MULTI CAMERA AUTOMATIC VIDEO EDITING

Stanislav Sumec
Brno University of Technolohy, Božetěchova 2, 612 66 Brno, Czech Republic

Abstract: Current technology makes possible to record various events of a human live, such as meetings simultaneously with several video cameras. Large amount of data is obtained from the each recorded event. However, a problem with presentation of such data in a suitable way occurs. This paper describes an algorithm that can be used in a compact videos generation from several source video streams according to different aspects and requirements.

Key words: video editing; meeting room; multi view; activity evaluation; technical and aesthetical aspects; editing rules.

1. INTRODUCTION

Meetings are an integral part of a human's everyday life. Sometimes, the meetings are informal and their content does not need not to be remembered. But many events exist, which have to be preserved for a future processing. A traditional solution of this problem is manual transcription of the meeting. However, this solution needs a lot of human work and it is time consuming. Other problem is that the transcription does not accurately represent what is happening in the meeting room. Finding of the desired information is also difficult. Fortunately, a modern technology can help. Contents of the meeting can be recorded using one or more video cameras and microphones and then it can be stored in a digital library. Electronics sensors and expansion of communications make also possible to have the meeting for a long distance. But various new problems with obtained data may occur. Even if data analysis and searching of meeting database is omitted, the problem of data presenting by an acceptable way still remains, especially if more cameras is used in the meeting room to recording pictures from different views. Acceptable solution for the viewer may be watching

K. Wojciechowski et al. (eds.), Computer Vision and Graphics, 935–945.

program generated from the meeting similarly to the ordinary television programs. This means that only picture of one camera or blended picture of several cameras is chosen and shown at the each moment of the meeting. The camera selection has to respect various aspects, which guarantee accuracy and interest of the presented information. This can be called an automatic video editing. The same technique can be also used for reduction of dataflow in teleconferences, where several cameras are placed in the meeting rooms. Only one picture of the selected camera can be transmitted to remote participants of the meeting, thereby saving the line bandwidth.

Various works[1-5] focused on the automatic video editing have been already presented. But most of these works had different object of interest. Their goal was usually a summarizing of the meeting recorded with only one camera or other events recorded with more cameras but these events have been already edited e.g. television news or discussions. Other works[6] are interested in the automatic video editing of home video recordings. Different methods, such as scene segmentation, camera motion analysis, speech analysis, shot detection, etc., are applied in these works.

This paper presets a current state of work aimed at a design of an algorithm providing the automatic video editing of the meetings recorded by several cameras. Results of designed algorithm should be compact programs, which respect various aspects both technical and aesthetical. The editing should be also adjusted to desired information. This means that viewer could determine which person or happening in the meeting room is preferred to be included in the generated program. Designed algorithm is tested on a multimodal meeting corpus[7] recorded in IDIAP. This corpus contains the meetings recorded in the meeting room with three fixed cameras. Transcriptions of the meetings are also available. Figure 1 shows setup of the meeting room and pictures obtained from the cameras.

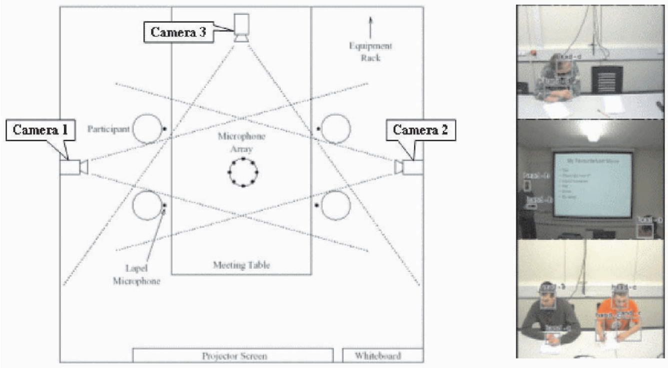

Figure 1. Meeting room setup and pictures from the cameras.

2. ALGORITHM

The function of the proposed algorithm can be formulated as a problem of one camera or of several combined cameras selection in each time point of the recorded meeting. The image from the selected camera has to preferably represent what is happening in the meeting room according to different (user specified) aspects. The first reflected aspects should be technical. Satisfaction of these aspects warrants that produced video contains as much of the relevant information as possible. For example, active speakers or gesturing participants will preferably be shown. However, experiments have shown that satisfaction of the technical aspects is not enough to produce "a good" output video. For example, problems may occur during a discussion of several participants because the cameras can be switched too fast or, on the contrary, during a monologue of one participant when the camera can be focused long time to the speaking person; therefore, some aesthetical aspects have to be included in the video editing algorithm to eliminate of these problems.

The main idea of the proposed solution is that a methodology of the video editing can be described through a set of various rules. An application of these rules frame by frame to the whole meeting produces a scenario that can be used for generation of the final video. The function of the designed rules should as good as possible model work of a human editor. The goal of the work is to design such rules and the methodology how to "put these rules together" and create the automatic video editing algorithm for the meetings recorded with several cameras.

The rules can be divided into two basic classes according to the information, that can be processed in the rule. The rules of first type (called A) can use only the "past" data – the data that are obtained from the events which occur before the time point being processed. For the second type rules (called B) it is possible to use data from the whole meeting. If the algorithm uses only the first type rules, it can be applied for live video editing e.g. in live broadcasting of the meetings or in teleconferencing. The rules of second type can be useful for offline video editing e.g. in digital meeting library; it is clear that first (A) type rules can be also utilized for this purpose but better result can be achieved with the second one type rules because editing can reflect the "future" events. The goal of the rules application is assignment of weight to every camera in the meeting room. After the weight of all the cameras is known, the camera with the highest weight is selected.

Technical aspects of video editing are mainly represented using so-called additive camera rules. These rules evaluate a measure of interest of the events on every camera. Resulting weights of the additive rules evaluated for given camera are summed up so the aggregated weight describes the "interest measure" of the given camera. The additive camera rules evaluate e.g. activity of meeting participants or other interesting things such as slides projecting; however, additional rules can also guarantee some aesthetical aspects (as presented later). Other set of rules contains so-called multiplicative camera rules. These rules can be used for suppression or stimulation of the some camera weight. Their application is important for satisfaction of the aesthetical aspects of video editing. Aggregated weight of the given camera, computed by the additive camera rules, is further multiplied by all of the multiplicative rules to obtain the final weight of the given camera. The whole algorithm works in the following way: 1) Source video streams of all cameras are simultaneously processed frame by frame from the beginning to the end of the meeting. 2) Additive and multiplicative rules are used for weight evaluation of every camera in given time point. 3) The image from the camera with the highest weight in given time point is selected and presented as output in the given time. If the largest weight is common for more cameras, then the camera can be selected randomly from these cameras. Figure 2 shows a block diagram of the proposed algorithm.

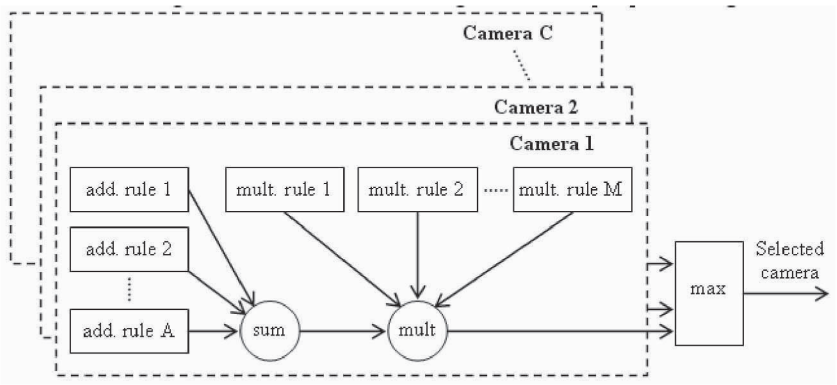

Figure 2. Video editing algorithm.

One of the additive camera rules is dedicated to activity evaluation of meeting participants. The meaning of the word activity is a measure of interest (importance). This rule uses other so-called person rules for evaluation of several aspects of human activity. Each of these rules gives weight according to some aspect of human activity. The activity of one person is evaluated so that all possible person rules are applied to a given person and a maximum weight or a scaled sum of the weights is treated as the person weight. The resulting weight of the described additive camera rule is computed as sum of the activity evaluated for every person multiplied by the visibility of given person on certain camera. Figure 3 shows a block diagram of how the activity of the meeting participants on given camera is evaluated. Persons' visibility is computed from the position of its head. Skin color detection is used to finding participant's head and hands positions. Detected objects on all cameras are labeled by participant's identification so the position of every participants head is known. The visibility is computed so that the participant with his/her head placed on higher position is better visible because if the head is higher it can be assumed that bigger part of the person is visible and also if the head is placed near to the middle of the image in direction of X axis the participant is usually better visible. Figure 3 also shows one possible visibility function. Each pixel represents the head in the corresponding position on the source image and its brightness determines the visibility of the head in this position. Brighter pixels represent positions of heads with better visibility and those darker represent heads, which are visible worse. The visibility function is evaluated from the average position obtained from several consequent frames.

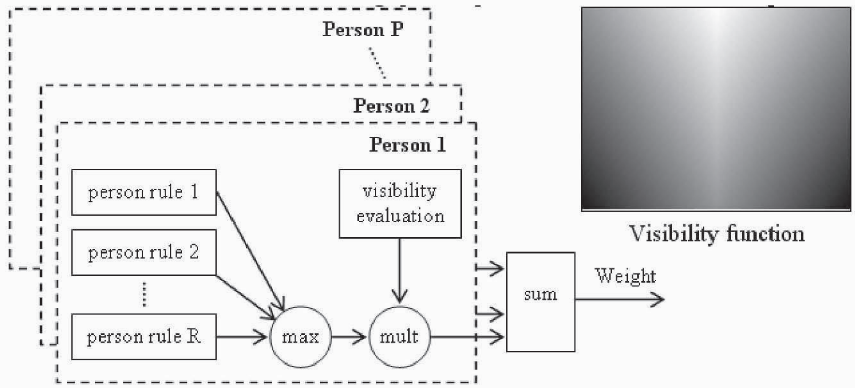

Figure 3. Additive rule for person activity evaluation and visibility function.

The first significant aspect of the participant's activity is the information about whether the participant is speaking or not. The source data for these rules is obtained from meeting transcription or, as it is planed in the future, from automatic speaker identification. The two following premises were supposed in a design of the speaking rules: More important person is that person who starts speaking the first, also more important is that person who is speaking longer. These premises follow from requirements to select camera with the participant before this participant starts speaking and preferring of long speaking participants. Various functions can be used to model speaking rules. Figure 4 shows examples of the tested speaking rules. The speaker activity is represented by two rules on graphs a) and b). The "past rules" are type A rules and can be used in live video editing because it is possible to evaluate its value from sooner data. The "future rules" and "future past rules" from graph c) and d) can only be applied in the offline editing because it is necessary to know when the participant starts speaking in the future. The best results with subjective measures were achieved with the exponential rules, which are presented at the graph a). But more precise evaluation is planed to determine shape and parameters of the speaking rules.

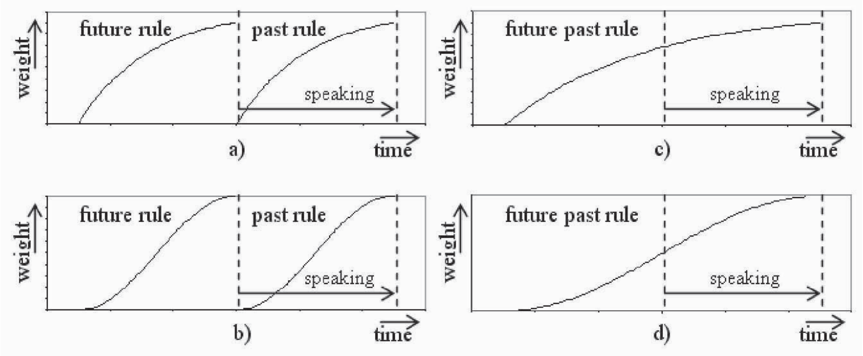

Figure 4. Modeling of speaking rules.

Others person rules describe the aspects of participant's physical activity. Detected objects with participants head and hands are used as source data for these rules. Motion of the head and the hands are described using two separate rules. One assumption was used during the design of these rules such that it is more important to show particular participant if he is gesturing by his head or hands. This activity evaluation is based on participant's velocity because it can be assumed that if participant is gesturing, the velocity of his body parts increases[8]. The velocity of the given object can be estimated from object position differences between two following frames or time window which gives better results for the activity measurement. If the velocity is computed from differences of several following frames a noise is reduced and an obtained value represents participant's activity during longer time period. The position of time window with regard to the evaluated time determines the type of obtained rule. If only differences of previous frames are used, the resulting rule has type A otherwise the rule type is B. A threshold can be used to a restriction of minimal activity.

As mentioned earlier, the additive camera rules can be also used to simulate some aesthetical aspects. Currently the rule handling periodic alternating of cameras is designed as additive rule. Its function is to add a little weight to cameras, which were not selected during long time period. This cause changes of cameras if no other activity is detected. Other additive camera rule can be used for example to simulate random activity.

Others aesthetical aspects of video editing are provided by the multiplicative rules. Some premises were determined for their design. At first, if certain camera is selected, it has to be selected at least for minimum given time period. This avoids quick camera changes that are not acceptable for the viewer. Next premise says that one camera can be selected at most for a given time period to guarantee an interest of the produced video. Two basic multiplicative rules were established for satisfaction of these requirements. The first one called anti-quick is designed so that it suppresses all cameras except an actually selected. Rule weight is zero and suppression is total till given time and then it is gradually reduced. Figure 5a shows shape of this rule. Two significant constants determining duration of total suppression t_1 and time of almost minimal suppression t_2 have to be defined. Similarly to this rule, the second multiplicative rule called anti-lazy rule works. However, application of this rule suppresses the weight of the actually selected camera. The rule does not suppress the selected camera for a given time period. After the time elapses, this camera is suppressed "more and more". The result of the rule application is that too long shots will be shortened. Figure 5b presents graph of the anti-lazy rule.

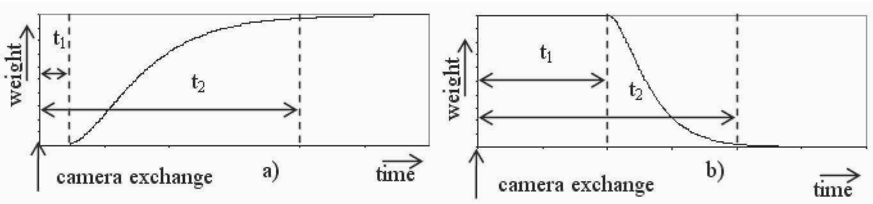

Figure 5. Shape of anti-quick and anti-lazy rule.

The proposed algorithm was currently implemented in Prolog. Source data is represented as facts with time stamps and the rules are implemented as clauses. Big advantage of the implementation in this language is an easy extraction of source data from the database and a possibility of using backtracking mechanism. Evaluation of this experimental implementation is described in section 4.

3. ENHANCEMENTS

The proposed algorithm can be further extended. The first possibility is adding of new rules. E.g., if new aspects of the video editing are obtained, the corresponding rules can be added and the rest of the algorithm can remain unchanged. Another possible extension is implementation of new person rules describing high level participant’s activity. Dialog pair detection and participant’s head orientation will be also included.

Other extension of the algorithm is its adjustability according to the desired information. This can be easily done by an amplification of the selected rule weight. The viewer can specify what activity of which meeting participant is preferred; e.g., the viewer always wants to see participant C when he is speaking. The weight of the corresponding rule can be amplified, which causes a preferring of the given participant in the produced output.

A summarization of the meeting can be done, too. This means that the viewer can specify only which activity he wants to see and also a maximum acceptable length of the produced program. The algorithm then works so that if a desired activity is not detected in the given time, the pictures of cameras in this time are not included in the produced program. The length of the program can be restricted, too. If the produced program is longer than is acceptable the algorithm can be applied repeatedly so that the pictures from the cameras in the time are not included when the desired activity is less than the threshold. If the length is still too much, the threshold is increased and the procedure is repeated as many times as needed till the desired length is obtained.

All of the cameras in the experimental meeting room are fixed and for the whole time they “watch” the same view. But the viewers are accustomed from television to focusing of the cameras when somebody’s activity should be emphasized. This effect can be replaced by zooming of the camera pictures. But maximum possible zoom is limited by the resolution of the cameras. However, as experiments showed, double zoom is sufficient for great improvement of produced program. The proposed algorithm supports so-called virtual cameras. The virtual camera is defined as a zoomed slice of the physical camera picture. Center of the slice and the selected zoom determines the picture of the virtual camera. The virtual cameras can be preset fixed to seat positions of meeting participants or can trace every person in the meeting room. The weight of virtual cameras is computed as well as of physical cameras. But visibility of persons is computed only from the slice on corresponding physical camera. If the virtual camera is evaluated as camera with the highest weight, it can be directly selected. But more authentic, especially if previous selected physical camera is corresponding to the selected virtual camera, is usage of an incremental zooming and changing of a virtual camera center. The obtained result is an image that looks like from the focusing physical camera. Several camera rules are used to support of these properties.

4. EVALUATION

Evaluation of the results of the algorithm is really difficult because it depends on the viewer's subjective opinion. The aesthetical aspects can be the best evaluated by humans. But this evaluation can be time consuming and such examination needs a lot of people. Evaluation of all the available meetings in the corpus is planed later, when the development of the algorithm will be finished. Other way is application of methods for machine evaluation of aesthetic measures with uses e.g. Birkhoff's aesthetic measure[9] or some probability or entropy based approaches. But some suitable methodology for aesthetical measurement of generated video files has to be at first designed. However, some experiments have shown that chosen methodology works and that the produced programs are acceptable for the viewers. While evaluation of the aesthetical aspects is problematic, the technical aspects can be evaluated well. Each technical aspect can be evaluated so that the whole produced program is compared if the camera with the person that meets an examined aspect is selected in given time. E.g., showing of speaking participants can be evaluated this way. Such experiments have been done to examine whether the produced program satisfies the technical aspects. Few hours of meetings recorded with three cameras were automatically edited using live editing rules and offline editing rules. Two virtual cameras were additionally used for each physical camera. Periods, in which someone was speaking and this person was or not visible in the generated program, were measured. The percentage of successfully selected cameras according to the technical aspects can be computed from the obtained results. Further, the randomly edited programs were evaluated the same way. The aesthetic rules were included in all experiments, too, but they were not examined primary. Their usage is essential e.g. in randomly edited programs because plain random switching of camera in each frame produces program that is visually unacceptable for the viewer. So the aesthetical rules were rather used in all experiments for better confrontation of the results. Table 1 shows the obtained results. Each number represents a ratio of number of frames in which the speaking participants are visible on generated program according to the total number of frames in which somebody is speaking.

Table 1. Percentage of successfully selected cameras.

	Random	**Live**	**Offline**
Physical cameras only	44,50%	75,60%	72,36%
With virtual cameras	44,28%	76,22%	74,96%

As it can be seen, the programs generated using the person's activity evaluation in the live and offline editing well satisfy the technical aspect of showing speaking. The obtained results would be better if no other technical aspects were used. However, physical activity of the participants was measured, too, because programs produced this way are more interesting for the viewer. Further feature that can be seen is that the percentage of the live editing is a little higher than of the offline editing. This can be caused by the fact, that in the live editing is the camera focused to the speaking person after a moment in which the participant starts speaking, but in offline editing, the camera is switched before the speaking begins and so other actually speaking participant could not be visible for a while; but offline generated programs will be probably evaluated as visually better then the live generated programs.

5. CONCLUSION AND FUTURE WORK

The algorithm for automatic video editing of meetings recorded by several cameras was proposed. This algorithm stands on an idea that the video editing can be described using several rules. Possibility of its applications and some its enhancements were described. The experiments we have carried out show that chosen methodology is feasible.

An examination of designed rules parameters and including of new rules especially person activity rules is planed as a future work. Some of the high level information, such as dialog pairs and more precise recognized gestures will be used in the future algorithm. A methodology for evaluation of the aesthetical aspects of output videos should be also designed. Finally, wide examination of the algorithm results is planed, too.

ACKNOWLEDGEMENTS

The paper has been founded by EU IST Program project IST-2001-34485[10]. It is part of CPA-2: the Cross Programme Action on Multimodal and Multisensorial Dialogue Modes.

REFERENCES

1. D. M. Russell, A Design Pattern-based Video Summarization Technique: Moving from Low-level Signals to High-level Structure, In Proceedings of the 33rd Hawaii International Conference on System Sciences, 2000, p. 3048.

2. L. He, E. Sanocki, A. Gupta, J. Grudin, Auto-Summarization of Audio-Video Presentations, Microsoft Research (MSR-TR-99-22), Redmond, USA, 1999.
3. A. Waibel, M. Bett, F. Metze, K. Ries, T. Schaaf, T. Schultz, H. Soltau, H. Yu, K. Zechner, Advances in Automatic Meeting Record Creation and Access, In Proceedings ICASSP, Salt Lake City, USA, 2001.
4. M. A. Smith, T. Kanade, Video Skimming for Quick Browsing based on Audio and Image Characterization, School of Computer Science, Carnegie Mellon University, Pittsburg, Technical Report CMU-CS-95-186, 1995.
5. L. Xu, J. Zhu, F. Stentiford, Video Summarization and Semantics Editing Tool, In Storage and Retrieval for Media Databases, Proc SPIE Vol. 4315, San Jose, 2000, pp. 242-252.
6. A. Girgensohn, J. Boreczky, P. Chiu, J. Doherty, J. Foote, G. Golovchinsky, S. Uchihashi, L. Wilcox, A Semi-automatic Approach to Home Video Editing, In Proceedings of UIST '00, ACM Press, November 5th, 2000, pp. 81-89.
7. Multimodal Media File Server; http://mmm.idiap.ch.
8. I. Potucek, S. Sumec, Participant Activity Detection by Hands and Face Movement Tracking in the Meeting Room, In Proceedings of CGI 2004, IEEE Press, June 2004, pp. 632-635.
9. G. D. Birkhoff, Aesthetic Measures, Cambridge Massachusetts' University Press, 1933.
10. The MultiModal Meeting Manager Project; http://www.dcs.shef.ac.uk/spandh/projects/m4.