ACTION SEGMENTATION AND RECOGNITION IN MEETING ROOM SCENARIOS

Frank Wallhoff, Martin Zobl and Gerhard Rigoll

Munich University of Technology
Department of Electrical Engineering and Information Technology
Institute for Human-Machine-Communication
Arcisstraße 21, 80290 München, Germany
{wallhoff,zobl,rigoll}@mmk.ei.tum.de

ABSTRACT

In this proposal a novel implementation to find and recognize person actions in image sequences of meeting scenarios is introduced. Such extracted information can be used as the basis for content based browsing and the automated analysis of meetings.

The presented system consists of four major functional blocks: The detection of a person, the feature extraction to describe actions, and a sophisticated segmentation approach to find action boundaries. The fourth module consists of a statistical classifier. Beside the functionality of these blocks, the image material for training and testing purposes is briefly introduced.

1. INTRODUCTION

The automated generation of meeting transcriptions has become focus of several EU funded projects, such as the MultiModal Meeting Manager (M4) [1]. Typical tasks within these projects cover recording, representation and browsing of meetings [2]. Furthermore the aspects of tracking the focus of attention [3] and multimodal recognition of group actions [4] are investigated. Group actions, such as consensus and discussion can be derived by the combination of single person actions like rising a hand or speaking.

To derive the desired high level informations, one has to find and recognize the activities of each participant in a meeting session first. In our work the recognition task is solved by a self learning statistical classifier being trained on low-level global motion features. These features have already been the basis for a stable recognition system in conjunction with body gestures [5]. These features can furthermore be used for finding the temporal segmentation of a given sequence by applying a sophisticated implementation of the Bayesian information criterion (BIC). The performances of the introduced modules is measured on M4 data recorded under typical meeting scenario constraints.

2. FINDING ACTION REGIONS

In a first step, people taking part at the meeting have to be found. For this purpose we use a combination of a skin color detection approach (see [6]) and an artificial neural network based face detection algorithm similar to that presented by Rowley and colleagues [7]. These two detection queues are combined using a condensation algorithm to track the faces over the time as introduced by Blake et al. [8].

Having a robust estimation of the localization of the participants' faces, we consider a region around the center of the face to be the action region. This is a rectangular box with geometrical dimensions empirically derived from the face sizes, symmetrically around the center of the face. These regions are covering the entire head, the upper part of the body together with both arms. Example images are depicted in Figure 1.

All found regions describing the participants' actions are further processed in independent queues.

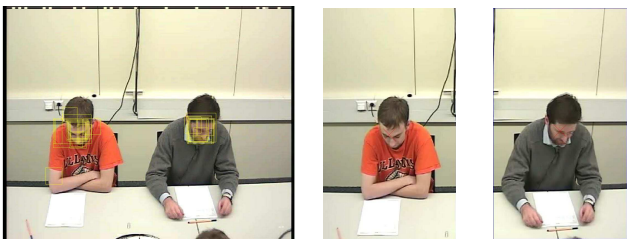

Fig. 1. Typical image in a meeting with detected faces (a) together with left (b) and right (c) action region

3. FEATURE EXTRACTION

Since person actions are always coupled with motion, a series of global motion feature vectors $\vec{X}$ is computed to represent the underlying action information. As shown in numerous previous approaches, building the difference image is an efficient and effective method to extract motion infor-

0-7803-8554-3/04/$20.00 ©2004 IEEE.

mation in a static or slow changing environment. The difference image sequence $I_d(x,y)$ is built by subtracting the pixel values at equal positions (x,y) of two adjacent frames in the image sequence. The thresholded absolute gray values $I_d(x,y)$ represent the intensity of motion for each spatial position (x,y) in the difference image and therefore characterize a specific action.

The computation of the center of mass of the difference image relative to the face $\vec{m}(t) = [m_x(t), m_y(t)]^T$ delivers the *center of motion* in x- and y-direction.

$$m_x(t) = \frac{\sum\limits_{(x,y)\in R_i} x|I_d(x,y,t)|}{\sum\limits_{(x,y)\in R_i} |I_d(x,y,t)|} \quad m_y(t) = \frac{\sum\limits_{(x,y)\in R_i} y|I_d(x,y,t)|}{\sum\limits_{(x,y)\in R_i} |I_d(x,y,t)|} \tag{1}$$

To consider the dynamic of a movement or the acceleration, the changes of the center of mass $\Delta m_x(t) = m_x(t) - m_x(t-1)$ and $\Delta m_y(t) = m_y(t) - m_y(t-1)$ are also computed.

Additionally, the mean absolute deviation of a pixel (x,y) relative to the center of motion $\vec{\sigma(t)} = [\sigma_x(t), \sigma_y(t)]^T$ is used to describe the motion.

$$\sigma_x(t) = \frac{\sum\limits_{(x,y)\in R_i} |I_d(x,y,t)|(x - m_x(t))}{\sum\limits_{(x,y)\in R_i} |I_d(x,y,t)|}$$
$$\sigma_y(t) = \frac{\sum\limits_{(x,y)\in R_i} |I_d(x,y,t)|(y - m_y(t))}{\sum\limits_{(x,y)\in R_i} |I_d(x,y,t)|} \tag{2}$$

With this kind of feature it is possible to distinguish between an action where large parts of the body are in motion (e.g. "get-up") and an action concentrated in a smaller area, where only small parts of the body move (e.g. "nodding"). This feature can be also considered as *wideness of motion*.

The last important feature describing a motion, is the *intensity of motion* $i(t)$, which can simply be expressed by the average absolute value of the motion distribution:

$$i(t) = \frac{\sum\limits_{(x,y)\in R_i} |I_d(x,y,t)|}{\sum\limits_{(x,y)\in R_i} 1}. \tag{3}$$

A large value of $i(t)$ represents a very intensive motion of parts of the body, and a small value characterizes an almost stationary image.

By introducing these features to represent the activity within an action region, the complexity and dimension of the high dimensional pattern can be dramatically reduced to a 7 dimensional vector, while preserving the major characteristics of the currently observed motion.

$$\vec{x}_t = [m_x, m_y, \Delta m_x, \Delta m_y, \sigma_x, \sigma_y, i]^T, \tag{4}$$

4. SEGMENTATION USING THE BAYESIAN INFORMATION CRITERION

For the automatic temporal segmentation of a feature vector sequence, an efficient variant of the BIC approach presented by Tritschler and Gopinath [9] is deployed.

The idea is as follows: A sliding window, beginning at feature s with the length n within a sequence of images, is scanned for an action boundary. If no boundary is found, the length of the window is enlarged and the process is repeated until a boundary is found. Hereafter the start of the window is moved to the found segment boundary. This process is repeated until the end of the given stream is reached.

Inside a window with n frames $x_s, \ldots, x_{s+n}$, a boundary at position $i \in \{s+4, \ldots, s+n-4\}$ is arbitrarily placed, so that two segments arise. In a second step, it is tested whether it is more likely that one process Φ_1 has produced the output $x_s, \ldots, x_{s+n}$, or that two different processes Φ_{21} and Φ_{22} have generated the two segments' output $x_s, \ldots, x_{s+i}$ and $x_{s+i+1}, \ldots, x_{s+n}$ respectively. To represent these processes, the covariance matrices Σ of the features $x(t)$ as well as the energy of the feature vector $e(t)$ are used. The decision rule to test a feature segment at a given discrete time i is:

$$\Delta BIC_i \overset{!}{<} 0 \quad \text{where } \Delta BIC_i \text{ is the minimum} \tag{5}$$

$$\Delta BIC_i = -\frac{n}{2}\log\|\Sigma_w\| + \frac{i}{2}\log\|\Sigma_f\| + \frac{n-i}{2}\log\|\Sigma_s\| + \frac{1}{2}\lambda\left(d + \frac{d(d+1)}{2}\right)\log n \tag{6}$$

In this equation the variable Σ_w denotes the covariance matrix of all feature vectors $x_s, \ldots, x_{s+n}$ respectively their energy $e_s, \ldots, e_{s+n}$, where Σ_f and Σ_s are the covariance matrices of the features or their energy of the first and the second segment. The dimension of the feature vector is given by d, which is 7 for motion features and 1 for the energy. According to the theory, the penalty weight λ should be 1. But in praxis it has turned out, that this weight heavily influences the sensitivity of the segment finder.

The experiments have shown, that best results can be achieved by using the energy of the feature vectors instead of the features themselves. Using the entire vector resulted in a highly sensitive and partly unstable segmentations. Another empirical observation is, that a initial window length of the ΔBIC with $n = 15$ gives optimal results. This value seems to be a good trade-off between the duration of actions and the time between them. A good segmentation on a typical sequence of actions is depicted in figure 2 ($\lambda = 4.5$).

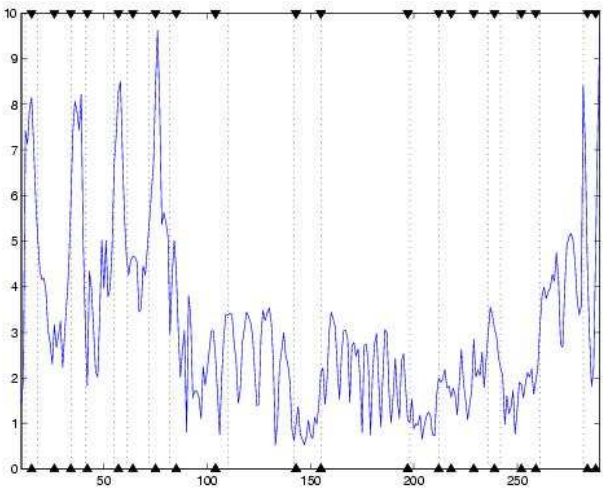

Fig. 2. Segmentation boundaries: hand labeled (triangles) and those found by BIC (dashed line)

5. OBEYED VIDEO MATERIAL

The training and testing material was acquired for the use within the *MultiModal Meeting Manager* project (M4 [1]) in a smart meeting room. In this room two cameras are mounted on two opposite walls capturing the participants sitting in front of a row of tables (see Figure 1). A third camera records an area with a whiteboard and a projection canvas as shown in Figure 3.

Six actions are performed by several persons during several sessions with this camera setup. A part of the gathered material is used for the training of the action models (see section 6 below), another portion is considered to be the evaluation material. Presently the vocabulary covers the following actions: "Stand up", "sit down", "nodding", "shaking head", "writing" and "pointing". It has turned out that mostly all important group actions can be derived from these elements.

Fig. 3. Whiteboard and canvas captured by third camera

6. ACTION MODELLING USING HMMS

In this section we briefly describe how actions can be modeled and recognized using Hidden Markov Models. The statistical HMM framework and its theoretical aspects are well covered by Rabiner in [10].

HMMs have shown superior recognition results in numerous classification scenarios due to their flexible time warping capabilities. This fact allows the modeling of actions respectively their observation sequences with different lengths. Furthermore these models have the ability to model several variations of the same action performed by a huge group of individuals by using several mixtures or a set of discrete distribution probabilities.

Unknown actions can be classified using the following maximum-likelihood decision:

$$M^* = \underset{M \in \text{all actions}}{\operatorname{argmax}} \quad P(X|M) \tag{7}$$

In this equation, X represents an unknown feature vector sequence of an unknown action and M represents one HMM from the set of all known actions. The classifier recognizes the performed action by finding the model M^* with the highest production probability $P(X|M)$.

Therefore the values of $P(X|M)$ for all models have to be computed. This task can be solved by using a Viterbi decoder. But before that, all model parameters have to be estimated first. These parameters are the state-transition matrices $\overleftrightarrow{A}$ and the production probabilities $P(x|S_i)$ in all states S_i for each HMM among the database.

For our system we have tested continuous multivariate Gaussian mixtures as well as discrete production probabilities. In the first case the real distributions of the features are approximated by a weighted sum (factor w_k) of several normal distributions $\mathcal{N}$.

$$p(\vec{x}_i|S_i) = \sum_{k=1}^{C_i} \mathcal{N}(\vec{x}_i, \vec{\mu}_k, \Sigma_k) \cdot w_k \tag{8}$$

The normal distribution describes the probability for a J-dimensional observation $\vec{x}_i$ in a certain state S_i. It is given by its mean-vector $\vec{\mu}$ and its covariance matrix Σ:

$$\mathcal{N}(\vec{x}_i, \vec{\mu}, \Sigma) = \frac{1}{\sqrt{(2\pi)^J |\Sigma|}} e^{-\frac{1}{2}(\vec{x}_i - \vec{\mu})^T \Sigma^{-1} (\vec{x}_i - \vec{\mu})} \tag{9}$$

Discrete probabilities $p(\vec{x}_i|S_i)$ are defined by a codebook that maps continuous observations to discrete labels (vector quantifier). This can be implemented by a K-Means Clustering algorithm.

All parameters above can be estimated using the well known Baum-Welch estimation algorithm and prior isolated action feature sequences. For our system we first train a common action model with all available training examples.

In a second step the action models are reestimated with their corresponding examples. Presently the possibility to reject unknown actions is not implemented.

7. EXPERIMENTS AND RESULTS

Goal of our experiments was to measure the performance of our present system on a part of the M4 meeting room datasets. Therefore the above explained feature extraction, segmentation and modeling techniques have been applied to several meeting videos. To test the segmentation- and the recognition modules separately, the segment boundaries were given manually. The results of a discrete HMM system (summarized in tabular 7) achieve an averaged recognition scores of over 86%. Discrete models have turned to be superior compared to continuous ones. Beside the class "shkaing head", all other show acceptable scores. The reason for this low result might be explained by the fact that the magnitude of the difference images is not very distinct for this class.

	sit down	stand up	nodding	shaking	writing	pointing	%c	%e
sit down	9	1	0	0	0	0	90.0	0.1
stand up	3	10	0	0	0	1	71.4	0.4
nodding	1	3	251	2	33	0	86.6	4.0
shaking	0	0	30	2	11	0	4.7	4.2
writing	0	0	28	0	515	7	93.6	3.6
pointing	4	1	4	0	6	57	79.2	1.5
Overall							86.2	

Table 1. Confusion matrix of recognized actions.

The automated BIC-based segmentation module showed first precise results using the energy of the computed feature vectors as depicted in Figure 2. Nearly all boundaries were found at the right position. It is possible to control the sensitivity of the penalty-parameter λ to find a good trade-off between a over-segmentation and boundary losses.

8. CONCLUSIONS AND OUTLOOK

A system for the automated segmentation and recognition of actions in meeting scenarios was presented. The main parts of the system are the feature extraction based on global motion features, a BIC based segmentation module and an action classification module using a statistical HMM approach. The performance of the system was evaluated on some video sequences from the M4 project. Recognition scores of up to 86% were achieved, which indicates, that the underlying global motion features carry most of the relevant informations of the performed actions. The temporal segmentation also showed acceptable results on the obeyed material.

In the future it is planned to investigate the use of feature vectors which describe the distribution of the motion within an action region in more detail. Furthermore a filler model will be added to the HMM-vocabulary in order to reject unknown actions. This will also improve the interaction between the segmentation- and the recognition-module, which has to be also improved.

9. REFERENCES

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

[2] A. Waibel, M. Bett, F. Metze, K. Ries, T. Schaaf, T. Schultz, H. Soltau, H. Yu, and K. Zechner, "Advances in automatic meeting record creation and access," in *In Proceedings of ICASSP-2001, Salt Lake City, Utah*, 2001.

[3] R. Stiefelhagen, "Tracking focus of attention in meetings.," in *In IEEE International Conference on Multimodal Interfaces, Pittsburgh, PA.*, 2002.

[4] I. McCowan, S. Bengio, D. Gatica-Perez, G. Lathoud, F. Monay, D. Moore, P. Wellner, and H. Bourlard, "Modeling human interaction in meetings.," in *ICASSP Proceedings (to appear), Hong Kong*, 2003.

[5] Stefan Eickeler, Andreas Kosmala, and Gerhard Rigoll, "Hidden Markov Model Based Continuous Online Gesture Recognition," in *Int. Conference on Pattern Recognition (ICPR)*, Brisbane, Aug. 1998, pp. 1206–1208.

[6] M. Soriano, S. Huovinen, B. Martinkauppi, and M. Laaksonen, "Skin detection in video under changing illumination conditions," in *Proc. 15th International Conference on Pattern Recognition, Barcelona, Spain*, 2000, pp. 839–842.

[7] H.A. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," in *IEEE Transactions on PAMI*, Jan. 1998, pp. 23–38.

[8] M. Isard and A. Blake, "Condensation – conditional density propagation for visual tracking," *International Journal of Computer Vision 29(1)*, pp. 5–28, 1998.

[9] A. Tritschler and R. Gopinath, "Improved Speaker Segmentation and Segments Clustering Using the Bayesian Information Criterion," in *Proc. EUROSPEECH*, Paris, France, 1999, vol. 2, pp. 679–682.

[10] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–285, Feb. 1989.