MVA '92 IAPR Workshop on Machine Vision Applications Dec. 7-9,1992, Tokyo

A Proposal of a Multimedia Cooperative Drama Scene Recognition System

Yoshitomo Yaginuma and Masao Sakauchi
Institute of Industrial Science, University of Tokyo
Roppongi 7-22-1, Minato-ku, Tokyo 106, Japan

Abstract

Several kinds of studies on color image analysis have been carried out, such as automatic key words extraction from color images and automatic classification of color image sequence. But, the recognition level of color images is not enough high to recognize all things from color images alone. On the other hand, the recognition of drama scene is necessary for efficient handling of color moving media, such as creation of a new scene modified from the original scene. In this paper, a multimedia cooperative drama scene recognition system using common concepts is proposed. This system employs so called fusion of the information from various media such as images and that from scenario and sound, in order to carry out higher recognition of a given drama scene. To show the effectiveness of this system, experiments were carried out using about 7 minutes TV drama scene. According to the results of the experiments, this system could successfully make a correspondence between images and scenario using common concepts, such as time, number of people, conversation pattern of two people, and duration time of the scene.

1 Introduction

Several kinds of studies on color image analysis have been carried out, such as automatic key words extraction from color images[1] and automatic classification of color image sequence[2]. But, the recognition level of color images is not enough high to recognize all things from color images alone. N.Nandhakumar combined thermal and visual images of the scene for the labeling of the image regions[3]. Fedric P.Perlant and David M.McKeown combined disparity map and monocular image to estimate three-dimensional scene structure[4]. On the other hand, the recognition of drama scene is necessary for efficient handling of color moving media, such as creation of a new scene modified from the original scene.

In this paper, a multimedia cooperative drama scene recognition system using common concepts is proposed. This system employs so called fusion of the information from various media such as images and that from scenario and sound, in order to carry out higher recognition of a given drama scene. To show the effectiveness of this system, experiments were carried out using 7 minutes TV drama scene. According to the results of the experiments, this system could successfully make a correspondence between images and scenario using common concepts, such as time, number of people, conversation pattern of two people, and duration time of the scene.

The details on the multimedia cooperative drama scene recognition system are discussed in section **2**. In section **3**, the method to make the correspondence between cuts and scene are shown. To show the effectiveness of this system, the results of the experiments are also shown in section **3**.

2 General features of multimedia cooperative recognition system

This system makes a correspondence between images and scenario using common concepts, which do not depend on the kinds of the media. This system is constructed by the Supervisor and the Analyzers(Fig.1). The Analyzers extract 'common concepts' from the media. The Supervisor exchange 'common con-

cepts' with the Analyzers, and carries out the fusion of the information extracted from the media.

For example, making correspondence between cuts and scene using 'time' is carried out as following. First, the Supervisor asks Analyzers the 'time'(Fig.1(a)). Second, the Analyzers send the 'time' information to the Supervisor(Fig.1(b)). Finally, the Supervisor fuses the 'time' information, and makes the correspondence between cuts and scene by the comparison of the 'time'.

3 Making correspondence between cuts and scene

The TV drama scene used in this experiment is about 7 minutes long. The scenario has 9 scene, and 22 cuts were extracted from the images. At first, the number of correspondence between scenes and cuts was about $100000(={}_{21}C_8)$. This system could reduce the number of correspondence between scene and cuts to 50 as shown below.

correspondence by 'time'

The cuts have the information of time. Therefore, the start time and the end time can be extracted from the cuts.

The scenario does not have the information of time. Therefore, the start time and the end time are estimated by the character number

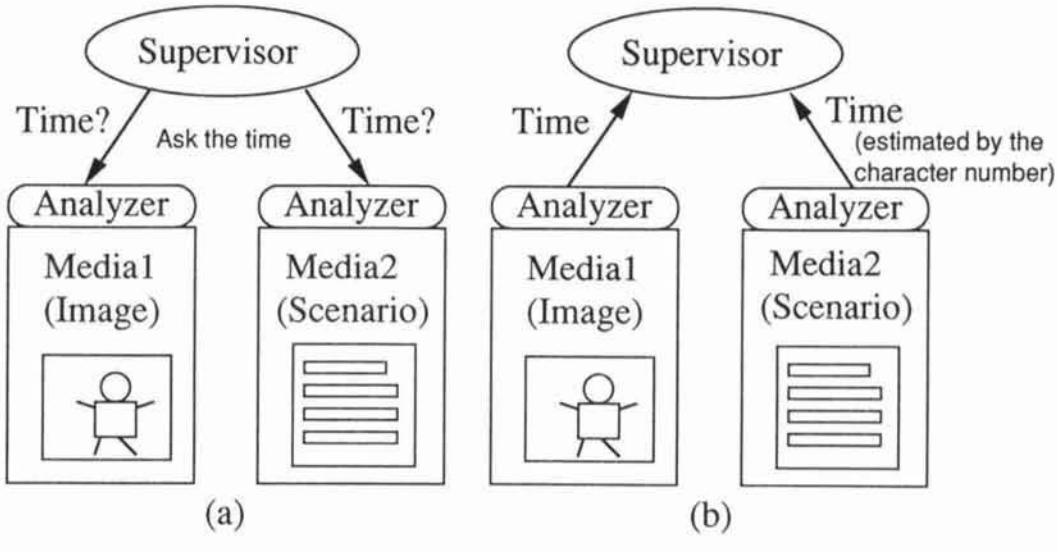

Fig.1 Cooperation using 'Common concepts'

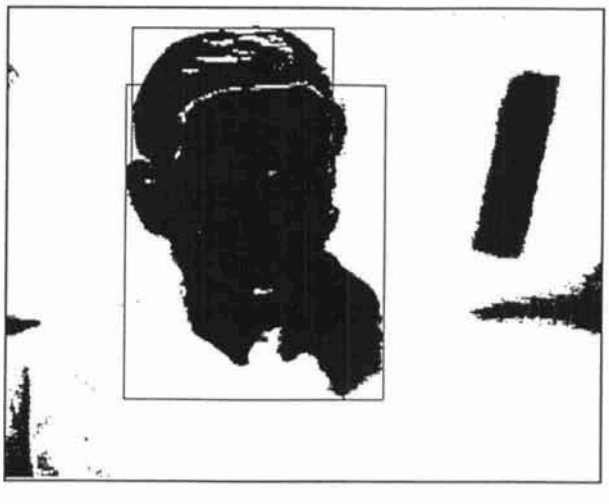

Fig.2 Extraction of human

in the scene. At first, the start time and the end time are estimated in proportion to the character number of each scene. As these time is not precise time, the start time is estimated by the subtraction of error time. Similarly, the end time is estimated by the addition of error time. (In this experiment, the error time is defined to be 20 sec.)

The correspondence between cuts and scene is roughly made by the comparison of these 'time'.

correspondence by 'human number'

The human number in the cuts is extracted by the scene analyzer. If there is no one in the cuts who is not described in the scenario, the human number extracted from the cuts is equal to or less than the human number extracted from the scenario.

The human extraction is carried out by the extraction of the face region and the hair region(Fig.2). At first, the RGB of the image is converted to HVC. The region near the red is extracted as a face region. The region with low brightness is extracted as a hair region. (This system intends to extract mainly Japanese people.) If there is a hair region above the face region, these two regions are regarded as a human. This algorithm could extract about 60% of the human.

The correspondence between cuts and scene is made by the condition that the human number extracted from the cuts are equal to or less than the human number described in the scenario.

correspondence by 'conversation'

If there is a conversation scene in the cuts, there must be a conversation scene in the scenario. As a conversation, the conversation pattern between two people is extracted(Fig.3). This conversation pattern is made by the camera change. In the cuts, the first cut and the third cut are similar, and the second cut and the fourth cut is similar. The similarity of the cuts is tested by the comparison of the color histograms of the cuts.

In the scenario, the names of people are aligned like A, B, A, B.

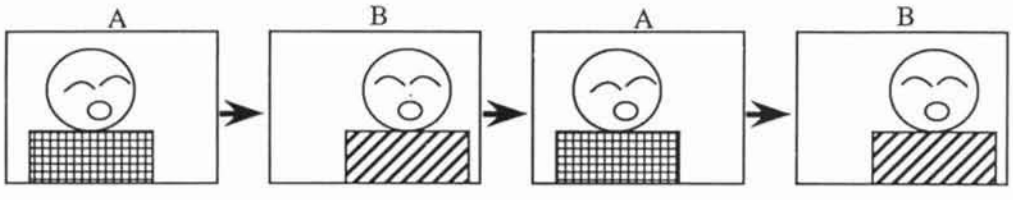

Fig.3 Conversation pattern between two people

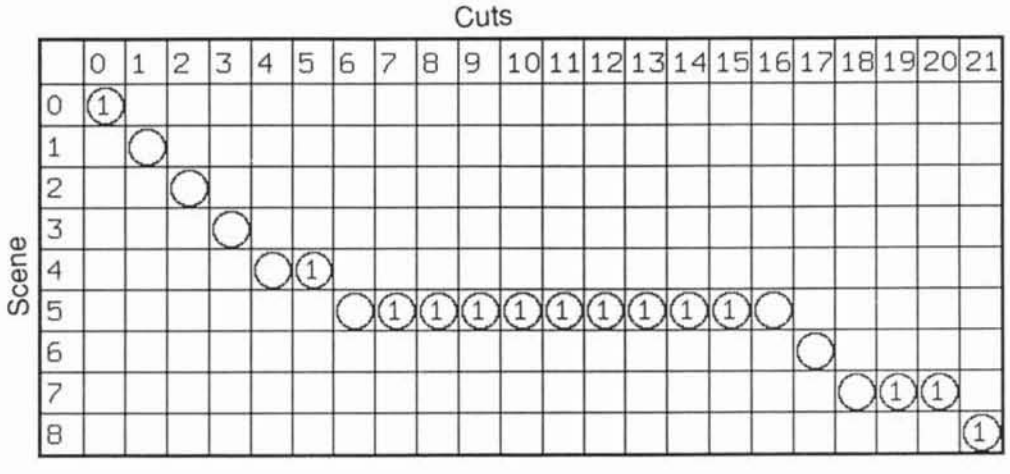

Fig.4 Results of correspondence between cuts and scene

The correspondence between cuts and scene is made by the matching of the conversation scene.

correspondence by 'duration time'
The cuts have the information of time. Therefore, the duration time of the scene can be extracted from the cuts.

The scenario does not have the information of time. Therefore, the minimal duration time is estimated by the character number in the scene. The minimal duration time is estimated by (character number) $\times$ (time needed to speak one character). In this experiment, the time needed to speak one character is defined to be 50msec.

The minimal duration time estimated by the scenario must be equal to or less than the duration time extracted from the cuts.

The correspondence between cuts and scene is made by the comparison of these 'duration time'.

The final results are shown in Fig.4. There remain 50 candidates as the matching results between cuts and scene. The ○'s show the true correspondence. The 1's show the recognition results which are determined identically. The 14 cuts out of 22 cuts are determined identically using this system.

4 Conclusion

In this paper, a multimedia cooperative drama scene recognition system using common concepts was proposed. This system fuses the information from various media such as images, scenario and sound using common concepts, and carries out higher recognition of the drama scene. Experiments were carried out to show the effectiveness of this system, and it was shown that this system could make a correspondence between images and scenario using common concepts, such as time, number of people, conversation pattern of two people, and duration time of the scene. It remains as a future work to find what can be done with this drama scene recognition system, and how to realize the fusion of more kinds of media.

References

[1] M.Sakauchi and J.Yamane: "Realization of fully automated keyword extraction in image database system", SPIE International Symposium on Optical Applied Science and Engineering (1992)

[2] Y.Gong and M.Sakauchi: "A Method for Color Moving Image Classification Using the Color and Motion Features of Moving Images", ICARCV'92 (1992)

[3] N.Nandhakumar : "A Phenomenological Approach to Multisource Data Integration: Analyzing Infrared and Visible Data", NASA Conference Publication 3099, pp.61-73 (1991)

[4] Fredric P.Perlant and David M.Mckeown : "Improved Disparity Map Analysis Through the Fusion of Monocular Image Segmentation", NASA Conference Publication 3099, pp.83-97 (1991)