AUTOMATIC MOBILE MEETING ROOM

[1]*Stanislav Sumec,* [2]*Igor Potúček,* [3]*Pavel Zemčík*

[1]sumec@fit.vutbr.cz, [2]potucek@fit.vutbr.cz, [3]zemcik@fit.vutbr.cz
Brno University of Technology, Faculty of Information Technology, Božetechova 2, Brno, Czech Republic

ABSTRACT

This paper presents methods for processing of the data recorded from meetings or meeting-like events. The traditional approach uses several video cameras to record such events. In this paper, we suggest to use a system with only one camera equipped with hyperbolic mirror, which allows capturing of a large portion of the space angle. The main advantage of this setup is its mobility, usage of no fixed mounted devices, easy installation, and low price. The acquired data need to be presented to the human in the appropriate manner, however, not in the "deformed" form obtained through the mirror. An intelligent video editing is proposed for this purpose. The ultimate goal is to show only such part of the recorded video that contains the relevant information.

***Keywords*:** meeting room, omni-directional vision, video editing, body parts detection.

1. INTRODUCTION

Meetings are important part of everyday human social life. It is generally suitable to retain information contained in the meetings for later use. Traditional approach of manual transcription is time consuming. Modern technology can help to automate meeting recording and processing. The goal is to develop suitable system for browsing, viewing, and searching information contained in meetings. Various systems, which use one or several fixed cameras, were proposed. But one common important disadvantage of these systems is their installation requirements, because cameras are usually mounted to the walls of the meeting room; also size of such systems is unacceptable for mobile application. We propose the solution for mobile meeting room with one camera with special mirror, instead of several standard cameras. The system is built from the off-the-shelf products. It is expected that the participants are located around a table s. figure 2. Such case is suitable for using of our system, which is then placed in the middle of the table. Each participant is captured from the frontal view. The processed video output from this system has panoramic view of all 360 degrees, so the recording equipment in our system consists of a standard laptop, one camera, and a set of microphones.

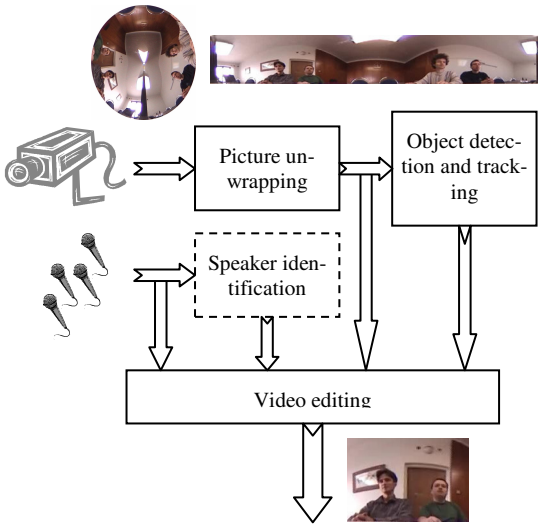

Figure 1. *System overview*

Figure 1 shows the overview of proposed system. Two types of information are used: speech

and participants behavior. We are only concerned with using information extracted from the video sequence. The whole 360 degree image is not suitable for the human observer. The main idea of our solution is to produce video that contains only selected viewports with the important events in the meeting room. We need to detect an activity in meeting room for this purpose. The behavior of meeting participants is expressed by movements, especially by hands and head movement. Other suitable information can be lip movement as complementary information to sound. The obtained data can be also used for searching the meeting databases. The output of the system is video composed from selected viewports of panoramatic view. This task is solved with an automatic video editing algorithm, which uses features detected in picture of the camera and also speaker identification data obtained from transcriptions.

2. OMNI-DIRECTIONAL SYSTEM

The system allows capturing large portion of the space angle - e.g. 120 x 360 degrees. This kind of systems is usually based on perspective camera and hyperbolic mirror on the mirror holder. Two options exist to set-up such system to use mirror above or below the camera. Much more suitable for our task is setup with mirror below the camera, because the area of the interest is mostly above the meeting table.

The output image has circular native stage and needs to be further processed. The main property of such image is non-uniform resolution in vertical axis. Mirrors are rotationally symmetric and can have different profiles. The image needs to be transformed into the panoramic view before detection and recognition. The parameters of transformation depends on the mirror profile equation.

The mirrors are usually designed with angular unit gain in the given distance from vertical axis. This property allows usage of simple unrolling algorithm. Simple transformation can caused distortion in vertical axis of the result-

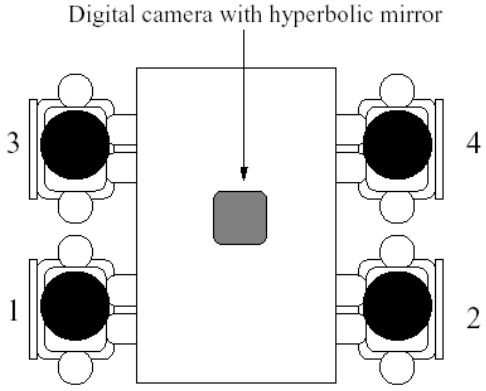

Figure 2. *Participant positions*

ing image in other cases. Therefore, the more sophisticated geometrical transformation is more suitable for mirrors, which have not unit angular gain. Both transformations use the rotationally symmetric property. The simplest transformation uses just "unwrapping" of the source image. The properties of the input captured omni-directional image which are center and radius of projected circle are known.

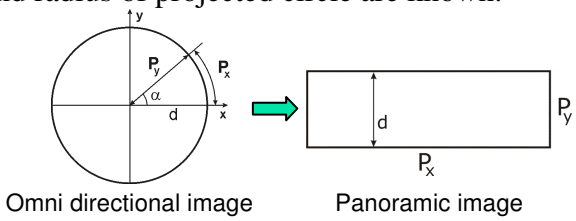

Figure 3. *Simple unrolling transformation*

The transformation of the output coordinates to coordinates of the captured image is following:

$$x_M = (d - P_y) * \cos\alpha + CenterX$$
$$y_M = (d - P_y) * \sin\alpha + CenterY \quad (1)$$

where angle is $\alpha = \frac{P_x}{d}$. Horizontal size of the panoramic view is computed as perimeter $W=2\pi d$ and vertical size is done by radius d. The main advantage of this transformation is its low computational cost.

Another approach is to use geometrical properties of the mirror for image projection on the

cylindrical plane around the major mirror axe. Due to the rotational symmetry of the system only the information about the mirror profile is used. The image formation can be expressed as

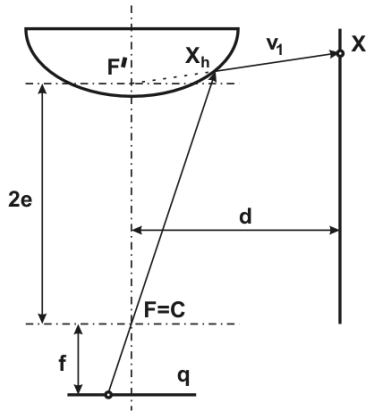

Figure 4. *Geometrical transformation*

a composition of coordinate transformations and projections [1]. We want to find the relationship between the real world co-ordinates and the camera image plane co-ordinates. The mirror coordinate system is centered at the focal point F and the hyperbolic mirror is defined by the equation:

$$\frac{(y+e)^2}{a^2} - \frac{x^2}{b^2} = 1, \qquad (2)$$

where a, b are mirror parameters and $e = \sqrt{a^2 + b^2}$ means the eccentricity. The camera center has to coincide with the second focal point of the mirror to preserve the single effective viewpoint. The geometry of the image formation of the omni-directional catadioptric camera is shown in.

3. AUTOMATIC CALIBRATION

The camera mounted to the holder with mirror is susceptible to the vibrations. These vibrations create undesirable movements in the transformed image, which is increasing in the upper border direction. Automatic detection of the image parameters can solve this problem. The output image from catadioptric system has circular shape, which is given by the mirror top view, figure 5. The transformation algorithm uses information about circle center and radius. Because input image consists of simple background and circular "omni image", it is easy to find border of the "omni image" with its properties.

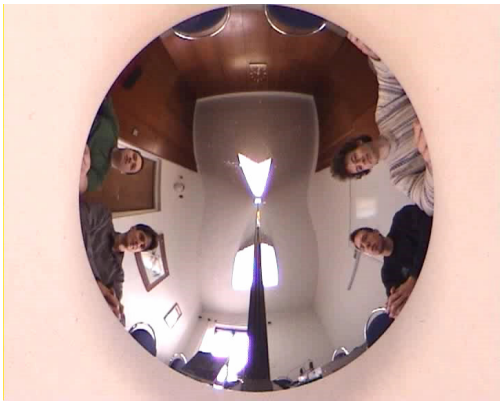

Figure 5. Captured omni-directional image

The goal is to extract pixels laying on the circle border. The first step is to use edge detector for circle border sharpening. The starting position of the detection is around the circle adjusted to the image center. The circle border is tested in the direction from outer to inner area border, because the image background contains almost no edges, see Figure 5.

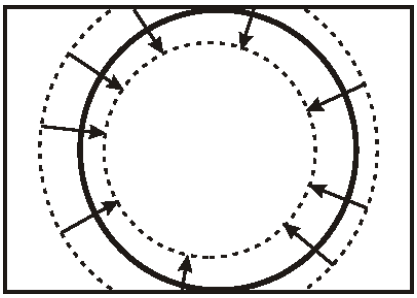

Figure 6. *Circle border detection*

The circle fitting algorithm is used for computing center and radius from pixels positions on the circle border. For accuracy improvement, the iteration method is used. The border interval is decreasing in each cycle until the number of detected points is lower than 1/5 of all possible points, which can be detected on the border. The resulting accuracy of the detected cen-

ter and radius is ±1.5 pixels, which is enough for automatic initial calibration.

4. BODY PARTS DETECTION

Human face detection, gesture recognition, and other applications for tracking of humans are based on the assumption that the areas of human skin are already detected and located. Color is the key feature for hands and face detection. The color-based methods are most often used for this task. Their main advantage is low computational cost. On a negative side, it is only a partial method because of its low reliability.

For distinguishing colors between the human skin and other image regions, the distribution of skin color must be known a priori within the normalized color space.

Figure 7. *Skin detection in transformed image*

The use of Gaussian mixture in an illumination normalized color space produced comparable results to a single Gaussian model [3]. Omni-directional system has different properties comparing to standard perspective camera. It is necessary to adapt standard methods for finding hands and faces for usage in omni-directional images. Our next goal is to analyze properties of omni-directional images, which includes creation of the color statistic depending on different lightning conditions and mirror geometry.

When only the chromaticity information is considered, a relative robustness against intensity changes is achieved. However, this will not solve all the problems related to illumination and camera calibrations: skin chromaticities depend on the prevailing illumination and camera calibration light source. The more these two lighting factors differ, the bigger shift in chromaticities. Moreover, the illumination color can be non-uniform over the face or over whole sensed area (in this case, even a proper calibration is not enough).

Skin colored blobs are extracted by using connected component analysis. Often the results of skin color detection either contain noise, or many colored objects create considerable clutter. Therefore is suitable to use morphological operators to remove noise and complete splitted objects. Each object can be described by the ellipse. Parameters are computed from object pixel positions by using ellipse fitting algorithm.

The human detected parts are then tracked and their correspondence is determined between two consecutive frames. The location prediction method is used for this task. Information about motion in previous frames serves for computation of the velocity and acceleration of tracked object. The output of body parts detection are trajectories of the detected objects over the whole video sequence.

5. VIDEO EDITING

We use an automatic video editing algorithm for output videos generation. The goal of this part of our system is to select relevant information from the whole unwrapped picture and present this information in a way that is feasible for human observer. This means that only some viewports can be displayed to the viewers. This method was chosen because a viewing of the whole omni-directional view or unwrapped picture is not habitual for humans and such videos could be difficult to watch. The video editing algorithm that was applied was primary designed for editing of recordings captured with several fixed cameras. It is, however, possible to use this algorithm for editing of omni-directional camera because it supports so called virtual cameras. Virtual cameras are defined by position and zoom of selected part

Figure 8. *Viewports placement example*

of source video so the viewports from unwrapped omni-directional picture can be easy represented as virtual cameras. Besides the fixed virtual cameras with the same position during whole recording period the moving virtual cameras are available for meeting participant following in the meeting room. Because our setup is intended for four participants we decided to use two fixed virtual cameras, which are watching opposite participants pairs, and one moving virtual camera for every participant. An example of virtual cameras placements can be seen in Figure 8. Fixed cameras represent distant shots and moving cameras provide detail shots dedicated for a highlighting of an important activity.

The video editing algorithm is based on evaluation of activities, which occur in the meeting room and also some simulation of human editor is included for the selection of the best camera and effects in given time point. Currently the features describing physical activity of meeting participants are evaluated from detected head and hands positions. Other participants' activities are deduced from the speaker identification. The problem of camera selection at given frame t from the recording with length l can be defined as function

$$\begin{aligned} c(t) &= f(t, \vec{o}, \vec{a}_1, \ldots, \vec{a}_n), \\ \vec{o} &= (c(0), \ldots, c(t-1)), \\ \vec{a}_i &= (a_i(0), \ldots, a_i(t-1)) \; or \\ \vec{a}_i &= (a_i(0), \ldots, a_i(l-1)). \end{aligned} \tag{3}$$

Vector $\vec{o}$ represents all selected camera till time point t. Measured activities are presented as vectors $\vec{a}_i$. It is supposed that camera selection process will be applied sequentially from the beginning to the end of recorded data. The video editing algorithm can be used in two basic applications. If activities' vectors contain data available only till time t, the output can be generated on the fly during the recording process and so live broadcasting of the meeting can be realized. On the other hand, the offline production of the output videos can use vectors compounded from activities computed during whole time period of the meetings. Better visual results can be achieved in the offline editing, because the camera can be also selected according to the events, which offer after the evaluated time point. Figure 9 shows the overview of the video editing part of the proposed system. Video editing algorithm is currently implemented using various rules. Some of these rules describe how to convert source features into data expressing importance of an appropriate activity. Other rules represent the video editing methodology that says what and when should be selected some of the viewports. This means that the viewport showing

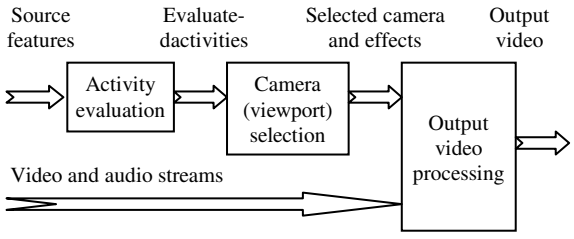

Figure 9. *Video editing*

the most important events is selected, but also the measure of desirability of the continuation of given shots is taken into account. Result of each rule is a weight representing weight of one aspect e.g. activity of one participant or importance of given camera selection according to video editing methodology. All of the rules are connected into the some kind of net-

work. The total weights describing importance of every camera are computed and the camera with the highest weight is selected. Please refer to [5] for more details. Last block of this step generate the output video stream from the source unwrapped picture and information about selected viewport and in some case about chosen visual effect.

A significant aspect of moving virtual cameras, which are mainly used in editing of the omnidirectional view, is a method for determination of their positions. We compute this position from the head positions, because these cameras should follow particular participants. However, the head positions have to be filtered to avoid an undesirable noise. Elementary filtration of high frequencies in heads motion can be used but visual results of such virtual cameras are not satisfactory. Their position is changed relatively often and sometimes only by few pixels so disruptive effects in the output video can be coused. However, the camera position should be changed fast enough, if e.g. somebody stand up and go through the meeting room. We apply an additional criterion to determine, whether the camera should move to new position. This criterion says that the virtual camera changes its position if the difference of filtered participant head position in current time point and certain future time point is bigger than some threshold. The camera position is than changed in the direction of the head position in the future. Application of this criterion is limited only on case that "future" data are available in evaluation time point. Figure 10 shows example of virtual camera position evaluation. Graph a) represent vertical positions of participant's head, filtered values shows graph b), and finally graph c) shows positions after an application of the criterion.

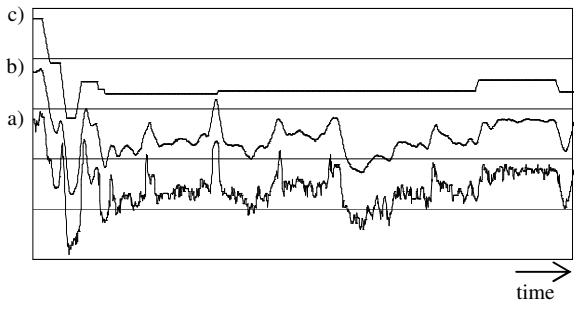

Figure 10. *Virtual camera position fitration*

An effect applied when a new viewport is selected can play an important role in visual impression of the produced video. It is possible to swap virtual cameras directly but in some cases is better to change the viewport fluently, especially if an actual viewport and new one show close parts of the source picture. Focusing of common cameras can be simulated in this way. Some interpolation of position and zoom of the viewport have to be applied. We use a linear interpolation of zoom and nonlinear interpolation of virtual camera position with respect to dimensions of an actual visible viewport. If the viewport with position X_1, Y_1 and zoom Z_1 should be transformed into viewport with parameters X_2, Y_2 and Z_2, the interpolation will be following:

$$\begin{aligned} Z &= (1-i) * Z_1 + i * Z_2 \\ j &= \frac{i * Z_2}{Z} \\ X &= (1-j) * X_1 + j * X_2 \\ Y &= (1-j) * Y_1 + j * Y_2. \end{aligned} \tag{4}$$

This method was chosen to avoid undesirable motion that occurs in the output video if both zoom and position is interpolated linearly. Figure 11a illustrates an example of the linear interpolation applied when one bigger viewport is transformed into smaller viewport with greater zoom. A motion of the bounding rectangle of the visible area is displayed. It can be seen that some pixels of source picture dramatically

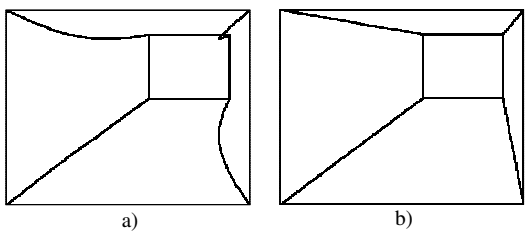

Figure 11. *Virtual camera zooming*

change the move direction in the output viewport during the effect. Figure 11b shows this motion if the nonlinear interpolation of the position is used. Such results act much more authentically.
The whole system can produce generic full length cut of the recorded meeting offline. It is also possible create video according to specific query given by the viewer. The selected participants or activities can be highlighted due to strengthening of an appropriate rule weight. Further it is possible to produce the shortened version of the meeting. A summarization method based on skipping of segments with low importance is applied. In addition video editing algorithm can be used for live broadcasting of the meetings in real-time.

6. CONCLUSION

The low cost and mobile meeting room designed for recordings and presentation of relevant information to the human was presented. The methods for image transformation, participants detection, and intelligent video editing were used for this task. The video processing is necessary when meeting events need to be presented. Image quality of such system with common digital camera is not same as with fixed mounted camera systems, but hardware requirements are much lower as the price. However, the results can be comparable with standard systems, if the HDTV camera with better resolution is used. This kind of meeting monitoring system is suitable for casual meetings at different places and conditions.

7. ACKNOWLEDGMENTS

The paper has been founded by EU-HLT Program project 506811-AMI.

8. REFERENCES

[1] Nayar, S., Baker, S., A Theory of Catadioptric Image Formation, Department of Computer Science, Columbia university, Technical Report CUCS-015-97.
[2] Svoboda, T.: Central Panoramic Cameras Design, Geometry, Egomotion, Center for Machine Perception, Faculty of Electrical Engineering, Czech Technical University, September 30, 1999.
[3] Jones, M., Rehg, J.: Statistical Color Models with Application to Skin Detection, Cambridge Research Laboratory, Computer Vision and Pattern Recognition (CVPR99), Ft.Collins, CO, 274-280, June, 1999.
[4] Martinkauppi, J., Soriano, M., Laaksonen, M., Behavior of skin color under varying illumination seen by different cameras at different color spaces, Proc. SPIE Vol. 4301 Machine Vision in Industrial Inspection IX, January 19-26, San Jose California.
[5] Sumec, S. Multi Camera Automatic Video Editing, In: Proceedings of ICCVG 2004, Warsaw, PL, 2004.