Key Frame Detection Method from News Program using Common Features to News Topics

Kohei Sato[1], **Masaru Sugano**[2], **Hitomi Murakami**[3], and **Atsushi Koike**[4], Non-members

ABSTRACT

A hard disk drive and a flash memory have become high capacity, so that we are now in the situation where a lot of TV programs can be recorded. However, while convenient, it is not easy to watch all the recorded programs in a limited time. Therefore, an efficient TV browsing application is needed.

In this paper, we propose a key frame detection method for TV news program to allow a viewer to efficiently browse TV news programs and/or particular TV news topics. Instead of using features specific to certain news topics, our proposal is to extract important features which are common to news topics, and to generally detect the unique key frame which best represents each news topic. We show the validity of our proposed method by a simulation experiment.

Keywords: Key frame, News Topics, Shot, Detection

1. INTRODUCTION

A hard disk drive and a flash memory have become high capacity, so that we are now in the situation where a lot of TV programs can be recorded. However, while convenient, it is not easy to watch all the recorded programs in limited time. Therefore, efficient TV browsing application is needed. TV browsing application provides the functionality to allow a viewer to browse only the scenes that he or she wants to watch. Such application makes it time-efficient to understand more programs.

A TV program consists of a series of frames, which are filmed without the break, called a shot. A shot represents a sequence of frames captured from a continuous record from a video camera. One of the solutions to build the above TV browsing application is to provide a set of key frames representing the shots which construct the program, in order for a viewer to efficiently understand a TV program in a limited time.

Manuscript received on February 28, 2011 ; revised on March 13, 2011.

[1,3,4] The authors are with Department of Information and Computer Science, Seikei University Musashino, Japan., Email: dm116218@cc.seikei.ac.jp, hi-murakami@st.seikei.ac.jp and koike@st.seikei.ac.jp

[2] The author is with KDDI R&D Laboratories, Inc., Japan.,Email: sugano@kddilabs.jp

Key frame is the frame which can represent the remarkable content and information of the shot. It is said that shot segmentation and key frame extraction are the foundations of video analysis. In other words, specifying and detecting only a key frame that represents the shot is indispensable. However, a TV program usually contains a large number of shots. If a key frame is always detected from each shot, there is large quantity of key frames. Therefore selection of key frames is also necessary. Our motivation is to select a unique key frame which best represents each topic or scene. Selection of key frames differs according to the content and the purposes of a TV program. In this paper, we propose a key frame detection method for TV news program. By the conventional method, key frames were extracted based on the various objects which represent the news topic. It is thus necessary to specify many kinds of objects in advance as an important object. However, it is practically difficult because of the variations of the news topics.

Our method is based on recognition of the TV production techniques [1]-[2], and thus does not exploit an algorithm to recognize various important objects independently. That is, instead of using features specific to certain news topics, our proposal is to extract important objects and features which are common to news topics, and to generally detect the key frame which represents each news topic. We show the validity of our proposed method by a simulation experiment.

This paper is organized as follows. In Section II, related existing works are outlined. In Section III, the details of the preliminary experiment are described. Our proposed method is presented in Section IV. In Section V, the details of quantitative experiment are described. The experimental results and comparative analysis are shown in Section VI.

2. RELATED WORKS

So far, many algorithms for key frame detection have been reported. Here, we overview one of the typical approaches. Key frame detection by conventional methods is based on dividing the shot into sections at first, and then selecting key frames from sections using a clustering based approach. For example, it is unsupervised cluster algorithm [3]-[4]. In addition, there exists other method, such as a color histogram

method [5]-[6]. By these methods, key frames are detected from all the shots after shot segmentation. When one news topic is divided into N shots, there will be N key frames by the conventional methods. When applying these methods to TV news program, a sequence of key frames will be a summary of a news program. Although key frame is the frame which can represent the remarkable content and information of the shot, a lot of unnecessary frames are included. Therefore, those frames are not proper key frames with the most important information.

Figure 1 shows the difference between the conventional method and our method.

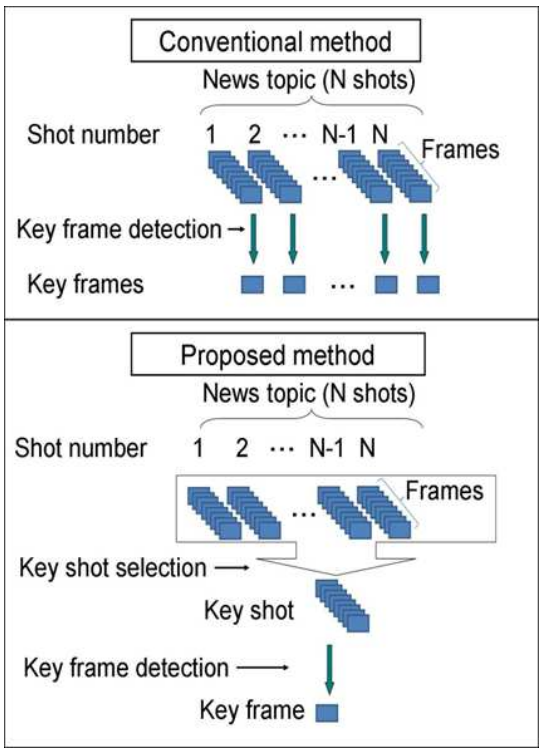

Fig.1: *The procedure of key shot detection*

3. PRELIMINARY EXPERIMENT

As described above, our goal is to detect only one key frame in a certain news topic which is composed of many shots. Towards development of the proposed algorithm, the preliminary experiment was performed. Firstly, we analyzed how the TV program is constituted before the experiment. Analyzing the composition of TV programs is to understand TV production technique, the shooting technique and the editing technique, and to interpret them as the specific features required in order to develop an algorithm. Here, a subject of research is the news program which is organized in similar way even if the broadcasters differ. Because the news programs resemble each other in composition, it is easy to understand the flow of programs.

The news program consists of plural different news topics. In the preliminary experiment, three typical news programs were selected, and each was analyzed manually how the shots are constructed. We used the video streams in MPEG compressed format, and chose the various temporal segments in each program in order to avoid biased contents. We paid attention to what is selected as features in the analysis. By investigating every shot, the change of camera work, the appearance of the telop, and the presence of the person are manually labelled. We performed the work to detect an important key frame by human subjective judgment.

Consequently, it turned out that the display time of a telop, the existence of a camera flash, the camera work and the display time of the object representing the content of news in the shots are useful features to detect an important frame within a certain topic. The telop in those topics synchronizes and continues with a sound even if a shot changes. So the existence of the telop is not relevant to a key frame. In addition, the camera flash is very effective in the topics such as the press conference, but is not always helpful for other topics. Therefore we decided to use i) the quantity of the camera work, ii) the stability of a motion in a shot, and iii) shot duration, which are common to the various topics. Here, the stability of a motion is derived from the display time of an important object, related to a news topic, captured with a small amount of camera motion.

It turned out that the quantity of camera work, the stability of the motion in a shot and the duration of the shot have a close relation to the average amount of total motion vectors, the variance and the number of the frames, respectively.

4. PROPOSED METHOD

According to the results from the preliminary experiment, the detailed algorithm was designed. As described in Section 2, the difference between our proposed method and the previous methods is that only one key frame is chosen from all shots in a certain news topic. This also means that our method selects only one representative shot which contains the key frame. Such a key frame of a representative shot serves as an important picture which can be used as a thumbnail image in the TV browsing application.

Our key frame detection algorithm has basically three steps. They are shot segmentation, shot selection, and key frame detection.

First, video streams are divided into shots by shot segmentation. Then only important shots are chosen among all those shots. Finally one key frame is selected from those chosen shots. The procedure is illustrated in Figure 2.

Since we put weight on shot selection, the video streams are segmented into shots beforehand this time. The shot selection algorithm that we developed

is divided into four processes.

i) Select shots with the top three of the long frame length.

Firstly, the number of frames is calculated, and the top three shots with the longest durations are chosen. In the preliminary experiment, all the shots with very short length turned out to be of no importance. The reason why we chose shots with the top three long lengths is that the important element in selecting shots is the length of a shot. We regard the length of a shot as the most significant element of all.

ii) Omit the shot with the smallest motion variance.

With the optical flow, the average amount of total motion vectors and their variance are calculated. The shot with the smallest motion variance is removed from the above selected three shots. The larger the variance is, the lower the motion stability of a shot is. In the preliminary experiment, all the shots with very low motion stability turned out to be of importance. Usually, a shot with low motion stability includes two types of segments; one is a segment with large motion and the other is that with almost no motion. For example, in a certain shot, the first part contains still frames and the second part contains the fast tilting operation. It is highly possible that the latter, i.e. a still segment, may contain a key frame. In the preliminary experiment, it is turned out that the first several frames of a shot with low motion stability cannot have many motion segments. Therefore we assume that the key frame is the first frame.

iii) Select the shot with lower average amount of total motion vector.

The shot with the lower average amount of total motion vectors is chosen from the remaining two shots. The lower the average amount of total motion vectors is, the smaller quantity of camera work exists. The reason for making the selection of the variance before the selection of the average amount of the total motion vectors is to give priority to the size of change of a shot. Here, a shot with change by the simple movement of the camera is excluded. This is because simple movements such as panning and tilting are not important.

iv) Key frame detection.

Finally, the beginning frame of the shot is extracted. It is impossible that camera switches in a shot of video stream. And the pattern of all the images that constitute the shot becomes similar without significant changes. Therefore, the beginning frame is simply detected. It becomes the key frame. Figure 3 illustrates the framework of our algorithm simply.

5. VALIDATION EXPERIMENT

To verify the validity of the algorithm, it is implemented by VC++ in Window Vista environment on Intel i5 (3.60GHz) with 4GB RAM. We used OpenCV as the video processing library. The computational

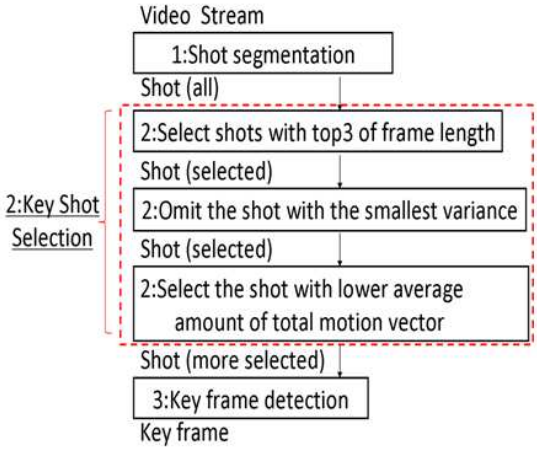

Fig.2: *The procedure of key frame detection*

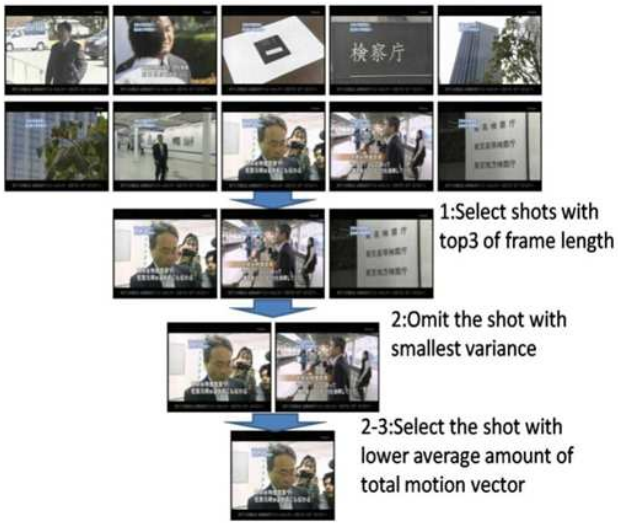

Fig.3: *Key shot selection*

simulation experiment which detects and specifies a key frame using these features was conducted for the several news topics. We used the video stream in MPEG compressed format, as in the case of a preliminary experiment. Eleven programs were recorded from the analog TV broadcasting using the capture board on the personal computer. The contents are various, such as interviews, speeches, and reports. Video streams were analyzed using the proposed algorithm.

6. RESULTS AND COMPARATIVE ANALYSIS

Figure 4 shows the key frames selected by manually and those detected by the proposed method. Each key frame corresponds to a certain news topic. As expected, the common characteristic was observed by a result in spite of different selection methods. When the obtained images of eleven topics by two methods

were compared, we got reasonable results in seven topics. However, there were some cases in which the different frame was chosen although the same shot was determined. That is because, in this experiment, we assume that the key frame is the first frame of a shot. As a result, it turned out that a key frame is not always the first frame. In addition, there also exists a video stream in which a key frame cannot be selected properly. For example, when two or more subjects are important in a news topic (e.g. the news topic of introducing an artist and his/her artwork), it is difficult to decide which subject is the most important. In addition, when the subject itself does not appear in a sequence of shots, while other stuff which is not so much related to the subject only appear, it is also difficult to identify the key frame. Therefore measurement for evaluation is very important. There is no key frame everyone chooses, so a frame selected by more participants is suitable as key frame. And to select more subjectively reasonable key frame, we conducted the subjective evaluation experiment by the questionnaire. This time, subjective evaluation method which we used is MOS (Mean Opinion Score). Evaluation with MOS criterion is further described in the next section. We want to pursue flexibility by using more news topics of more broadcasting stations. Although we thought that a telop was an important factor which constitutes a news program, it was not used this time. Therefore, how to use a telop as a feature is next theme.

7. EVALUATION OF A DETECTION RESULT

So far, various researches on key frame detection have been performed. However, evaluation criteria are various among individuals. At this time, there is no standard objective criterion for evaluation of a key frame detection algorithm. Therefore subjective evaluation using the questionnaire is performed much, not objective evaluation using the optical instrument.

We also sent out questionnaires to college senior and evaluate in the study.

7.1 Evaluation standards

The evaluation method of pictures can roughly be divided into two categories; subjective evaluation method and objective evaluation method.

Subjective evaluation is a method by manually and objective evaluation is a method by compared original picture with processed picture using numerical value. But objective evaluation for key frame is very difficult and evaluation of contents is limited to subjective evaluation method. The key frame of TV program is different among individual and there is no standard for the evaluation. So it is hard to evaluate objectively by compared numerical value. The relation of value acquired by object evaluation and evaluation by vision is not still known.

Fig.4: *key frames chosen by two methods*

Therefore subjectivity evaluation is performed much, not objectivity evaluation with numerical comparison. MOS (Mean Opinion Score) is used most widely in subjective evaluation method. The calculation method is as follows:

$$\mathrm{MOS} = \frac{1}{n}\sum_{t=1}^{n}\alpha_t \tag{1}$$

The number of rater is n, and α_t is evaluation value of person of rating. Although MOS is a very simple method, in order to raise accuracy, it is necessary to

increase the number of rater. The standard deviation of MOS is regarded as about 0.1 in the case of fifteen to twenty people.

There are two kinds of methods of subjective evaluation: a method which only a frame is evaluated, and that which two or more pictures are compared. In our case, since evaluation criterion is whether selected frames have important information or not, the method of evaluating only a frame is used. Evaluation of only a frame is often using rating scale method. This method is that a numerical value is given to a category ordered and a scorepoint is acquired from results.

Firstly Estimation divides into about five categories, For example, It is "very good", "good", "fair", "bad ", and "very bad". For example a highest score is set to 5 points, the lowest point as 1 point, points are given to each evaluation, so results are gotten from person of rating.

7.2 Evaluation experiment with questionnaires

Targeted at eight people (male: seven persons and female: one person), the subjective evaluation experiment by the questionnaire was conducted. Subjects of questionnaires are three news topics in three programs.

A result of questionnaires is indicated in Table 1. Sum in Table 1 are summation values for the answer of each question, and the perfect score is 40 points. When a number is large, the extracted key frame is highly regarded.

In the case of Figure 5, the upper frame is selected as key frame automatically, but some people think that the lower frame is better as a key frame than the upper frame. This is because the news topic has two key objects: album jacket and an artist.

In the case of Figure 6, a value of MOS is 2.00. This value is comparatively low. Because many people assumed that the lower image in which a person appears clearly is better than the upper image.

In the case of Figure 7, a value of MOS is 3.875. This value is comparatively high. This is because the key

Table 1: *Mean Opinion Score*

	Very good	good	fair	bad	Very bad	SUM	MOS
The Beatles	3	2	2	1	0	31/40	3.875
Iwamura	1	0	0	4	3	16/40	2.00
Higashikokubaru	3	1	4	0	0	31/40	3.875

(a)our method

(b)selected by subjective evaluation experiments

Fig.5: *News topic including two important subjects*

(a) our method

(b) selected by subjective evaluation experiments

Fig.6: *News topic including Iwamura player*

(a) our method

(b) selected by subjective evaluation experiments

Fig.7: *News topic including Higashikokubaru governor*

8. CONCLUSIONS

In this paper, we proposed a key frame detection method for TV news program. Instead of using features specific to certain news topics, our proposal was to extract important features which are common to news topics, and to detect the key frame which represent each news topic. These features make it possible to find the key frame in news topic regardless of these topics on the occasion of detection of an important shot. As a result of comparing with the manually detected frames, it turned out that the similar frames in terms of representing news topics were detected in many cases. Subjective evaluation with MOS for 2 news topics achieves the very high level of 3.85. Therefore, the evaluation experiment was able to show the validity of average amount of total motion vector, dispersion, and the shot length. It is concluded that our method is effective in detecting key frame from news programs.

References

[1] T. Takahashi, M. Sugano, and S. Sakazawa: "Automatic Thumbnail Extraction for DVR Based on Production Technique Estimation," *IEEE Transactions on Consumer Electronics*, vol.56, no.2, pp.888-894, May.2010

[2] T. Takahashi, M. Sugano, and S. Sakazawa: "Intuitive Thumbnail Extraction for Content/Scene Navigation," *ITE Technical Report*, vol.34, no.8, pp.13-16, Feb.2010[In Japanese].

[3] G. Ciocca and R. Schettini: "An innovative algorithm for key frame extraction in video summarization," *J. Real-Time Image Process*, vol.1, no. 1, pp. 69-88, 2006.

[4] R. Ogushi, K. Takeuchi, Q. Zhu, A. Kodate, and H. Tominaga : "Extraction of Keyframes from Video Sequence," *IPSJ*, pp.53-58, Mar.2001[In Japanese].

[5] G. Liu, and J. Zhao: "Key Frame Extraction from MPEG Video Stream," *ISCSCT '09*,pp.007-011, Dec.2009

[6] M. Smith and T. Kanade: "Video Skimming and Characterization through the Combination of Image and Language Understanding," *Proc. ICCV98*, pp.61-70, Jan.1998.

Kohei Sato received the B.S. degree in Faculty of Science and Technology from Seikei University in 2011. His research interest is image processing.

Masaru Sugano received the B.E. degree in electronics and communications engineering and the M.S. degrees in electronics, information and communications engineering from Waseda University in 1995 and 1997, respectively. Since 1997 he has been with Research and Development Laboratories of KDD, Saitama, Japan, and engaged in research on video coding and video/audio analysis for content-based search and retrieval. From 2005 to 2006 he was a visiting researcher at the Robotics Institute of Carnegie Mellon University. Currently he is a Research Engineer of Ultra-Realistic Communications Laboratory at KDDI R&D Laboratories, Inc.

Hitomi Murakami received Ph. D degree from Graduate School of Engineering,Hokkaido University, Japan in 1974. Then he joined KDD corporation and since then, he engaged himself in the research of digital satellite communication, digital TV transmission, TV signal transform technology. He became an executive CTO of KDDI corporation in 1997, and since then, president of multimedia technology businesses, president of R&D and president of network technologies. He was a chairman of TTC standardization and ITUSG9 in Japan. He was also the vice president of Society of Video Media in Japan. Presently, after his retirement of KDDI corporation, he is a professor of Department of information and computer science, Seikei University, Tokyo.

Atsushi Koike was born in 1961. He received the M.E. degree from Tohoku University in 1985 and the Ph.D. degree from Kyoto University, Japan, in 2002. He was working at KDDI R&D Labs, Saitama, Japan, from 1985 to 2009. He is currently a professor in the Department of Computer and Information Science, the Faculty of Science and Technology, Seikei University, Japan. His research interests include visual communication system, highly efficient coding of moving images, computer vision, medical information system. He received the Niwa-Takayanagi Achievement Award of Institute of Image Information and Television Engineers in Japan, 2007. He is a member of Communication and Signal Processing Society in IEEE.