Content Analysis for Home Videos

Naoko Nitta (member)†, Noboru Babaguchi (member)†

Abstract The popularity of hand-held video camcorders has increased the amount of poor-quality home videos captured by amateur camcorder users. This paper introduces the content analysis techniques, namely, techniques for segmentation, indexing, and static and dynamic representation generation, which have been developed to help viewers watch such poor-quality videos by considering the characteristics of home videos.

Key words: Home Videos, Segmentation, Indexing, Static Representation, Dynamic Representation

1. Introduction

The popularity of consumer hand-held video camcorders has dramatically increased the quantity of home videos which record personal events such as vacations, travels, weddings, and graduation ceremonies. However, people rarely watch the recorded videos afterward since they often contain many long uninteresting video segments. Further, the videos recorded by amateur camcorder users are often of poor quality resulting from lack of professional capture skills. In order to increase the value of home videos, two types of solutions have been mainly proposed. One solution is embedding computer assistance at the time of capture which supports camcorder users in capturing the videos of good quality in the first place.

The other solution is embedding computer assistance after capture, which supports viewers in watching the videos already captured by amateur camcorder users with video content analysis techniques. This paper introduces the existing techniques especially focusing on the latter type of solutions. Let us note here that many techniques for supporting viewers in watching professionally edited broadcast videos have been proposed; however, due to the differences in the characteristics of broadcast videos and home videos, techniques specifically designed for home videos have been developed.

As shown in Fig.1, in order to help viewers watch the videos captured by amateur camcorder users, mainly two types of video representation: static representation and dynamic representation are generated. A static representation is a collection of salient images or keyframes extracted from the original videos, where viewers can easily browse many videos without actually watching all of them. A dynamic representation is a new good-quality video created from original poor-quality videos, which viewers can watch more pleasingly. Both types of video representation can be generated by analyzing the content of original videos. This requires firstly to segment a video into smaller video segments each of which contains consistent content and then to index them with descriptions about their content. After discussing the characteristics of home videos, this paper introduces existing techniques for video segmentation and indexing which are designed especially for home videos, then discusses how to generate each type of representations.

2. Characteristics of Home Videos

A video is generally a sequence of frames, which are static images, and a sequence of frames which are captured consecutively by one camera is called a shot. As Fig.2(a) shows, a professionally edited broadcast video is generally composed of several scenes, each of which is created by sequentially arranging several shots unified
by their semantic content. Each shot in a broadcast video usually lasts a short period of time and contains consistent content.

On the other hand, a home video is captured with the camcorder users’ start and stop operations, and as a result, corresponds to a shot. However, since camcorder users often capture different targets by intentionally moving the camcorder without switching it off, a home video shot usually lasts a relatively long period of time and contains inconsistent captured targets. Further, due to camcorder users’ unskillful capture behaviors such as unintentional erratic camera operation and defocusing and poor environment such as weak lighting and varying illumination, a home video often contains segments of low perceived quality, which are rarely seen in broadcast videos. Therefore, especially for home videos, a subshot is often defined as a subsegment within a shot, each of which contains consistent camera motion, consistent captured target, consistent level of perceived quality, etc. as shown in Fig. 2(b).

The shot composition of a broadcast video can be defined for a specific genre with visually and semantically very similar shots. For example, a news video has many anchor shots each of which initiates or follows a news story, a sports video has many canonical view shots each of which initiates an event in the sports games, and a cooking video is composed of hand shots and face shots. Moreover, broadcast videos of a specific genre can contain specific types of sound effects such as crowd cheering sound in sports videos. Related textual information sources such as the textual transcript of audio or closed-caption, text overlay in a frame, and external textual information such as Internet news for news videos, online sports statistics for sports videos, and recipe texts for cooking videos, are often available for broadcast videos.

On the other hand, home videos are usually recorded freely and their content is diverse. Therefore, it is not easy to categorize them into genres and to determine specific sub-shot composition patterns according to the genres. Further, the audio quality of home videos is often degraded by the background noise of the recording environment, which complicates the audio signal processing such as automatic speech recognition and audio classification. Since home videos are raw, unedited video footage, no text is superimposed on frames and the related external textual information sources are rarely available for home videos. Therefore, while multimodal visual, audio, and textual information is often used for the content analysis of broadcast videos, the visual information tends to be the main information source for the content analysis of home videos. Thus, the following sections introduce content analysis techniques designed by considering the structures of home videos and their visual characteristics.

3. Segmentation and Indexing

This section firstly introduces one of the fundamental content analysis techniques, namely, segmentation and indexing, which can later be used to develop techniques for generating static or dynamic video representations of home videos.

In order to analyze the content of a video, a video firstly needs to be segmented into smaller video seg-
ments each of which contains consistent content. Since consecutive frames within a shot of professionally edited broadcast videos are expected to be quite similar in content, segmenting a video into shots is often the first step of content analysis for broadcast videos. Each shot can then be characterized and indexed with its low-level features such as color, motion, shot length, and audio volume, or mid-level features such as framing types (wide shot, mid shot, close-up, etc.), and audio types (music, speech, noise, etc.). Since videos of the same genre can contain visually and semantically very similar shots as discussed in Section 2, the shots in videos of a specific genre can also be indexed with high-level semantic features such as anchor shots in a news video, canonical view shots in a sports video, and hand shots and face shots in a cooking video. Related textual information sources can also be used to generate various semantic indices for shots.

On the other hand, since a shot can contain inconsistent content for home videos, either a frame or a sub-shot should be handled as the basic unit and indexed with its features. Let us note here that, although there are a few events often captured by amateur camcorder users such as wedding8) and birthday9) whose semantic content (e.g. wedding vows, ring exchange, etc. in a wedding ceremony) can be extracted with certain visual and audio characteristics (e.g. the bride’s white gown and veil, speech, applause, music, etc.), the diverse content, the poor audio quality, and the unavailability of the related textual information sources for home videos usually do not allow high-level semantic indexing. Therefore, the indices which can be automatically generated for frames or subshots in home videos are often limited to low/mid-level features which can be extracted regardless of the semantic content of videos. More concretely, focusing on the characteristics of home videos, home videos are often segmented and indexed considering how the consecutive frames were captured such as how the camera was moved or under what environmental conditions the frames were captured. The following are the four examples of how home videos are segmented and indexed.

Camera Motion

As discussed in Section 2, home videos are often captured by moving the camcorder. When we let the z-axis go through the axes of lens, and be perpendicular to the image plane x-y, four types of camera motion can be defined as: 1) panning, tilting, and rolling, resulted from camera rotations around the x-, y-, and z-axis, respectively; 2) tracking, booming, and dollying, resulted from camera displacement along x-, y-, and z-axis, respectively; 3) zooming(in/out), resulted from lens’ focus adjustment; and 4) still10).

The optical flow \((MV_x, MV_y)\), which is the 2-dimensional motion field for each point \((x, y)\) in a frame, can be computed for each pair of consecutive frames by examining how the feature points in the current frame move in the next frame. The computed optical flows represent the local motion, which is caused by the object motion, and the global motion, which is caused by the camera motion. Since moving the camcorder during the shooting would result in the global motion in one direction, the best transformation between each pair of frames can be computed by fitting the optical flows to the following 2-dimensional affine model. Note that dollying and zooming, panning and tracking, and tilting and booming cause visually similar effects and are often treated as single types of camera motion. Therefore, in the 2-dimensional affine model, \(zoom, rotate, pan, and tilt\) represent the amount of zooming or dollying, the rolling, panning or tracking, and tilting or booming, respectively.

\[
\begin{pmatrix}
MV_x \\
MV_y
\end{pmatrix}
= 
\begin{pmatrix}
zoom & rotate \\
-rotate & zoom
\end{pmatrix}
\begin{pmatrix}
x \\
y
\end{pmatrix}
+ 
\begin{pmatrix}
pan \\
tilt
\end{pmatrix}
\tag{1}
\]

According to the estimated camera motion, each frame can be indexed with the dominant type of camera motion or the estimated amount of each type of camera motion. Further, shots can be segmented into subshots each of which contain consistent dominant type of camera motion. Each sub-shot can then be indexed with the dominant type of camera motion10).

Perceived Quality

Since camcorder users can unintentionally move the camcorder or defocus during shooting under poor environment such as weak lighting and varying illumination, a home video often contains segments of low perceived quality. Therefore, segmenting a shot into subshots with consistent perceived quality would help viewers watch only the subshots of sufficient quality. For example, this is realized by focusing on the three primary properties of camera motion: speed, direction, and ac-
Camcorder User Intention

The camcorder users often intentionally move the camera in a specific way to capture a specific target. For example, they use zoom-in and then keep the camera still to focus on a specific target of interest, or pan the camera to get a large view of the scene, to follow a moving object, or to search for targets of interest\(^5\)\(^\text{(11)}\). Therefore, a shot can be segmented into subshots with consistent captured targets by looking at how the camera was moved. Then, the segmented subshots can be indexed according to the camcorder users’ intention. For example, subshots are indexed as one of the following seven categories: Static Scene, which captures a static object or scene with few object motion, Dynamic Event, which tracks a dynamic event with obvious object motions, Close-up View, which is a close-up of a face, Beautiful Scenery, which records beautiful scenery, Switch Record, where captured targets are switched with relatively fast camera motion, Longtime Record, which records large portion of an event, and Just Record, which are the complement set of other categories\(^5\)\(^\text{(12)}\).

Viewer Attention

Attention is the cognitive process of selectively concentrating on one aspect of the environment while ignoring others. Generally speaking, the amount of attention from viewers can be estimated according to the video content, which is not necessarily related to how the video was captured. For example, salient regions, which have the visually distinguishable, conspicuous image components, attract more human attention than others at the first glance. These are usually high contrast regions, or regions with significantly different appearance compared to their surroundings\(^5\)\(^\text{(13)}\). Similarly, salient motion regions, which have irregular motion direction and velocity compared to their surroundings both spatially (within a frame) and temporally (between consecutive frames), attract human attention\(^5\)\(^\text{(14)}\). Other examples of attention grabbers include human faces\(^5\)\(^\text{(15)}\)\(^\text{(16)}\), moving objects\(^1\)\(^\text{(17)}\), objects in the center of a frame, and objects which are about to enter the camera view after the camera motion\(^1\)\(^\text{(16)}\)\(^\text{(17)}\). Such visual attention grabbers are often called regions of interest (RoIs). Additionally, specific sounds such as salient sounds, speech, and music also attract viewers’ attention\(^5\)\(^\text{(18)}\). Based on the above observations, the amount of attention viewers may pay to the video content can be estimated by the visual and audio features of each frame such as salient regions, salient motion regions, faces, camera motions, audio volume, and audio type. Then, each frame can be indexed with the importance ranking representing the estimated amount of attention\(^5\)\(^\text{(1)}\).

Another way to estimate the viewers’ attention to the video content is to observe viewers’ response such as eye movement and facial expression when they watch the video. When viewers show the positive reactions such as large head motions, few eye blinks and saccades, and positive facial expressions including smiling and laughing, the corresponding frame can be indexed with higher interest score than other frames\(^1\)\(^\text{(19)}\). This approach can estimate the amount of attention only after the viewer watches the video; however, can consider the differences in the amount of attention toward the content among different viewers.

Consecutive frames which are estimated to attract attention more than a threshold are indexed as an interesting subshot, while other consecutive frames can be indexed as an uninteresting subshot\(^1\)\(^\text{(14)}\).
4. Static Representations

Since video is a temporal media, the time which equals to the length of the video is required to view its content. In order to help viewers browse many videos efficiently, static representations of videos are often used. This type of representations has often been used based on the ideas that consecutive frames often have only little variation both in their visual and semantic content; and therefore, a collection of salient images or keyframes extracted from the original video can sufficiently represent the content of the video. Some work has proposed to select the keyframes without considering the structure of home videos. For example, keyframes can be selected so that the original video can be best reconstructed from the selected keyframes\(^{(1)}\). However, most of the existing work has focused on the structure of home videos to develop static representation generation techniques specifically designed for home videos. The following are the three examples of such static representations which can be generated based on the segmentation and indexing techniques discussed in the previous section.

Perceived Quality

After shots are segmented into subshots with consistent perceived quality, the keyframes (the first frames, the middle frames, etc.) of only the subshots of good quality can be presented to viewers. Another approach for generating quality-based static representation is to present the perceived quality of each subshot. For example, the keyframe of each subshot can be presented with face icons ranging from smiling to frowning expressions with different background color representing five levels of perceived quality\(^{(2)}\).

Camcorder User Intentions

An example of the intention-based static representation is to present keyframes of subshots with a specific type of intentions specified by the viewer\(^{(2)}\). The keyframe can be selected according to the intention, for example, a keyframe can be selected from the hold segment\(^{(3)}\) or at the end of the zoom segment\(^{(4)}\) of the zoom-and-hold subshot. Another approach for generating intention-based static representation is, instead of extracting a frame as a keyframe of each subshot, to generate an image or to select frames which represent the content of the subshot more effectively according to the intentions. For example, for subshots with a panning motion, keyframes can be selected by considering the coverage of the captured environment and the panning speed\(^{(5)}\). The keyframes can also be stitched together to generate a panoramic image so that the captured scene can be viewed in a single image. Zooming-in motion followed by a still motion indicates that the user zoomed in to record an object of interest. Therefore, super-resolution enhancement is applied to the keyframe of the still subshot to generate an image with the object at higher resolution. Subshots with considerable object motion indicates that the user is trying to capture the motion. Therefore, a collection of images which represents the action in the shot is extracted\(^{(6)}\).

Viewer Attention

A static representation generated in terms of the former two points of view would still present keyframes of most subshots regardless of the importance of their content. Selecting only the important subshots would more effectively present the content of videos. Intuitively, the frames which would attract the viewers’ attention more than others should be selected as keyframes and presented to viewers. For example, attention curve can be defined as a time series of the estimated amount of attention value associated with each frame. As the frames which attract most attention from viewers, keyframes can be extracted from crest peaks in the attention curve\(^{(7)}\). Selecting the crest peaks in the order of the estimated amount of attention allows viewers to change the number of keyframes to be presented.

Additionally, the temporal information of each subshot and the similarity among subshots are often considered for all types of static representation. For example, subshots which are captured at the same location can be assembled in a group and arranged vertically, while the groups are arranged horizontally in their chronological order\(^{(2)}\). Such grouping is also useful to present the hierarchical structure of the original videos\(^{(2)}\). Another way of presenting the temporal information of each subshot is to use the information bar, which is a translucent overlay for the keyframe, to show the position and length of the subshot. Considering the similarity among subshots, visually similar subshots can be clustered and visualized as piles of their keyframes, which are then presented row-by-row in their chronological order\(^{(2)}\).
5. Dynamic Representations

Another way to help viewers watch home videos more pleasingly is to create *dynamic representations*, which are new videos created by editing original videos. There are mainly two approaches to create the dynamic representations: 1) improving the quality of subshots/shots themselves, and 2) creating a new video of good quality by combining multiple subshots so that the created video would be a subshot sequence in the same way as a professionally edited broadcast video is a shot sequence as shown in Fig. 2(a). The two approaches are introduced separately in the following sections.

### 5.1 Subshot/Shot Quality Enhancement

The following are the three examples of how the quality of each subshot/shot can be improved focusing on the three points of view introduced in previous sections.

#### Perceived Quality

In order to improve the perceived quality of subshots which have been degraded by camcorder users’ unintentional erratic camera operations, the motions in the subshots can be stabilized by smoothing the camcorder motion path and then truncating the missing areas after aligning the video frames along the smoothed camcorder motion path. Motion blur can be removed by transferring sharper image pixels from neighboring frames to corresponding blurry image pixels\(^26\). In estimating the camcorder motion path, considering the possibility of missing area completion and the salient regions helps to avoid cutting out the salient regions in the subshots\(^27\).

#### Camcorder User Intention

As professional video producers intentionally use different framing and camera motion to influence the viewers’ interpretation of the content of shots, the framing and camera motion of a subshot can be changed under certain conditions. For example, if the original subshot captures the entire scene with camera rotation or displacement, the stabilized panoramic video can be constructed by stitching the frames within the subshot together. A rectangular region can be manually specified in the start and end frames of each type of the camera motion as shown in Fig. 3(b). Rectangular regions for other frames can linearly be interpolated. Cropping the rectangular region from each frame of the panoramic video can generate a subshot with different framing and camera motion as shown in Figs. 3(a) and (c)\(^28\).

#### Viewer Attention

The framing and camera motion can also be manipulated in the viewers’ point of view. Without the manual intervention, regions of interest discussed in Section 3 can be used to automatically determine the framing and the camera motion of a subshot so that the RoI would always be presented in a visually pleasing way\(^29\).\(^30\). The techniques for *video re-targeting*, which are originally designed for manipulating the videos to fit into a small display\(^31\)\(^32\), can be used when a subshot contains several distantly-positioned RoIs.

Since camcorder users often record an event from beginning to end, home videos tend to contain enough information to tell the viewer a complete story; however, they also tend to contain redundant segments. Such redundant segments often make home videos boring and reduce their overall aesthetic interest. As professional video producers manipulate motion in videos to exaggerate their emotional impacts and control viewers’ perception of the captured events, one way to improve the quality of a shot is to manipulate the playback speed of its subshots to produce motion rhythms in the shot. For example, after segmenting a shot into interesting and uninteresting subshots, uninteresting subshots can be sped up while interesting subshots can be slowed down to maximize the aesthetic interest of the original shot\(^14\). While this approach only manipulates the temporal configuration of the frames within a shot, *video condensation* changes the spatial configuration of a frame so that RoIs in different frames can be presented in the same frame\(^35\). This technique can increase the importance of the shot while shortening its duration, so that the viewer can watch the shortened shot to grasp the content of the original shot in a shorter time.

### 5.2 Subshot Sequence Creation

For subshot sequence creation, mainly two approaches have been proposed: automatic one, where computers automatically create a dynamic representation of good quality, and semi-automatic one, where computers assist amateur video editors in selecting and arranging subshots according to their preferences on an interface by presenting supportive information to improve the quality of the created dynamic representation.
Note that the subshots can be used after their quality is improved by subshot quality enhancement techniques introduced in Section 5.1.

(1) Automatic Subshot Sequence Creation

One of the main purposes of creating dynamic representations by combining multiple subshots is to generate a video summary which can convey the overall content of the original videos in a short period of time. Based on the ideas that the most important subshots should be included in a video summary while sufficiently excluding redundant subshots, a subshot sequence is often created as a dynamic representation by selecting subshots based on their perceived quality, camcorder user intention, and viewer attention in the similar way to how the keyframes are selected to create a static representation as discussed in Section 4 and by arranging them chronologically.

In order to further improve the quality of the dynamic representation of home videos, many work have tried to automatically synthesize music clips and the subshot sequence. Most approaches are music-centric, which is to create a subshot sequence which matches a given music clip. A common strategy is to firstly extract music boundaries by analyzing the beat, onset, or tempo of the music clip or by detecting the temporal variance in the frequency domain of the signal within the music clip and then to synchronize them with subshot boundaries by extracting a video segment from each subshot so that the duration of the selected video segment would fit with the duration of the music segment between the two consecutive music boundaries. One way to select the video segments is to use the peaks in the estimated amount of attention. Further, speech segment in the audio track of the video can be extracted by detecting the long pause between sentences. According to the extracted speech segment, the boundaries of video segments can be shifted so that the selected video segments would not interrupt the speech within a sentence. Further, several types of transition effects such as cut, fade-in/out, dissolve, and wipe can be used to connect the two consecutive video segments. According to the certain heuristically predefined rules related to the similarity of the two consecutive video segments and the beat strength or the tempo of the music segments, the transition type and duration can be determined so that the visual and audio changes at the video segment boundaries would be correlated to each other to give consistent impression to viewers.

Then, more researchers started to focus on the compatibility between subshots and music segments. For example, a music-video-like video can be automatically generated by aligning visually similar subshots to repetitive music segments in a music clip such as prelude, interlude, and coda. Focusing more on the aesthetic compatibility between subshots and music segments, the best subshot for each music segment can be selected by considering the co-occurrence relationship between low-level features of subshots and music segments. For example, the best subshot for each music segment can be selected so that the brightness of the subshot matches the timbre of the music segment and then time-warping can be used to change the speed of each subshot so that the video velocity would synchronize with the beat of the corresponding music segment. Alternatively, the best subshot for each music segment can be selected so that the amount of motion...
in the subshots and the tempo of the music segment would be correlated\(^40\). Such approach is related to computational media aesthetics\(^41\), which is an algorithmic study of a variety of visual and audio elements in media based on their use in the film grammar, and often uses heuristically predetermined rules about how certain low/mid-level visual and audio features should be correlated to each other to evoke certain impression in viewers’ mind\(^42\).

While these work tried to synthesize subshots to a music clip, video-centric approaches have also been proposed, which synthesize several music clips to a subshot sequence by selecting the best music clip for a subshot or a subshot sequence also based on heuristically determined rules about the aesthetic compatibility between subshots and music clips, for example, how the dynamic, motion, and pitch of subshots and music clips coincide with each other\(^43\). Other work tried to statistically examine which audio-visual features actually correlate with each other in the professional created documentary videos to determine a policy for selecting the most suitable music clip for a shot taken by a camera mounted on a car\(^44\).

Other than the co-occurrence relationships between visual and audio features, the film grammar also defines how certain types of shots should be arranged in sequence in an edited video. Such rules help to create a subshot sequence which is not necessarily a video summary. For example, original video subshots are firstly characterized by their framing types, duration, camera motion, etc. and the syntax rules such as two subshots of extremely different framing types, e.g. close-up and long shots, should not be connected, the duration of a still subshot should be up to 15 seconds, and a still subshot precedes panning/zooming subshots\(^45\), are used to automatically select and arrange subshots. Further, home videos can be repurposed so that the created dynamic representation delivers a specific intent such as cheer, serenity, gloom, and excitement, which is specified by the user. Low-level features such as lighting, color, framing type, and duration, of each subshot can be manipulated to deliver a specified intent based on the film grammar\(^46\).

Since various syntax rules can not be easily provided to the system manually, example-based approach, which learns the rules from the professionally edited video examples by using machine learning techniques, has been proposed\(^47\)–\(^49\). The patterns in the audio features at the audio boundaries synchronized with the shot boundaries, in the audio and visual features of the synchronized audio and shot boundaries for each type of transition effects, in the audio and visual features of the synchronized music segments and subshots, and in the visual features of the subshots sequentially arranged to construct a scene or a video are learned from a set of professionally edited video examples of a specific genre/intent. Additionally, by comparing original shots and the video segments which are extracted by professional editors from the original shots and then used in the edited videos, the visual features of the video segments which should be selected and the audio features of the selected video segments whose audio volume should be changed are learned. By subjectively comparing the subshot sequence created after each process, namely, subshot sequence generation, music selection, music boundary extraction, video segment selection, transition effect selection, and audio volume adjustment, all of which are based on the learned patterns, the first two are concluded to be the most important processes to effectively improve the aesthetic quality of the dynamic representations.

(2) Semi-automatic Shot Sequence Creation

While the automation makes the dynamic representation generation by combining multiple subshots faster, it usually involves depriving humans of control, which is not always desirable\(^50\). Since the selection of subshots and music clips should largely depend on personal preferences, certain amount of control in selecting them should be given to humans. Many video editing tools such as Adobe Premier\(^50\) and Windows Movie Maker\(^51\) have been developed to enable amateur video editors to create a new video from multiple home videos and music clips; however, the task is still time-consuming and the quality of the created videos depends on the editors’ levels of technical or artistic skills.

Therefore, the techniques to support amateur editors in creating a new video from multiple home videos and music clips have been proposed. Static representations discussed in Section 4 are generally used to present subshots to editors. Further, in order to support editors in the selection and arrangement of subshots and music clips, a storyboard or a template for the target video is presented where editors can use drag-and-drop function to sequentially arrange subshots. Providing the template describing the correlated temporal structure of the subshot and music clip sequences such as the temporal duration and the hierarchical structure of shots and scenes helps editors easily create a video of good qual-
A template can also describe the visual and audio characteristics of each subshot and music clip. Such a template can be created either manually or automatically, for example, based on the example-based approach, by considering the film grammar. In order to provide more support in the subshot selection, the subshots can be presented based on their characteristics. For example, subshots are presented based on the similarity of their emotional identity so that a video with continuous emotional mood can easily be created. Subshots can also be presented based on their suitability to the given template determined by their visual features so that the editors can select the subshots which fit the template without actually watching the subshots. Fig.4 shows an example of the user interface which provides such template-based support. A template for the shot and scene sequences is presented on the bottom left of the interface so that a subshot and a music clip can be selected for each shot and scene of the target video, respectively. The subshots are presented based on their perceived quality, suitability to the template, and chronological order on the top left of the interface. The editors can drag and drop the selected subshots to the boxes on the right side of the interface which are linked to the template to create a subshot sequence. A list of the suitable music clips for each scene is also presented so that the editors can select the one they like from the list.

6. Conclusion

This paper introduced content analysis techniques for segmentation, indexing, and static and dynamic representation generation specially designed for home videos. Although these techniques are mainly based on content-independent low/mid-level features of videos due to the diversity of the content of home videos, they have succeeded in helping viewers watch the originally poor-quality home videos to some extent. Since the ground truth for these techniques cannot be clearly defined, their performance is typically evaluated using the results of a subjective study configured by each research work. Therefore, as future work, the development of high-level semantic content analysis techniques and well-established evaluation measures would be desired. Additionally, the development of low-cost techniques would be a new direction due to the recent emergence of mobile devices such as smart phones or pads.

References


Naoko Nitta Naoko Nitta received the B.E., M.E., and Ph.D. degrees in engineering science from Osaka University, in 1998, 2000 and 2003, respectively. She is currently a Lecturer in the Division of Electrical, Electronic and Information Engineering, Graduate School of Engineering, Osaka University. From 2002 to 2004, she was a research fellow of the Japan Society for the Promotion of Science. From 2003 to 2004, she was a visiting Scholar at Columbia University, New York. Her research interests are in the areas of video content analysis and image/audio processing. She is a member of the IEEE, the ACM, the IEICE, and the ITE.

Norobu Babaguchi Norobu Babaguchi received the B.E., M.E., and Ph.D. degrees in communication engineering from Osaka University, in 1979, 1981 and 1984, respectively. He is currently a Professor of the Division of Electrical, Electronic and Information Engineering, Graduate School of Engineering, Osaka University. From 1996 to 1997, he was a Visiting Scholar at the University of California, San Diego. His research interests include image analysis, multimedia computing and intelligent systems. He received Best Paper Award of 2006 Pacific-Rim Conference on Multimedia (PCM2006) and Fifth International Conference on Information Assurance and Security (IAS2009). He is on the editorial board of Multimedia Tools and Applications, Advances in Multimedia, and New Generation Computing. He served as a General Co-chair of the 14th International MultiMedia Modeling Conference (MMM2008), ACM Multimedia 2012, and so on. He is a Fellow of IEICE, a Senior Member of IEEE, and a member of ACM, IPSJ, ITE and JSAI.