ROBUST FACE TRACKING AND PERSON ACTION RECOGNITION IN MEETINGS

Sascha Schreiber and Gerhard Rigoll

Institute for Human-Machine Communication
Munich University of Technology
Arcisstrasse 16, 80333 Munich
{schreiber,rigoll}@mmk.ei.tum.de

1 ABSTRACT

In the last few years the automatic analysis of meetings gains more and more in importance due to a growing number of meetings taking place in any part of the world. However if someone of the invited people cannot join the meeting it would be desirable for this person to have a kind of browser that enables him to get an automatically generated summary of the meeting and to watch any section of the video recording that seems to be interesting for him. In order to provide such a tool, several steps have to be performed on the videostream. In the first part of analyzing the recorded video the person has to be detected in each frame of the video. For this detection we use a tracker based on the ICondensation algorithm [1]. This algorithm is an extension of the standard condensation method and additionally uses a so-called importance function to draw samples from. In our implementation we have chosen the skin color distribution as such an importance function. Now samples can be drawn both from the importance function on the one side and from the face distribution on the other side. The great advantage of this extended algorithm is, that a track, that has been lost, can be found again by sampling from this importance function. This reinitialization is performed in every frame with a certain probability, i.e. some samples are drawn from the importance function to avoid a loss of the track, and the remaining samples are drawn from the face probability distribution with standard condensation sampling. The tracker produces an estimation of the person's location in the videoframe. With this information we now can define a rectangle around the person (including head and hands) called action region, where we suppose gestures will happen. In this region so-called global motion features [2] are calculated for each frame and we obtain one feature stream for each person, which is representing the motion in the action region. After that the BIC approach is used for the automatic temporal segmentation and finally these segments are classified by Hidden Markov Models, which have been trained with isolated gestures covering the actions "stand up","sit down","shaking head","nodding","writing" and "pointing". Applying this method to the data recorded for the M4 project, we have obtained an average recognition performance of 86%.

2 OCCLUDED GESTURES

A major problem in meetings we are opposed is partial occlusion of persons as it will occur, when one human passes in front of another. These occlusions will cause great distortions on the feature stream calculated from the video sequence and the recognition performance of gestures dramatically decreases. An approach for handling the effect of occlusions is to apply a Kalman filter [3] to the features. Hereby two different possibilities arise:

1. One general Kalman filter, which is trained for all gestures to be recognized, is used for compensating the distortions. After this filtering the feature stream is segmented and can be classified by a HMM.

2. For each gesture a specialized Kalman filter trained only on one gesture is applied. Thus we obtain n different feature streams, which are segmented as mentioned above. In a final step each stream of segmented features is given to its corresponding HMM in order to be classified.

Thereby in our approach the parameters of the Kalman filter have been trained with unoccluded data using the EM algorithm. Applying Kalman filtering to the disturbed feature stream, results have shown, that the average recognition performance can be increased from about 56.7% for occluded and unfiltered gestures up to 60.12 % corresponding to a relative improvement of about 7%.

3 REFERENCES

[1] Michael Isard and Andrew Blake, "ICONDENSATION: Unifying low-level and high-level tracking in a stochastic framework," *Lecture Notes in Computer Science*, vol. 1406, pp. 893–908, 1998.

[2] Martin Zobl, Frank Wallhoff, and Gerhard Rigoll, "Action recognition in meeting scenarios using global motion features," in *Proceedings Fourth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS-ICVS)*, Brisbane, Mar. 2003, pp. 32–36.

[3] G. Welch and G. Bishop, "An introduction to kalman filter," *UNC-CH Computer Science Technical Report*, 1995.