Video summarisation based on the psychological content in the track structure

Tsuyoshi MORIYAMA
Institute of Industrial Science, University of Tokyo
7-22-1 Roppongi, Minato-ku
Tokyo, 106-8558 Japan
Tel. +81-3-3402-6231 ext.2645
Fax +81-3-3402-5078
moriyama@sak.iis.u-tokyo.ac.jp

Masao SAKAUCHI
Institute of Industrial Science, University of Tokyo
7-22-1 Roppongi, Minato-ku
Tokyo, 106-8558 Japan
Tel. +81-3-3402-6231 ext.2645
Fax +81-3-3402-5078
sakauchi@sak.iis.u-tokyo.ac.jp

ABSTRACT

In this study, the method to summarise TV drama based on the psychological unfolding is proposed. In past researches, video summarisation has been achieved almost by way of choosing representative frames in visual track and never based on the mental side of video content. In this work, the video structure which consists of audio & visual tracks (actor's utterances, BGM, background sound, effect sounds and shots) is modeled and the track-structure-based video summarization method is proposed. To extract the temporal feature patterns of the track structure correspondent with the specific psychological content, each tracks is first quantified by calculating existence ratio etc., and second the intra/inter track feature patterns are determined based on empirical knowledges. The proposed method is implemented and examined in the subjective experiment.

Categories and Subject Descriptors

H.3.1 [**Information Search and Retrieval**]: Content Analysis and Indexing—*abstracting method*; H.4.0 [**Information Systems Applications**]: General; H.1.2 [**Models and Principles**]: User/Machine Systems—*Human factors*

General Terms

Human factors, Theory

Keywords

multimedia, video summarisation, track structure, TV drama, time series pattern

1. INTRODUCTION

Digital broadcast is beginning, and many audio & visual resources will be delivered to consumers in a few years. In addition to efforts to produce multimedia contents, the technologies which provides consumers with capability of editing, searching and summarising vast amount of multimedia data grow its importance increasingly.

So far, many kinds of researches have conducted concerned with video processing [1, 2]. Rui *et al.* [3] categorized them into four fields: namely *video analysis*, *video representation*, *video browsing* and *video retrieval*. In the field of video summarization, especially, as one of video representation above, a typical strategy is extracting shots, followed by determining representative frames or tracking objects [4], indexing them by some physical or semantic attributes, and providing description-based browsing and retrieving [5~7]. This is so-called frame-based method. In another approach, object-based method was proposed by Günsel [8], it is not based on the frame structure of video but on the *video object* prescribed in MPEG-4 standard.

Meantime, in the case of TV drama especially, since mental unfolding is often more essential than behavior of objects, then it should be more effective for audience to matter the key relevant to express mental unfolding in addition to semantic elements. On the other hand, past studies have mainly focused on only visual content of multimedia data except for several studies [7, 10~12], and audio track, especially speech of actors, has been used as just a sort of captions before, it has never been related to mental content of video.

From the viewpoint of the physical aspects of video, TV drama is produced based on mainly two factors of *montage* (editing components) and *mise-en-scene* (studio sets) [9], then it can be assumed that *montage* corresponds with combination and arrangement of track structure (cutting, incorporating BGM and sound effects etc.) and it contributes to produce psychological unfolding of drama video. Though object-based and frame-based approach in past studies treated result of *montage* indirectly, track-structure-based approach will more directly describe the psychological content of drama video. As mentioned by Vasconcelos [9], the emotions of the audience is related to the physical structure of a TV drama which is consciously structured by the movie makers in order to express some psychological unfolding with personalities of character's and emotion, therefore it also seems to be suitable for modeling mental perception of the audience fundamentally.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
ACM Multimedia Workshop Marina Del Rey CA USA
Copyright ACM 2000 1-58113-311-1/00/11...$5.00

In this work, a video summarization method is proposed and implemented by formulating *montage* correspondent with the emotions of the audience. It is newly based on the mental content of drama video. In the following section, the track structure is mentioned and the objective summarization strategy is presented in Section 2 and 3. The proposed method is implemented and examined in Section 4, followed by conclusion with future work in Section 5.

2. THEORY

2.1 Structure of TV drama

The physical structure of TV drama video can roughly be divided into visual component and audio component. The visual component has been focused on in most past researches, through the hierarchical model of *video - scenes - shots - key frames - objects - segments* when describing the spatial features such as shape and geometrical information of the object and the temporal features such as the trajectories of the moving object. Saraceno [10] used audio track as the support of cut detection based on the difference energy of the luminance between adjacent frames for video news and advertisements, while the proposed method enthusiastically takes advantage of audio channel for TV drama summarization. The physical tracks of a TV drama is illustrated in Fig.1. The visual component is represented by *shot* track alone here, which are caused from camera works, and the audio component is divided into four tracks, namely *speech by actors*, *background sounds*, *effect sounds* and *BGMs*. Picture effects such as 'dissolve' in shot boundaries are observed but the middle point of the transitional part is simply regarded as the cut point here. In terms of the definition of each tracks, the music as the *mise-en-scene* (in the studio set) is not considered as BGM, but the music superimposed on the video is defined as the BGM. In such superimposed sounds, those which don't have any melodies and rhythms are defined as effect sounds. Audio components which don't belong to anything above are grouped as the background sound.

This track structure model is considered from the viewpoint of *montage* components related to the psychological unfolding of drama video, so it is not clear whether it is sufficient to describe the physical aspect of various kinds of drama video, and to apply it to other sort of video such as TV news. It is intended to include all components which can be considered important for the time being.

2.2 Acquisition of the relation information

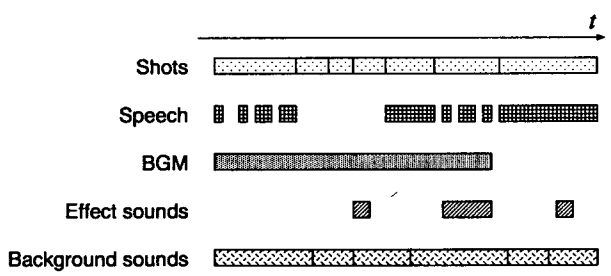

Figure 1: Track structure of the drama video

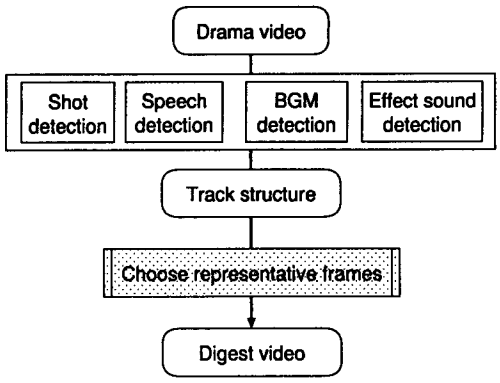

Figure 2: Track-structure-based video summarization

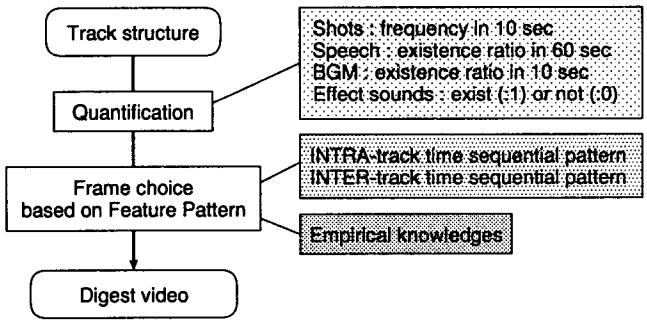

Figure 3: How to choose representative frames

In this study, each local track structure of video sequence is related to its psychological impression (climax, tense and so on). This relation information is applied to put the weights on scenes in drama, and the candidate frames of digest video are determined based on the weights.

3. TRACK STRUCTURE BASED VIDEO SUMMARIZATION

3.1 Observation of track structure

Fig.2 and Fig.3 shows the outline of the proposed system. It is assumed that each tracks are provided individually as standardized in MPEG-4 (AV objects). Then the track structure is extracted manually in this paper.

Based on the assumption that temporal distribution of whether each track exists or not is treated as the only one factor by which the audience affected about the psychological impact of the scene, accordingly quantification indicated in Fig.3 is conducted here.

3.2 Determination of representative frames

Quantification in 3.1 means to encode or to parameterize the track structure into time sequential patterns, in other words, one dimensional vector on the temporal axis. In this section, the feature patterns correspondent with psychologically important places are determined based on '*rules of thumb*' mentioned below;

Shot: Frequent cutting produces strong tense impression

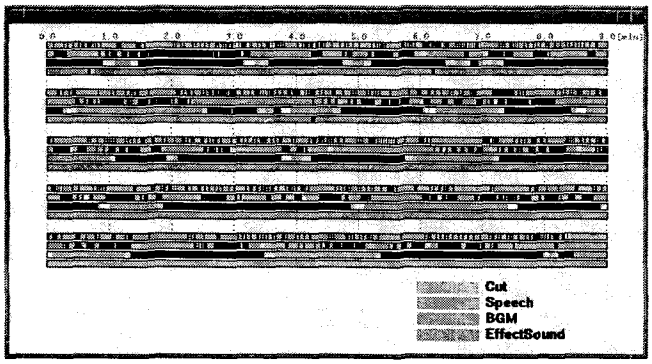

Figure 4: Example of determined representative frames (*'Dangerous relation'*)

Speech: The beginning (topic presentation) and the end (topic conclusion) of a frequent part is important

BGM: Just before moving from non-BGM to BGM is important (coloring neighboring scenes)

Effect sound: Effect sound adds importance by its existence.

By the way, there are definitely some interaction among tracks in convention about *montage*, it is necessary to determine the empirical rules for combination of track patterns based on conventional knowledge. In this study, however, it will be achieved after the 1st step of clarifying the role of each tracks for *montage* in terms of psychological direction. On the other hand, it is, of course, supposed that not only combination of intra-track rules but inter-track rule is originally significant, so they must be investigated simultaneously.

4. IMPLEMENTATION

In this section, the proposed video summarization method is examined for two instances of Japanese trendy dramas, 'Dangerous relation' and 'Neighbor smiles suspiciously'.

4.1 Determined representative frames

First, the track structure illustrated in Fig.1 was extracted, then each 4 one dimensional on/off patterns (black part of Fig.4) was quantified to a time series parameter based on the features mentioned in 3.2. Next, the representative frames were determined in each tracks (white part of Fig.4). In *shot* track for example, the standard deviation σ is calculated for the frequency histogram and 3σ is determined as the threshold that if the number of cuts excess it, it can be judged cuts occur statistically frequently in the neighbor and tense place. Determined representative frames (5 seconds from it) based on such '*rules of thumb*' in each tracks are shown in Fig.4 superimposed on the extracted track structure(A drama video of 45 minutes length is depicted in 5 banches (9 minutes in each banches), 1 banch consists of 4 track mentioned before), and each top frames are shown in Fig.5. In Fig.5, an example in case of being based on *shot* track is indicated, in which audio tracks are accompanied with each truncated clips.

4.2 Result of video summarization

The full time length of each TV drama piece is 44'28" (Fig.4) and 43'12" respectively. Accordingly, concerned with the length, it was reduced remarkably as indicated in Table 1.

In the case of making digest video, which is a consumer aimed technology, it is naturally necessary to evaluate this method by statistical approach through a subjective evaluation expriment. Accordingly the proposed method is compared with a past simple strategy in the following section concerned with this viewpoint.

4.3 Comparison with another method

As a competitor, a method in which frames at fixed intervals of 606 and 1178 frames respecitively are gathered as the representative frames is chosen. It is denoted as DEF like in [12] while the proposed method is denoted as TSB (Track-Structure-Based method). These intervals are so determined as to become totally the same length as the digests made by TSB.

The subjective experiment was conducted, in which 5 students watched two digest video without being informed which TSB was, and answered better one and evaluated about following points; namely 1) how it retains original contents, 2) naturality of truncation and 3) correlation with scene boundaries (meaningful unity).

From the result of subjective evaluation, it was clarified that 1)TSB retains drama content well than DEF, 2)TSB roughly extracts scene boundaries, nevertheless it simply extracts frames at which cuts statistically frequently occur, and 3)DEF truncates disagreeable points than TSB, it causes in giving an impression of unnaturalness and confusion. Moreover, in spite of DEF shows more number of frames than TSB in the same total length (Fig.5), it doesn't contribute to retain the psychological unfolding and density as a result.

5. CONCLUSION

In this paper, the method to digest TV drama based on the psychological unfolding was proposed and implemented. The structural model of a TV drama was mentioned, and the quantification method to describe temporal features of each tracks and how to select the representative frames was presented. As the result of implementation, it is indicated that the proposed method is able to retain psychological unfolding and scene boundaries in addition to reducing temporal length.

This summarization method is innovative from the viewpoint that it treats mental side of multimedia content, and it is fundamental for audience with very simple algorithm and little computational cost. Besides it proposed a new basis in addition to the frame-based and object-based ap-

Table 1: Result of video summarization

drama#	*shots*	*speech*	*BGM*	*effect sounds*
dange~1	3'10"	2'55"	5'55"	0'15"
neigh~2	1'50"	2'40"	4'40"	7'14"

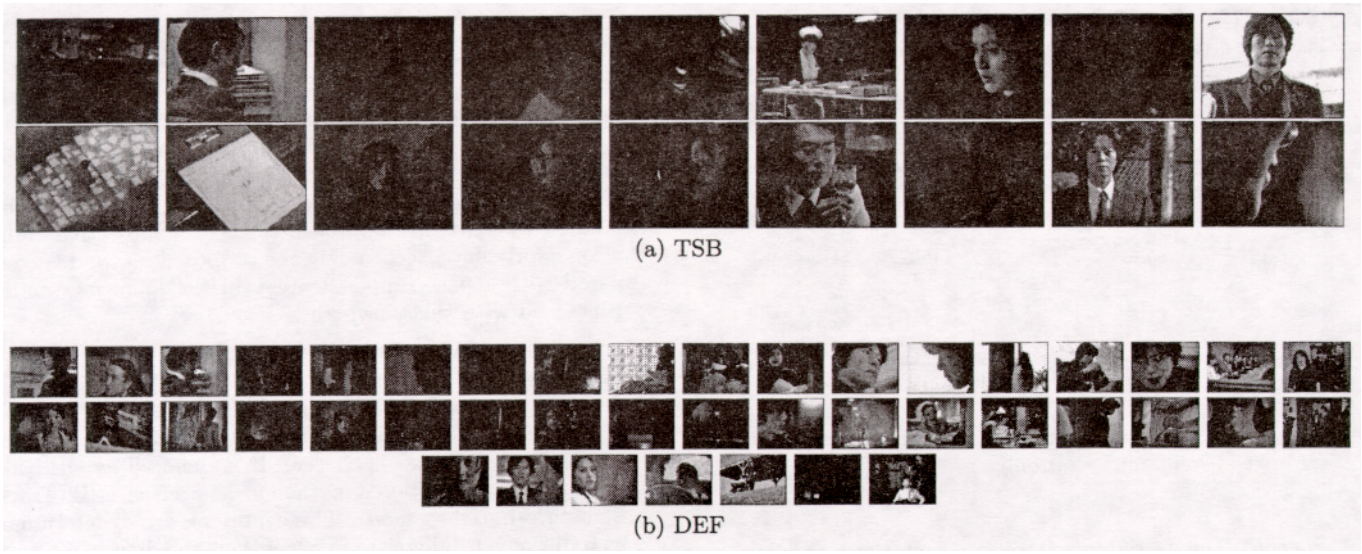

Figure 5: An example of determined representative frames ('Dangerous relation')

proach, the track structure based approach.

In the future work, in addition to polishing intra-track rules up, inter-track rules are necessary to be developed. Moreover, it is effective to implement it in the form of a plug-in working in EPG system.

6. ACKNOWLEDGEMENTS

This work is (partially) supported as the Grant-in-Aid for Creative Basic Researh #09NP1401: "Research on Multimedia Mediation Mechanism for Realization of Human-oriented Information Environments" by the Ministry of Education, Science, Sports and Culture.

7. REFERENCES

[1] Idris, F., Panchanathan, S., *Review of image and video indexing techniques*, Journal of visual communication and image representation, Vol.8, No.2, pp.146-166, 1997.

[2] Roth, V., *Content-based retrieval from digital video*, Image and Vision Computing 17, No.7, pp.531-540, 1999.

[3] Rui, Y., Zhou, S.X., Huang, T.S., *Efficient access to video content in a unified framework*, IEEE International Conference on Multimedia Computing and Systems, Vol.2, pp.735-740, 1999.

[4] Man-Kwan Shan, Suh-Yin Lee, Content-based video retrieval via motion trajectories, Proc. of the SPIE, Vol.3561, pp.52-61, 1998.

[5] Candan, K.S., Golshani, F., Panchanathan, S., Park, Y., VIMOS: a video mosaic for spatio-temporal representation of visual information, 1998 IEEE Southwest Symposium on Image Analysis and Interpretation, pp.6-11, 1998.

[6] Pope, A., Kumar, R., Sawhney, H., Wan, C., *Video abstraction: summarizing video content for retrieval and visualization*, Conference Record of Thirty-Second Asilomar Conference on Signals, Systems and Computers, Vol.1, pp.915-919, 1998.

[7] Ronfard, R., *Shot-level description and matching of video content*, Proc. of the SPIE, Vol.3229, pp.70-78, 1997.

[8] B. Günsel, A.M. Tekalp, P.J.L van Beek, *Content-based access to video objects: Temporal segmentation, visual summarization, and feature extraction*, Signal Processing 66, pp.261-280, 1998.

[9] Vasconcelos, N., Lippman, A., *Bayesian modeling of video editing and structure: semantic features for video summarization and browsing*, Proc. 1998 International Conference on Image Processing, Vol.3, pp. 153-157, 1998.

[10] Saraceno, C., Leonard, R., *Audio as a support to scene change detection and characterization of video sequences*, Proc. IEEE International Conference on Acoustic, Speech and Signal Processing, MDSP3L.1, Vol.4, pp.2597-2600, 1997.

[11] Nam, J., Alghoriemy, M., Tewfik, A.H., *Audio-visual content-based violent scene characterization*, Proc. International Conference on Image Processing, Vol.1, pp.353-357, 1998.

[12] Smith, M.A., Kanade, T., *Video skimming and characterization through the combination of image and language understanding*, Proc. 1998 IEEE International Workshop on Content-Based Access of Image and Video Database, pp.61-70, 1998.

[13] Gelgon, M., Bouthemy, P., *Determining a structured spatio-temporal representation of video content for efficient visualization and indexing*, Proc. 5th European Conference on Computer Vision, Vol.1, pp.595-609, 1998.