Users' Background Analysis: To better validate the performance, we further analyze the responses of different types of users from the following two aspects.

Firstly, users' personal preferences on comics may affect their judgement. To avoid bias, we evaluate the results from the perspective of participants' preference on comics. As shown in Figure 14(a), we have three types of users: those who prefer manga, those who prefer traditional western comics, and those without any preference. The average scores of the 6 questions are summarized in Figure 14(b)-(d). Overall, our comics score higher than those of Movie2Comics on each of the 6 questions, for all the three types of users. The paired-sample, two-tailed t-test demonstrates that, there are statistically significant differences in Q1-Q5 (all p-value $\ll 0.001$) for each group. It should be admitted that users' preference indeed influences their choice of the comics. This can be seen from the pie charts of Figure 14 that 88% of the users who prefer manga finally select our comics with manga-style layout, while it is only 77% for the users who prefer traditional western comics. As for Q1-Q5, the advantage of our approach over Movie2Comics for the users who prefer traditional western comics is not remarkable, compared with the other two user groups.

Secondly, users' familiarity with comics may also affect their responses. As shown by Figure 15, we analyze the results of three types of users: those who read comics very often, those who read comics sometimes and those who know nothing about comics. Again, our method performs better in each group of users. For Q1-Q5 of all the three types, there are statistically significant differences in the subjects choosing our method over Movie2Comics (all p-value $\ll 0.001$).

For the detailed scores given by each type of users on different video clips, we refer the reader to our supplemental material.

3) Feedback: In addition, the feedback from some subjects show that more details should be provided in the comics we produced, such as the context of conversation. As our approach tries to mimic real comics which focus more on speakers and their conversations, it pays less attention to non-speakers. This is considered as a limitation of our current implementation, and can be conquered by modifying our strategy on key frame merging.

1520-9210 (c) 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2015.2474263, IEEE Transactions on Multimedia

E. Limitations

Our approach works only for conversational videos, such as TV series or movies. Subtitles are generally needed to facilitate informative frame extraction, otherwise we have to resort to speech recognition.

On the other hand, our framework relies on speaker detection to identify speakers. Although the speaker detection algorithm we used has proven to be robust and accurate (over 90% accuracy) for a variety of TV/movie types, manual efforts are still needed to check the detection results in order to guarantee the quality of the comics. This is another limitation of our approach.

V. Conclusion and Future Work

We have presented a new approach that conveniently converts a video sequence with conversation between speakers into comics with manga-style layout. Our approach computes a set of parameters concerning layout geometry, visual content in each panel, and word balloon placement, relating to the display of a comic page. Except for user assistance for correcting the errors of speaker detection, our approach works in a content-aware manner and does not require any further user interaction. Experiments, comparisons, as well as a user study demonstrate the effectiveness of our approach.

In this work, we propose the ability to jointly optimize the visual content and word balloon placement in a unified framework. Our framework is specifically designed for videos with subtitle files. Thus it is not applicable to those videos without providing subtitles, such as legacy TV series. This is a limitation of our method. Besides, our user study has shown that more contextual information may be helpful for the users to better understand the story. To tackle this problem, we would like to embody more background information by modifying our key-frame merging strategy in future.

Acknowledgment

The authors would like to thank the anonymous reviewers for their positive and constructive comments. We also thank Benjamin Wilson Chidester and Michael Cannon Lowney for helping proofread the whole paper.

References

[1] Y. Cao, A. B. Chan, and R. W. Lau, "Automatic stylistic manga layout," *ACM Transactions on Graphics*, vol. 31, no. 6, p. 141, 2012.

[2] D. Kurlander, T. Skelly, and D. Salesin, "Comic chat," in *Siggraph*, 1996, pp. 225–236.

[3] A. Shamir, M. Rubinstein, and T. Levinboim, "Generating comics from 3d interactive computer graphics," *Computer Graphics and Applications, IEEE*, vol. 26, no. 3, pp. 53–61, 2006.

[4] M. Wang, R. Hong, X.-T. Yuan, S. Yan, and T.-S. Chua, "Movie2comics: Towards a lively video content presentation," *IEEE Transactions on Multimedia*, vol. 14, no. 3, pp. 858–870, 2012.

[5] Y. Hu, J. Kautz, Y. Yu, and W. Wang, "Speaker-following video subtitles," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 11, no. 2, p. 32, 2014.

[6] J. Kyprianidis and J. Döllner, "Real-time image abstraction by directed filtering," *ShaderX7–Advanced Rendering Techniques, Charles River Media*, 2009.

[7] H. Winnemöller, J. E. Kyprianidis, and S. C. Olsen, "Xdog: an extended difference-of-gaussians compendium including advanced image stylization," *Computers & Graphics*, vol. 36, no. 6, pp. 740–753, 2012.

[8] W.-I. Hwang, P.-J. Lee, B.-K. Chun, D.-S. Ryu, and H.-G. Cho, "Cinema comics: Cartoon generation from video stream." in *GRAPP*, 2006, pp. 299–304.

[9] D.-S. Ryu, S.-H. Park, J.-w. Lee, D.-H. Lee, and H.-G. Cho, "Cinetoon: A semi-automated system for rendering black/white comic books from video streams," in *Computer and Information Technology Workshops, 2008. IEEE 8th International Conference on*. IEEE, 2008, pp. 336–341.

[10] J. Preu and J. Loviscach, "From movie to comic, informed by the screenplay," in *ACM SIGGRAPH 2007 posters*. ACM, 2007, p. 99.

[11] S. Uchihashi, J. Foote, A. Girgensohn, and J. Boreczky, "Video manga: generating semantically meaningful video summaries," in *ACM Multimedia*. ACM, 1999, pp. 383–392.

[12] J. Boreczky, A. Girgensohn, G. Golovchinsky, and S. Uchihashi, "An interactive comic book presentation for exploring video," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2000, pp. 185–192.

[13] J. Calic, D. P. Gibson, and N. W. Campbell, "Efficient layout of comic-like video summaries," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 7, pp. 931–936, 2007.

[14] L. Herranz, J. Calic, J. M. Martínez, and M. Mrak, "Scalable comic-like video summaries and layout disturbance," *IEEE Transactions on Multimedia*, vol. 14, no. 4, pp. 1290–1297, 2012.

[15] B.-K. Chun, D.-S. Ryu, W.-I. Hwang, and H.-G. Cho, "An automated procedure for word balloon placement in cinema comics," in *Advances in Visual Computing*. Springer, 2006, pp. 576–585.

[16] M. Toyoura, T. Sawada, M. Kunihiro, and X. Mao, "Using eye-tracking data for automatic film comic creation," in *Proceedings of the Symposium on Eye Tracking Research and Applications*. ACM, 2012, pp. 373–376.

[17] Y. Cao, R. Lau, and A. B. Chan, "Look over here: Attention-directing composition of manga elements," *ACM Transactions on Graphics (Proc. of SIGGRAPH 2014)*, vol. 33, 2014.

[18] T. Chen, P. Tan, L.-Q. Ma, M.-M. Cheng, A. Shamir, and S.-M. Hu, "Poseshop: human image database construction and personalized content synthesis," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 19, no. 5, pp. 824–837, 2013.

[19] A. Agarwala, A. Hertzmann, D. H. Salesin, and S. M. Seitz, "Keyframe-based tracking for rotoscoping and animation," *ACM Transactions on Graphics*, vol. 23, no. 3, pp. 584–591, 2004.

[20] J. Wang, Y. Xu, H.-Y. Shum, and M. F. Cohen, "Video tooning," in *ACM Transactions on Graphics*, vol. 23, no. 3. ACM, 2004, pp. 574–583.

[21] H. Winnemöller, S. C. Olsen, and B. Gooch, "Real-time video abstraction," *ACM Transactions On Graphics*, vol. 25, no. 3, pp. 1221–1226, 2006.

[22] A. Bousseau, F. Neyret, J. Thollot, and D. Salesin, "Video watercolorization using bidirectional texture advection," *ACM Transactions on Graphics*, vol. 26, no. 3, p. 104, 2007.

[23] A. Hanjalic and H. Zhang, "An integrated scheme for automated video abstraction based on unsupervised cluster-validity analysis," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, no. 8, pp. 1280–1289, 1999.

[24] Y.-F. Ma, X.-S. Hua, L. Lu, and H.-J. Zhang, "A generic framework of user attention model and its application in video summarization," *IEEE Transactions on Multimedia*, vol. 7, no. 5, pp. 907–919, 2005.

[25] C.-W. Ngo, Y.-F. Ma, and H.-J. Zhang, "Video summarization and scene detection by graph modeling," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 2, pp. 296–305, 2005.

[26] Y. Fu, Y. Guo, Y. Zhu, F. Liu, C. Song, and Z.-H. Zhou, "Multi-view video summarization," *IEEE Transactions on Multimedia*, vol. 12, no. 7, pp. 717–729, 2010.

[27] T. Wang, T. Mei, X.-S. Hua, X.-L. Liu, and H.-Q. Zhou, "Video collage: A novel presentation of video sequence," in *Multimedia and Expo, 2007 IEEE International Conference on*. IEEE, 2007, pp. 1479–1482.

[28] T. Chen, A. Lu, and S.-M. Hu, "Visual storylines: Semantic visualization of movie sequence," *Computers & Graphics*, vol. 36, no. 4, pp. 241–249, 2012.

[29] Y. Hu, J. Ren, J. Dai, C. Yuan, L. Xu, and W. Wang, "Deep Multimodal Speaker Naming," in *Proceedings of the 23rd annual ACM international conference on Multimedia*. ACM, 2015.

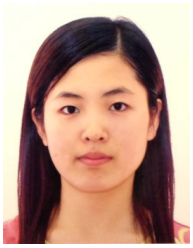
[30] B. T. Truong and S. Venkatesh, "Video abstraction: A systematic review and classification," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 3, no. 1, pp. 1–37, 2007.

[31] D. B. Goldman, B. Curless, D. Salesin, and S. M. Seitz, "Schematic storyboarding for video visualization and editing," in *ACM Transactions on Graphics*, vol. 25, no. 3. ACM, 2006, pp. 862–871.

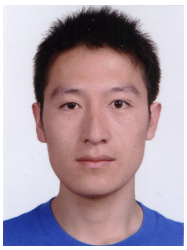
1520-9210 (c) 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2015.2474263, IEEE Transactions on Multimedia

[32] C. Barnes, D. B. Goldman, E. Shechtman, and A. Finkelstein, "Video tapestries with continuous temporal zoom," *ACM Transactions on Graphics*, vol. 29, no. 4, p. 89, 2010.
[33] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.
[34] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, "Global contrast based salient region detection," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 409–416.
[35] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," in *ACM Transactions on Graphics*, vol. 23, no. 3. ACM, 2004, pp. 309–314.
[36] J. S. Liu, *Monte Carlo strategies in scientific computing*. springer, 2008.
[37] W. K. Hastings, "Monte carlo sampling methods using markov chains and their applications," *Biometrika*, vol. 57, no. 1, pp. 97–109, 1970.
[38] J. Wang, L. Quan, J. Sun, X. Tang, and H.-Y. Shum, "Picture collage," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 1. IEEE, 2006, pp. 347–354.



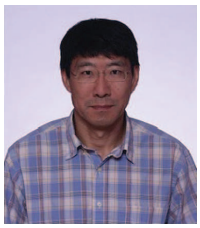
Guangmei Jing received her B.Eng degree from University of Science and Technology of China in 2010. She is currently a Ph.D. candidate in the Department of Computer Science, The University of Hong Kong. She worked as a research assistant in State Key Laboratory of CAD&CG, Zhejiang University, Hangzhou, China in 2010. Her research interests include image/video processing and computer vision.



Yongtao Hu received his B.Eng degree from Shandong University in 2010. He is currently a Ph.D. candidate in Department of Computer Science, The University of Hong Kong. He worked as a research intern at Internet Graphics Group in Microsoft Research Asia in 2010 and researcher assistant at Image & Visual Computing Lab (IVCL) in Lenovo Research & Technology, Hong Kong in 2014. His research interests include image/video processing/analysis and computer vision.

Yanwen Guo received the Ph.D. degree in applied mathematics from the State Key Lab of CAD&CG, Zhejiang University, China, in 2006. He is currently an associate professor at the National Key Lab for Novel Software Technology, Department of Computer Science and Technology, Nanjing University, Jiangsu, China. He worked as a visiting professor in the Department of Computer Science and Engineering, The Chinese University of Hong Kong, in 2006 and 2009, respectively, and the Department of Computer Science, The University of Hong Kong, in 2008, 2012, and 2013, respectively. He has been a visiting scholar in the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, since 2013. His research interests include image and video processing, vision, and computer graphics. He is the corresponding author of this paper.

Yizhou Yu received the PhD degree from the University of California at Berkeley in 2000. He is currently a full professor in the Department of Computer Science, The University of Hong Kong, and an adjunct professor at the University of Illinois, Urbana-Champaign. He received the 2002 National Science Foundation CAREER Award and the Best Paper Award at 2005 and 2011 ACM SIGGRAPH/EG Symposium on Computer Animation. He is on the editorial board of Computer Graphics Forum and International Journal of Software and Informatics. He is the program chair of Pacific Graphics 2009, Computer Animation and Social Agents 2012, and has served on the program committee of many leading international conferences, including SIGGRAPH, SIGGRAPH Asia, and International Conference on Computer Vision. His current research interests include data-driven methods for computer graphics and vision, digital geometry processing, video analytics, and biomedical data analysis.



Wenping Wang received the Ph.D. degree from the University of Alberta, Edmonton, Canada. He is a professor and the department head of the Department of Computer Science, The University of Hong Kong. His research interests include computer graphics, visualization, and geometric computing. His current research interests include mesh generation and surface modeling for architectural design. He is a journal associate editor of Computer Aided Geometric Design, Computers and Graphics, and IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, and the program cochair of several international conferences, including Pacific Graphics 2003, ACM Symposium on Physical and Solid Modeling (SPM '06), Conference on Shape Modeling (SMI '09), and the conference chair of Pacific Graphics 2012 and SIGGRAPH Asia 2013. He is a member of the IEEE.

1520-9210 (c) 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.