Participant activity detection by hands and face movement tracking in the meeting room

Igor Potucek and Stanislav Sumec
Brno University of Technology, Faculty of Information Technology
Bozetechova 2, Brno, 612 66, Czech Republic
Email: potucek@fit.vutbr.cz, sumec@fit.vutbr.cz

Abstract

For the purpose of Multimodal Meeting Manager Project (M4), an approach based on face and a hand tracking is proposed. The technique essentially includes skin color detection, segmentation, feature extraction and tracking detected objects. Our aim is to extract information from participant hands and face movement suitable for intelligent video editing and as additional information for speech recognition. An activity of meeting participants is evaluated for this purpose.

1. Introduction

In recent years many researchers have started to develop systems, which observe human activity for various purposes. This branch has wide usage as facilities control, smart rooms, meeting managers, videoconferencing, alarm systems etc. Two cues are frequently used: speech and body motion. We aimed for development of meeting rooms. There are usually used several video cameras, which are capturing all participants. Our system extracts information from video sequences for smart video editing and speech recognition. An important source of information is human face and hand. We can recognize individual participants when they are moving with hands or head. Hands movement serves to determination participant's activity for video editing.

We tested our algorithms on data from IDIAP. There is a meeting room for several participants with three video cameras. Two cameras are on the opposite sides and each camera takes two participants sitting at the table. Third camera is taking the board. We try to observe dependence of the hand and head moving of the talking person. The activity detection as movement of the hands and heads, tracking head looking direction is suitable information for recognition of meetings events occurring in the meetings on a higher semantic level. The paper is organized into two main parts. The first part describes used techniques for segmentation and recognition human body parts as hands and faces. There are also used methods for detecting specific face areas as eyes, nose and mouth, which can be suitable for recognizing looking direction and further processing for the speech recognition. The acquired data are then stored in the xml file for the next manipulation.

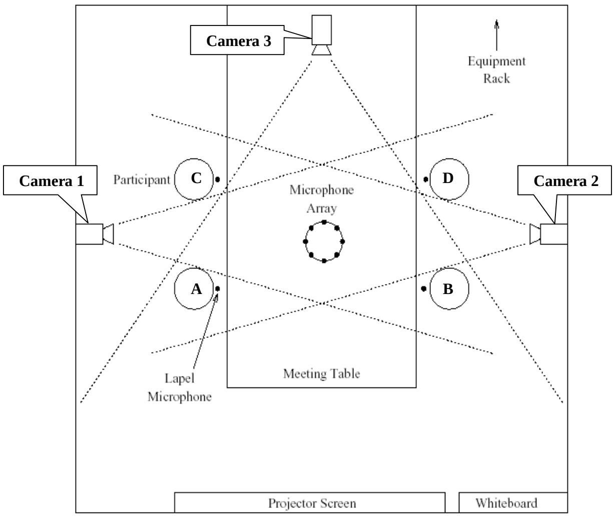

Figure 1. Meeting room setup

Second part describes method that we use to an assignment of detected object with skin color to particular participants and their parts of body as head or hands. Possibilities of participant activity evaluation are shown further. One example of practical application of obtained data is presented in an intelligent video editing algorithm.

2. Skin detection and segmentation

Color is the key feature for the face and hands detection. It is often used first stage method [3,4] of the human parts detection. Its main advantage is low computational cost. On a negative side, it is only a partial method because of its low reliability. This disadvantage don’t manifest in stable environments as meeting room. Appearance of the skin-tone color depends on the lighting conditions. Therefore we used normalized rg-color space that provides good solution to the problem of varying brightness. Normalized rg-color is computed from RGB values:

$$r = \frac{R}{R+G+B} \text{ and } g = \frac{G}{R+G+B} .$$

Various face color pixels are picked manually and then color class Ω_k is computed [3]. The color class Ω_k is determined by its mean vector m_k and the covariance matrix K_k by its distribution. We need to compute probability of each pixel in the image by this equation:

$$p(c \mid \Omega_k) = \frac{1}{2p\sqrt{\det K_k}} \exp\left(-\frac{1}{2}(c - m_k)^T K_k^{-1}(c - m_k) \right)$$

Often the results of skin color detection either contain noise, or many colored objects create considerable clutter in skin probability image and thus make skin regions not so clearly distinguishable. Therefore is used morphological operator dilatation and then erosion. The spatially separated groups of skin pixels are treated as separate objects.

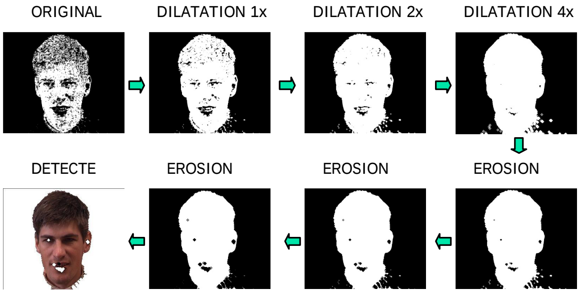

Figure 2. Object erosion and dilatation

It is used size threshold for removing noise objects. Size of the threshold is chosen according to ratio of the hands and face size to area of the whole image. The skin detection algorithm produces very good results for image segmentation for our used images because of simple background. It isn’t necessary implement another additional methods for segmentation.

3. Motion correspondence

We must keep information about objects for whole sequence of frames. The aim is to attach to an object in frame *t* its correspondent object in the frame *t+1*. We are then limited to prediction of the next object position from its previous motion. It is used Kalman filtering for estimating a new position from previous object movement. Prediction of object position is therefore based only on the positions in past frames. Thus we can define a boundary in which the searched object occurs. The boundary we use is circular and it is defined by the radius ***r***.

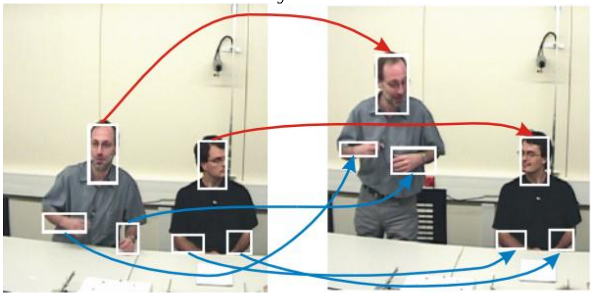

Figure 3. Object corespondence

The problem can arise when some of tracked objects disappear or appear on new position. Then we can only use face template matching or consider consistence of participants appearance in the meeting room. If this consistence is corrupted, then new person is coming. The motion correspondence keeps consistence between recognized objects in all frames. Then we can apply algorithms for recognizing individual objects.

4. High-level recognition

We are using the Gabor Wavelet method for detecting significant parts in the face as area with eyes, nose and mouth. For each person is used the set of templates for different look directions. This method is also capable to recognize single person faces by comparison the template matches.

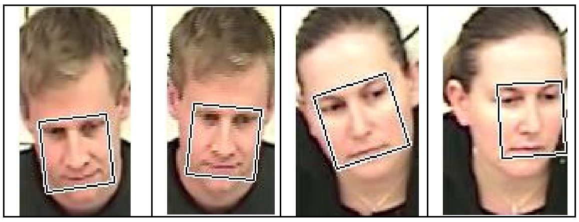

Figure 4. Face region detection

We have enough information for determine direction, angle and slope of the head. The slope in the side is computed as the slope angle of the inner rectangle covering eyes and mouth, which is parallel with vertical head axis. The look direction in vertical and horizontal direction is computed as ratio of left and right distances inner rectangle covering eyes and mouth and outer rectangle covering whole head. The additive information is from the type of the template. If it is used the template for side look, then is horizontal look direction defined constantly +90 or –90 degrees.

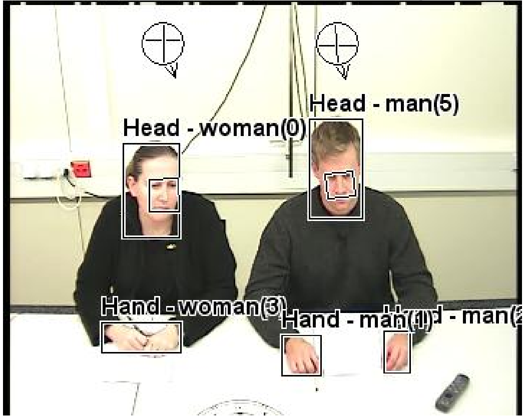

Figure 5. Looking direction

5. Object labeling

An identification of type and person has to be evaluated for every detected object. This process is called object labeling. Resulting object type can be head or face and hand. Product of person identification is an assignment of selected object to relevant meeting participant. Object labeling has to be consistent during whole meeting independently on an activity of the participants.

If each camera is evaluated separately a simple algorithm can be used to the object labeling. Known and unchangeable information about position of participant at the beginning of meeting uses this algorithm. Tracked image is divided into two same size parts with vertical line. On each side is sitting one participant and all objects belong to him. Than is supposed that the highest object is head of the participant. We can use the template matching for an improvement of this algorithm because it is possible to test if given object is really head of participant. Than two remaining lower objects are the hands.

However this algorithm does not always work. For example if one participant leaves its place and walks through meeting room the identification of its objects can be lost. Other problems can occur if two identified objects are merged together or one already identified object is divided into two separate objects. In general labeling of objects has to be evaluated for all new detected objects and for all transformed objects, which are already identified.

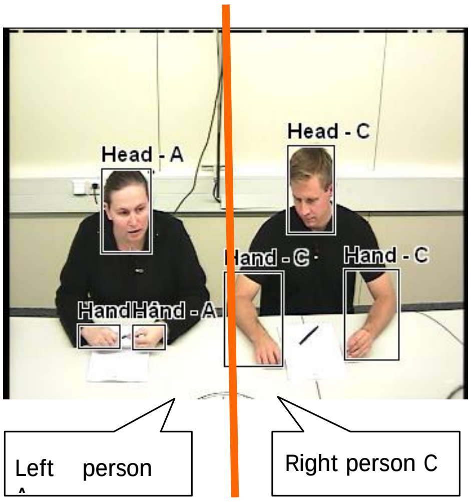

Figure 6. Labeling according to seat positions

More reliable solution of the object labeling is evaluation of all cameras simultaneously. A lot of additional information about a meeting room setup and participants can help during this computation. We use algorithm that is based on an elimination of impossible identification. At the beginning of meeting is evaluated labeling according to seat positions of participants using simple algorithm. A set of possible unused identifications in given time is evaluated for all unlabeled objects detected during the meeting. This set contains heads and hands, which are not assigned according to already identified objects on all cameras and possible number of participants. However other conditions given by meeting room setup have to be granted. For example it is clear that if one object on first camera is labeled as head of person B no object on second camera can be assigned to this person. Similar rules can be designed for relation between objects on first and third cameras or second and third camera. One of them for example says that if person is located on the left side of first camera it is impossible for this person to be on third camera at the same time. However if participant is located on the right side of first camera it can be assumed that unlabeled object on the left side of third camera belongs to the same participant. Some relations between objects on different cameras show figure 7.

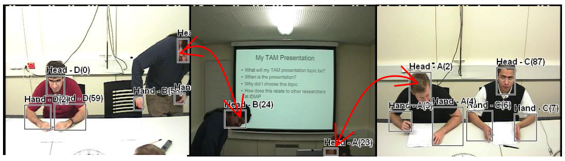

Figure 7. Relations between cameras

Set of possible identifications can be eliminated by other rules. There can be used known properties of human body as maximum distance between hands or head and hands of one person and also template matching can be used to discover if object contains face region. Resulting object identification is determined from eliminated set. If evaluated set contains only one member object is labeled by its identification. In other case when set contains more members with the same identification of participant labeling can be also easy done. If the set contains several identifications with different person assignment other rules for detection of merged or divided objects and finding of lost objects can be applied. But it is possible that using of all possible rules for set elimination does not help and set of possible identification is still to big or empty and labeling cannot be completed. In fact used algorithm works in this way. The sets of possible identification are computed for unlabeled objects in given frame and eliminating rules are applied. If at least one object is successfully labeled and other object remain unlabeled in this step process of evaluation and elimination is repeated. This is done for all frames of the meeting from its beginning to the end.

Function of described algorithms used to skin color object detection and its labeling was verified on video meeting corpus recorded in IDIAP. Some results obtained from several meeting are shown in table 1. Average number of objects with skin color occurred on all cameras during the meetings, number of detected objects and number of correctly labeled objects is shown.

Skin color objects	73155
Detected objects	66578
Detection effectivity [%]	91,01
Correctly labeled object	63587
Labeling effectivity [%]	95,51

Table 1. Experimental results

6. Evaluation of participants activity

Date about heads and hands positions of meeting participants can be used in several tasks for example in an intelligent video editing, voting detection or speaker identification. Basic information is localization of each participant at meeting room. This knowledge can be obtained from an objects assigned to given participant on all cameras. In addition it is possible to evaluate an activity of participants according to position change of their objects during the meeting. Velocity of every object in given time can be computed because its trajectory and frame rate is known. In this computation we use length of object trajectory obtained from several previous frames divided by size of this window. This technique gives better results for activity evaluation because its results represent velocity during longer time with suppressed noise. If length of object trajectory between frame *f* and frame *f-1* is *s(f)*, number of frames in window is *W* then velocity *v* in frame *f* is following.

$$v(f) = \frac{\sum_{i=f-W+1}^{f} s(i)}{W}$$

Influence of chosen window size is shown in figure 8 where velocity of one participant's hand during short time is displayed. Two peaks can be seen on shown graph. These two peaks represent time when participant greatly changed position of his hands and also his increased activity. On original video stream can be seen that participant gestured in the time relevant to peaks in graph of velocity. Threshold can be used to better distinction whether selected person is gesturing or not. It is possible to assign this threshold invariable for each type of object and camera or compute it according to average value of given object velocity for each participant on each camera.

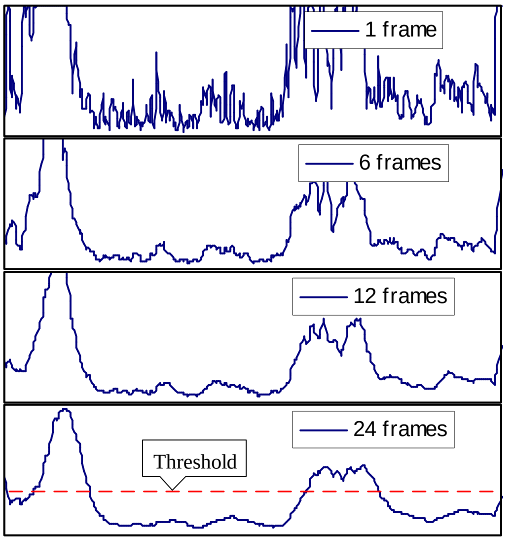

Figure 8. Influence of window size to velocity

Similar to velocity computation can be evaluated direction of object movement in given time as vector obtained from actual object position and its position on previous frame. Computation of direction during several previous frames gives better results as well as in velocity evaluation. Figure 9 shows graph of head object direction computed with window size 24 frames. The same participant and part of meeting was used as in velocity example. First graph represents x-axis of direction and second graph y-axis. Increased activity can be detected from this graph too. Participant moved his head down and then up in the same time, when maximum and minimum values of y-axis are reached. This method can be used for example in an agreement detection where direction changes of head object are significant and in a voting detection where analyzing of hand movement can be performed. If only increased activity detection is required threshold can be used.

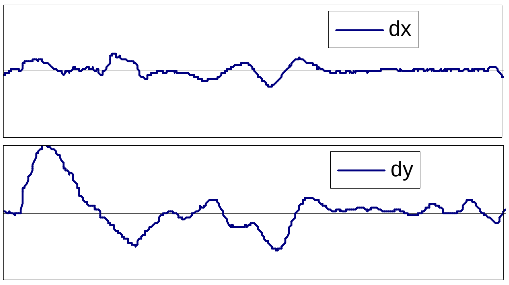

Figure 9. Direction of object movement

If template matching is used to face recognition additional information about look direction and head slope can be included in evaluation of participant's activity. For example agreement gestures can be detected from these data. It is also possible to determine other participant at who is evaluating participant looking. Example of obtained results can be seen in figure 5.

Resulting activity of meeting participant has to be compounded from results computed for all objects assigned to this participant. If head and hands objects are detected activity of person can be obtained as sum of particular object activity multiplied by relevant weights or maximal activity of participant's objects can be chosen.

Currently we use computed activity of meeting participant in an intelligent video editing algorithm. The goal of this algorithm is select in given time only one camera whose output is displayed. Selected camera should as well as possible represent happening in meeting room and whole cut should keep some rules of video editing. Velocity of objects and their position together with speaker identification are used in this algorithm. An example of data processed in video editing algorithm is shown in figure 10. Each line represents one aspect of participant activity. Lines are divided into three rows each for one camera. Each three lines represent activity of one participant. First of these lines means information whether person is speaking. The second line represents head object velocity and third one velocity of heads objects. Resulting cut is shown at fourth row.

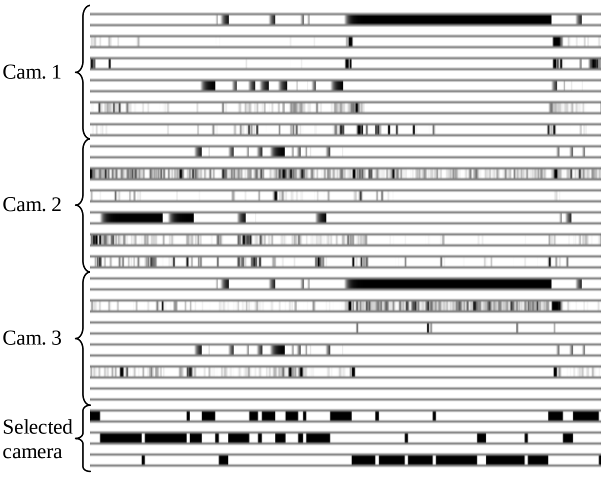

Figure 10. Map of participants activity

7. Conclusions

The paper deals with the possibilities of tracking persons in the meeting room with three cameras and about usage obtained information for the next processing. The used skin detection method is enough for image segmentation, because of the simple and unchangeable environment. Correspondence determination and face recognition by the help of Gabor Wavelet Transformation gives us enough information about participant's activity. Some methods for this measurement are presented, which use velocity and direction computation of participant's head and hands objects. Approach for estimation of the looking direction is proposed.

We plan to improve our algorithms for skin color object detection with better recognizing of human parts of body as further work. Algorithms to evaluation of activity of participants can be in addition extended to detection of some particular gestures.

Acknowledgements

The paper has been prepared as a part of the solution of EU IST Program project IST-2001-34485. It is part of CPA-2: the Cross Programme Action on Multimodal

and Multisensorial Dialogue Modes, and is linked to the activity on Human Language Technologies. For more inforamtion refer to [1].

References

[1] The MultiModal Meeting Manager (M4) Project Homepage. http://www.dcs.shef.ac.uk/spandh/projects/m4/.

[2] Jordao, L., Perrone, M., Costeira, J.,P., Santos-Victor, J., Active Face and Feature Tracking, 10th International Conference on Image Analysis and Processing, pp.572, September 1999.

[3] Jahne, B., Geisler, P., Handbook of Computer Vision and Applications – Systems and Applications, Academic Press, 1999.

[4] Jones, J., M., Rehg, J., M., Statistical Color Models with Application to Skin Detection, Cambridge Research Laboratory Technical Report Series, CRL 98/11, December 1998.

[5] Eiserloh, P.: An Introduction to Kalman Filters and Applications, Assault and Special Projects Naval Air Warfare Center, China Lake, 2002.

[6] McCowan, I., Bengio S., Gatica-Perez D., Lathoud G.: Automatic Analysis of Multimodal Group Actions in Meetings, IDIAP-RP 03-27, May 2003.

[7] Zobl M., Wallhoff F., Rigoll G.: Action in Meeting Scenarios using Global Motion Features, 4th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, March 2003.

[8] Stiefelhagen R.: Tracking Focus of Attention in meetings, In IEEE International Conference on Multimodal Interfaces, Pittsburgh 2002.