A Method for Structuring Digital Video

Jae Yeon Lee*, Se-yoon Jeong*, Ho-Sub Yoon*, Kyu-Heon Kim*, Younglae J. Bae*, Jong-whan Jang**
*Image Processing Division, Systems Engineering Research Institute
PO Box 1, Yusung-gu, Taejon, KOREA, 305-600
**Division of Computer, Information, Communication and Electronics Engineering, Paichai University,
439-6, Doma-dong, Seo-Gu, Taejon, Korea
Tel) +82-42-869-1454, Fax) +82-42-869-1479, e-mail) leejy@seri.re.kr

Abstract

For the efficient searching and browsing of digital video, it is essential to extract the internal structure of the video contents. As an example, a news video consists of several sections such as politics, economics, sports and others, and also each section consists of individual topics. With this information in hand, users can more easily access the required video frames. This paper addresses the problem of automatic shot boundary detection and selection of representative frames (R-frames), which are the essential step in recognizing the internal structure of video contents.

In the shot boundary detection, a new algorithm that have dual detectors which are designed specifically for the abrupt boundaries (cuts) and gradually changing boundaries respectively is proposed. Compared to the existing algorithms that mostly have tried to detect both types by a single mechanism, the proposed algorithm is proved to be more robust and accurate.

Also in the problem of R-frame selection, simple mechanical approaches such as selecting one frame every other second have been adopted. However this approach often selects too many R-frames in static shots, while drops important frames in dynamic shots. To improve the selection mechanism, a new R-frame selection algorithm that uses motion information extracted from pixel difference is proposed.

1. Introduction

The efficient searching and browsing of digital video are very important topics in multimedia service. Due to the large size of digital video and its sequential characteristics, the access to the required information is very tedious and time-consuming. To solve the problem, the hierarchical data models[1]-[3] as shown in Fig. 1 that reflect the internal structure of the contents have been proposed. Based on this model, a news video may be divided into several sequences that are mapped to politics, economics, sports and other sections. Also the sequences can be divided into several scenes that are mapped to individual topics. With this structure information of the video contents, the accessibility can be greatly improved.

Currently the above internal structure information is mainly extracted by manual operations. Especially the construction of sequence and scene level almost completely depends on human, which seems inevitable because the classification of these levels is based on high level semantic information that is hard to extract automatically with the current state of the art. However the recognition of shot level that is defined as an unbroken sequence of frames from one camera[4], is a relatively feasible problem and a lot of researches have been concentrated[1]-[8], but still remains many problems to be solved due to the diversity of video contents.

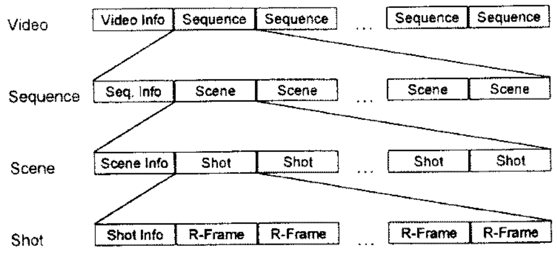

Fig. 1. Internal Structure of Video Contents

R-frames, a lower level than the shot in the hierarchy, are one or several frames that can stand for the contents of the shot. The reasonable selection of R-frames is important in two respects. One is because they are used as visual indices in video browsing. If the selection was not appropriate, user may miss the existence of the necessary information in browsing. Also in considering image content based searching, the image features are extracted from the R-frames only, because it is cumbersome to extract the features from all the frames where so similar frames appear in succession. Hence unreasonable selection of R-frames can result in failure in searching.

This paper addresses the above two problems, a robust shot boundary detection algorithm and an R-frame selection algorithm.

2. Related Works and the Idea of the Proposed Algorithm

Generally shot boundaries are divided into two types. One is an abrupt change in one frame as shown in Fig. 2, which is called a cut. Another type contains gradual boundaries that are transited from one shot to the next in sequence of several frames as shown in Fig. 3.

The most intensively investigated topic in shot boundary detection is to find out a robust feature that can detect cuts without being confused by the fast motion, camera operation or sudden brightness changes. For the purpose, there have been proposed many algorithms that utilize the pair-wise pixel differences, histogram differences, edge differences or a variance of the difference histogram[4]-[6]. These approaches are reasonable because the difference between the prior and the posterior frame of the boundary is large in general. However because of the diversity of video contents, they have weaknesses in certain situations respectively. For example, pair-wise pixel differences tend to generate false positive boundary when fast motion exist, while histogram differences often misses the shot boundaries where the shots of similar atmosphere succeeds[4][5]. And a variance of the difference histogram is robust against brightness changes[6].

In this paper, a combined feature of the above features is used to detect cuts with the expectation that the features can compensate the weaknesses of each other. For the combination, a neural network of back error propagation model is adopted.

As another problem, there are gradual boundaries that are difficult to detect with the above features that are based on the differences between consecutive frames. In this arena, a dual threshold method proposed by Zhang et al may be the representative one[9]. This method adopts two thresholds. One is a high threshold that declares shot boundary if that threshold is exceeded. Another is a low threshold that initiates accumulating the histogram differences. As can be found in this method, many researchers have tried to detect the cuts and gradual boundaries with a single mechanism.

However due to the large difference in the characteristics of each type, it seems more reasonable to solve each problem with respective approach. In this paper, a new detection algorithm that has dual detectors that are specifically designed for each type is proposed. Furthermore instead of simply merging the results of the detectors, the proposed algorithm adopts an arbitration module called a delayed decision module. This additional process not only arbitrates the results of each detector but also checks if the detected shot is reasonable.

For the problem of R-frame selection, mainly very simple approaches such as selecting a frame every other second or selecting the first and last frame of the shot have been adopted[10]-[12]. However these approaches often select too many R-frames in static shots, while drops important frames in dynamic shots.

In this paper, motion information is considered as a key feature for the R-frame selection. However, due to the computational cost of general motion analysis, the pixel difference that is relatively sensitive to motion, is used for the purpose.

3. The Proposed Algorithm

The flow of the proposed algorithm is shown in Fig. 4. It has dual detectors that are designed to detect cuts and gradual boundaries, respectively. Each detector observes the video stream carefully based on its own decision mechanism. However instead of declaring boundary by themselves, their

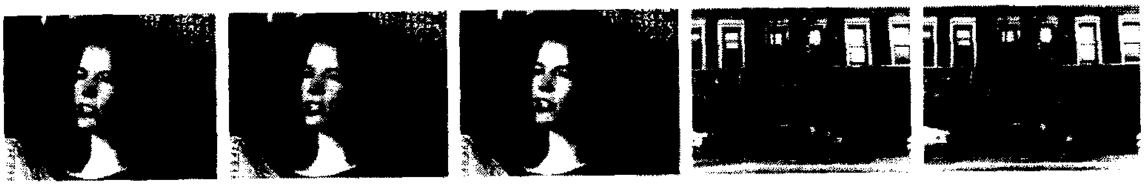

Fig. 2. An Example of Cuts

Fig. 3. An Example of Gradual Boundaries

observation results are reported to the delayed decision module.

The delayed decision module has two roles. One is to arbitrate the results of the two independent detectors if some conflicts exist. Another is a kind of post-processing that determines if the reported candidate of boundary is reasonable.

The R-frame selector selects one or several frames that can stand for the contents of the shot. The number of selected R-frames is determined by analyzing the contents based on the pixel difference and the information provided by the gradual boundary detector as shown in Fig. 4.

3.1 Cut Detector

As described in section 2, the features that are proposed by many researchers have weaknesses in different situations of video contents. Hence by combining them appropriately, it is expected to detect shot boundaries more accurately. In selecting the features for the combination, not only the effectiveness in shot boundary detection but also the processing time for calculating each feature should be considered.

In this paper, pixel difference, histogram difference and a variance of difference histogram those are defined as the following respectively, are used for the combination.

1) pixel difference[4][5]

$$D_p = \frac{\Sigma(I_m(x,y) - I_{m+1}(x,y))}{N} \qquad (1)$$

(Here, $I_m(x, y)$ is the pixel intensity of m-th frame at coordinate (x, y), N is the total number of pixels in a frame)

2) histogram difference[4][5]

$$D_h = \frac{\sum_{i=1}^{K} |H_m(i) - H_{m+1}(i)|}{K} \qquad (2)$$

(Here H_m is the histogram of m-th frame, K is the range of pixel values)

3) a variance of difference histogram[6]

$$D = \frac{\sum_{i=1}^{K} (H_{m,m+1}(i) - H)^2}{K} \qquad (3)$$

(Here, $H_{m,m+1}$ is the difference histogram calculated between m-th and (m+1)-th frame, K is the range of the difference values, and H is the mean of $H_{m,m+1}$).

In Fig. 5, the structure of the adopted neural network is shown. The input layer has three neurons that correspond to the above three features, one hidden layer of three neurons, and the output layer has two neurons that correspond to the shot boundary and continuing frames respectively.

For the training of the neural network, a news video of approximately 1,500 frames is used. This data contains no gradual boundaries but cuts. The number of cuts is 32. To prohibit the preponderance of training, the features for cuts are repeatedly applied to the network 50 times.

3.2 Gradual Boundary Detector

Among gradual boundaries, only the ones where a solid color frame passes by as in fading or curtaining effect are considered here. On the contrary to the cut detection that focus on the relation between two consecutive frames, the status of the individual frame is observed to watch if the sequence reaches a solid color frame. As a metric of the observation, a pixel variance of the

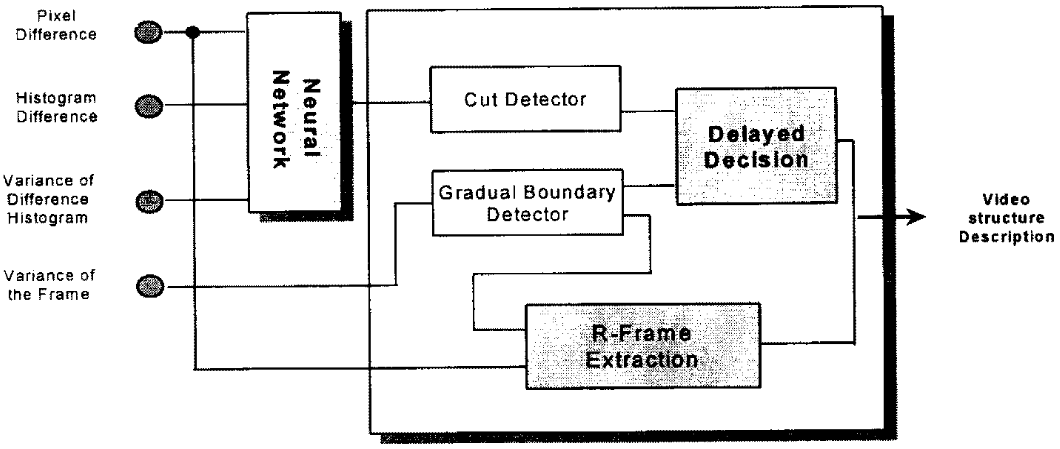

Fig. 4. The Proposed Algorithm

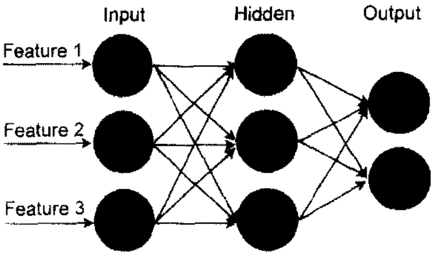

Fig. 5 Configuration of Adopted Neural Network

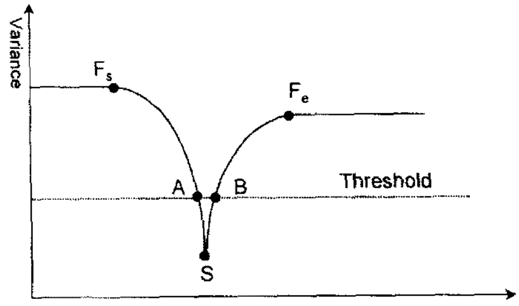

Fig. 6. Detecting Gradual Shot Boundary

frame image is adopted, which can be easily calculated from the histogram already calculated by the cut detector by the following formula.

$$V = \frac{\sum_{i=1}^{K} H(i) \bullet (i - M)^2}{N} \qquad (4)$$

(where, H is the intensity histogram of the image, K is the range of pixel values, N is the total number of pixels and M is the average intensity value of the pixels which also can be easily calculated from the histogram H).

When a sequence is fading out, the pixel variance may decrease rapidly and may increase again if fading in starts as shown in Fig. 6. The purpose is to declare shot boundary at the point S in the figure. For this purpose, once the variance goes down under the pre-determined threshold (shown in dotted line), the subsequent variance are saved in an array until the variance goes up over the threshold again. That is, the variances from the point A to the point B are saved in an array. And the minimum value in the array is searched to find the point S.

3.3 Delayed Decision

The most important role of this module is to check the validity of the detected shot. Fig. 7 shows an example of a sudden brightness change that is caused by the camera flashes of the press reporters. In this case, the cut detector naturally reports two consecutive frames as boundary candidates, which result in one frame length shot. To avoid this kind of false detection, the delayed decision module checks the length of the shot candidates. If the length is shorter than a pre-determined threshold, delayed decision module commands the cut detector to check if the prior and the posterior frame of the shot (the 1st and 3rd frame of Fig. 7) have sufficient difference to declare a cut. If it is the case, the frames are considered as a kind of gradual boundary and only one boundary is declared eliminating the peculiar small length shot. Otherwise, the reported shot boundaries are ignored (in case of Fig. 7, the boundaries are ignored). With this additional process, a lot of unreasonable shot boundaries could be eliminated.

3.4 Selection of R-frames

As described earlier, motion information could be of great help in selecting meaningful frames. Fig. 8 can be a good example. In the sequence of frames of the figure, a camera is very quickly zooming in to capture the face of the man. In this sequence, it is reasonable to think that the quickly

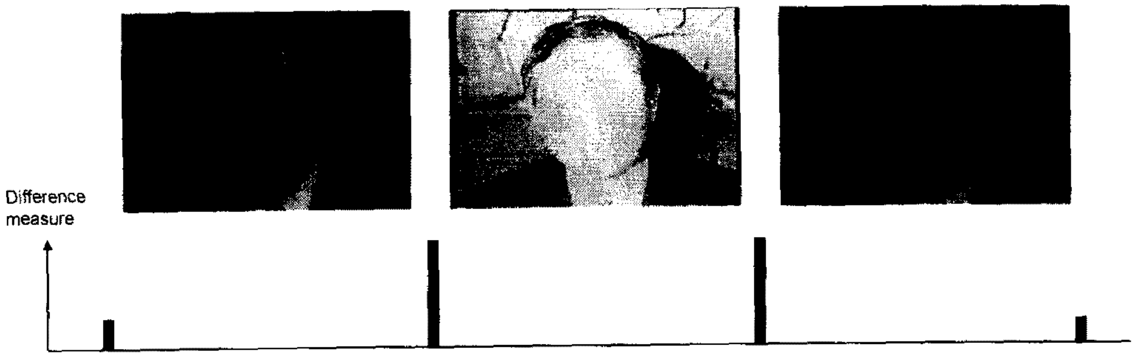

Fig. 7. False Detection due to Sudden Brightness Change

moving frames are only transient status and the finalized status is the theme of the shot. If motion information is available, the final frame can be selected as an R-frame.

However general motion analysis requires a lot of processing time. In the proposed algorithm, the pixel differences that are known to be sensitive to motion is used. Especially because the pixel difference is already calculated for the cut detection, this information is available without additional computation.

The proposed algorithm classifies the changing pattern of the pixel differences into three types of Fig. 9 and select R-frames in the following manner for each pattern. In the figure, the black dots show where the R-frame is selected.

Pattern A : This type is a very static shot, that any frame can stand for the contents. In this case, the center frame of the shot is selected as an R-frame.

Pattern B :The start and end of the transient status is selected as R-frames.

Pattern C : This case has no special clue to select an R-frame. In this case, a fixed period approach is adopted.

Also fading should be considered for the reasonable R-frame selection because it is not desirable to select the frames in the process of fading in or fading out. For this reason, the frames between the point Fs and Fe of Fig.6 are excluded from the target of R-frame selection described above.

4. Experiment

The proposed algorithm is tested on digital videos of MPEG and AVI format. The experimental data contains approximately 100,000 frames (55 minutes) in total. There are approximately 800 cuts and 30 gradual boundaries in the experimental data.

For the comparison, the algorithms using the following features are implemented.

1) A pixel difference,
2) A histogram difference,
3) A variance of difference histogram
4) the proposed algorithm

The algorithms are evaluated based on the following values.

$$Recall = \frac{Correct}{Correct + Missed} \quad (5)$$

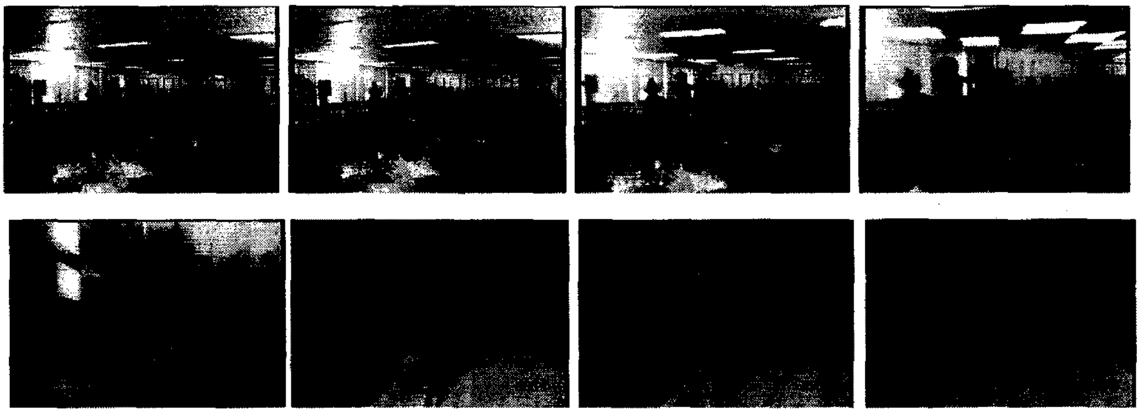

Fig. 8. Fast Zoom In

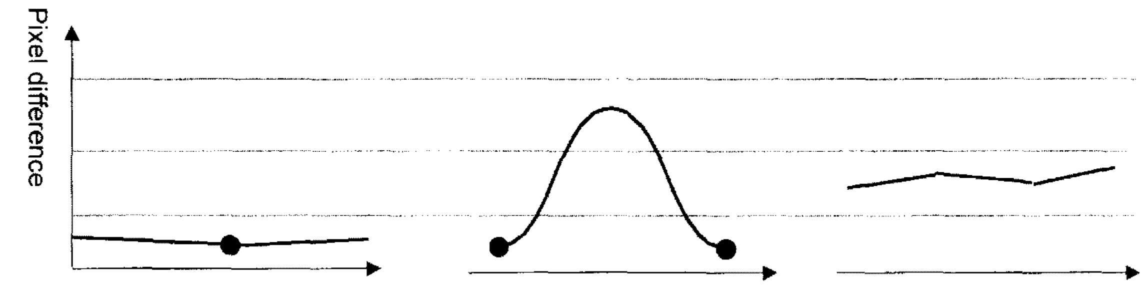

Fig. 9. Changing Pattern of Pixel Differences

Table 1. Comparison of Accuracy

	Recall	Precision
Pixel Difference	0.883	0.815
Histogram Difference	0.774	0.724
Variance of Difference Histogram	0.892	0.853
The Proposed Algorithm	0.974	0.942

$$Precision = \frac{Correct}{Correct + FalsePositive} \qquad (6)$$

The experimental result is shown in Table 1. As shown in the result, the proposed algorithm shows far better results than the cases that each feature is applied individually.

The validity of R-frame selection is very difficult to evaluate quantitatively because it is a subjective problem. However for the very static shots, the proposed algorithm selected only one frame compared to the 3 to 20 frames of the existing algorithms. Also owing to excluding the gradually changing regions, no cumbersome frames are selected when a gradual boundary exist.

5. Conclusion

In this paper, a new shot boundary detection and R-frame selection algorithm that are essential in the automation of internal structure extraction are proposed.

On the contrary to the existing shot boundary detection algorithms, the proposed algorithm adopts dual detectors that are specifically designed for detecting cuts and gradual boundaries respectively. Also it has a delayed decision mechanism that checks the validity of the detected shots, which resulted in a much improvement in overall accuracy.

Also contrary to the existing R-frame selection algorithms that adopt simple mechanical approaches like a fixed period sampling, the motion of the content is considered to select meaningful frames. The adopted criteria for the decision is a pixel difference that is already calculated in the shot boundary detection, such that almost no additional processing time is required to solve the problem.

[Reference]

[1] Liusheng Huang, John Chung-Mong Lee, Qing Li, Wei Xiong, "An Experimental Video Database Management System Based On Advanced Object-Oriented Techniques", Proc. SPIE, Vol. 2670, pp. 158-169, Feb. 1996.

[2] Di Zhong, Hong Jiang Zhang, Shih-Fu Chang, "Clustering Methods for Video Browsing and Annotation", Proc. SPIE, Vol. 2670, pp.239-246, Feb. 1996.

[3] H. J. Zhang, C. Y. Low, S. W. Smoliar, J. H. Wu, "Video Parsing, Retrieval and Browsing : An Integrated and Content-Based Solution", Proc. ACM Multimedia '95, pp. 15-24, Nov. 1995.

[4] J. S. Boreczky and L. A. Rowe, "Comparison of Video Shot Boundary Detection Techniques", IS&T/SPIE, Vol. 2670, pp.170-179, Feb. 1996.

[5] G. Ahanger, T. Little, "A Survey of Technologies for Parsing and Indexing Digital Video", Journal of Visual Communication and Image Representation, Special Issue on Digital Libraries, Vol. 7, No. 1, pp. 28-43, Mar. 1996.

[6] Jae Yeon Lee, Byung Tae Chun, Younglae J. Bae, "A research on the methods for efficient video segmentation", The 9th Workshop on Image Processing and Understanding, pp. 287-291, 1997.

[7] H. J. Zhang, S. W. Smoliar, "Developing Power Tools for Video Indexing and Retrieval", Proc. SPIE, Vol. 2185, pp. 140-149, Feb. 1994.

[8] E. Ardizzone, M. La Cascia, D. Molinelli, "Motion and Color Based Video Indexing and Retrieval", Int. Conf. on Pattern Recognition, ICPR, Wien, Austria, Aug. 1996.

[9] A. Zhang, S. Multani, "Implementation of Video Presentation in Database Systems", Proc. SPIE, Vol. 2670, pp. 228-238, Feb. 1996.

[10] M. La Cascia, E. Ardizzone, "JACOB: Just a Content-Based Query System for Video Database", Proc. ICASSP-96, pp. 1-4, 1996.

[11] E. Ardizzone, M. La Cascia, D. Molinelli, "Motion and Color Based Video Indexing and Retrieval", Int. Conf. on Pattern Recognition, ICPR, Wien, Austria, Aug. 1996.

[12] R. W. Picard, "A Video Browser that Learns by Example", MIT Media Laboratory Technical Report #383. 1995.