# A Neural-Field-like Approach for Modeling Human Group Actions in Meetings

Stephan Reiter and Gerhard Rigoll
Technische Universität München
Institute for Human-Machine Communication
80290 Munich, Germany

## Abstract

*In this paper we investigate a new architecture for recognizing human group actions in meetings. These group actions provide a basis that enables effective browsing and querying in a meeting archive. For this task we propose an architecture that was inspired by the neural field theory. Our approach is particular, because contrary to other methods, we present all features to our classifier in parallel. The experiments show, that our system has comparable results to existing sequential techniques.*

## 1. Introduction

Analysis of meetings is a task that some research groups have begun to deal with only recently. Meanwhile a number of groups are concerned with developing a meeting recorder or a meeting browser system. In the meeting project at ICSI [8], for example, the main goal is to produce a transcript of the speech. At CMU the intention is to develop a meeting browser, which includes challenging tasks like speech transcription and summarization [12] and the multimodal tracking of people throughout the meeting [2, 11]. In the European research project M4 the main concern is the construction of a demonstration system to enable structuring, browsing and querying of an archive of automatically analyzed meetings.

Due to the complex information flow of visual, acoustic and other information sources in meetings (e.g. from documents or projectors) the segmentation of a meeting in appropriate sections represents a very challenging pattern recognition task, which is currently investigated by only a few research teams.

Goal of the described work here is, to divide a meeting into segments with the length of several seconds, the so called group actions. A common approach is to present the features in a sequential order. This is done for example in [7, 9, 13, 10]. There various standard techniques for pattern recognition like Hidden Markov Models (HMM), Bayesian Networks, Multilayer Perceptrons (MLP) and Support Vector Machines (SVM) are used. Here we propose a new approach by presenting all features in parallel in one instant. In doing so, events from the very beginning of a meeting can have influence on an event at the end of the meeting and the other way round.

The paper is organized as follows: Section 2 describes the meeting data. In Section 3 the neural-field-like system is introduced. Section 4 then describes the features that were used. Finally in Section 5 the results of our system are given.

## 2. The Meeting Data

For our research we used the public available meeting corpus from IDIAP that is described in [7]. This corpus consists in special scripted meetings that were recorded in the IDIAP Smart Meeting Room. This is a room equipped with fully synchronized multichannel audio and video recording facilities. Each participant has a close-talk lapel microphone attached to his clothes. Additionally a microphone array on top of the table was used. Three television video cameras provide PAL quality video signals that were recorded onto separate digital video tape recorders.

Each recorded meeting consists of a set of predefined group actions in a specific order that was defined in an agenda. The appearing group actions are:

- Monologue (one participant speaks continuously without interruption)
- Discussion (all participants engage in a discussion)
- Note-taking (all participants write notes)
- White-board (one participant at front of room talks and makes notes on the white board)
- Presentation (one participant at front of room makes a presentation using the projector screen)

A total of 53 scripted meetings with two disjoint sets of meeting participants were recorded. 30 of them were used for the training, the remaining 23 videos were used for the evaluation of the system. In each meeting there were four participants at six possible positions: four seats plus white-board and presentation board.



*0-7803-9332-5/05/$20.00 ©2005 IEEE*

## 3. The Neural-Field-like system

For the task of segmenting the meetings into group actions we propose a new approach that is based on the theory of the neural fields, first analyzed by Amari [1]. Meanwhile some other researchers make use of the neural fields in various applications [5, 3]. Here the idea is to present the features of a whole meeting to the neural field simultaneously and get a segmentation and classification as output. In this way elements from the end of a meeting can have influence on elements at the beginning, which should increase the robustness of the classification task. The typical equation for a neural field is denoted in eq. 1.

$$\tau \frac{\partial u(x,t)}{\partial t} = -u(x,t) + \int w(|x-y|) f[u(y)] dy + h + s(x,t) \tag{1}$$

Thereby the output $u(t)$ describes the average membrane potential at time t of a neuron. The average activity is given by the transfer function $f[u(t)]$, which can be of any arbitrary style. The input for each neuron is defined by the term $s(x,t)$. The strength of the connections $w(|x-y|)$ between two neurons is defined here as a function of their relative distance. $\tau$ describes the time constant of the dynamics and $h \geq 0$ the resting level to which the output $u(x,t)$ relaxes in the absence of any input.

As our inputs will be stationary, the time dependency of $s(x,t)$ does not apply. So the input becomes $s(x)$ only. Furthermore we are only interested in stable points, where the activations of the neurons are not changing anymore. This is the case, if $\frac{\partial u(x,t)}{\partial t} = 0$. With these two constraints, the equation for the neural field becomes

$$u(x) = \int w(|x-y|) f[u(y)] dy + h + s(x) \tag{2}$$

Since a computer cannot handle calculations in a continuous space, we have to discretize the neural field equation. It then becomes the following form,

$$\mathbf{u}(k) = \sum_{\xi} w(|k-\xi|) f[u_{\xi}(k-1)] + h + s(k) \tag{3}$$

with the index $\xi$ running from 1 to the number of all neurons in the field.

The architecture of our neural-field-like classifier is like the following: The features of an entire meeting are put in temporal order, are then cut into N frames and are provided to the classifier as input. The calculated output gives then a hint which group actions occur, in which order and with appropriate time stamp. The level of the activity (after quantization) corresponds to a specific group action. This is illustrated in figure 1.

An advanced investigation of equation 3 reveals a certain similarity to recurrent neural networks. The first layer of an Elman Network [4] for example has the following equation:

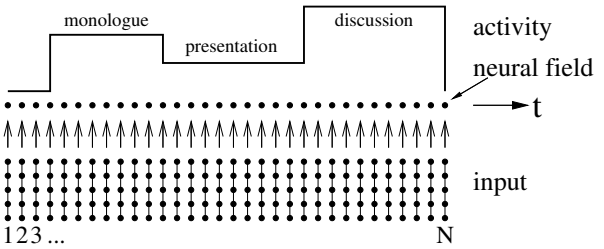


Figure 1: Schematic sketch of the architecture of the neural-field-like classifier

$$a_1(k) = \sum W_r a_1(k-1) + h + s(k) \tag{4}$$

where $W_r$ is the weight matrix, $s(k)$ is the input and $h$ is the bias. In equation 4 the assumption was made that the activity function is the identity function $f(x) = x$. Also $s(k)$ is an abbreviation of the weighted sum of the inputs $s(k) = \sum w_i s_i(k)$. With these assumptions it is possible to define an equivalent recurrent neural net, that has almost the same architecture as our proposed neural-field-like system, but has the great advantage that all known learning algorithms can be used. In figure 2 an equivalent recurrent neural net is shown that is used for our segmentation and classification problem. In this figure not all connections between the neurons are plotted. From each neuron there can be recurrent connections to all other neurons. We define the coupling length as the number of recurrent connections to one side of a neuron. For this type of neural net all training-algorithms for recurrent nets can be used. In our case we use a Jordan-Elman-Back-Propagation training.

For each frame $i \in [1 \dots N]$ there is one neuron in the recurrent neural net. The input of each neuron consists of six or twelve features from the speaker-turn detection and global-motion detection respectively, depending on whether we use an unimodal or a multi-modal approach. The output is binary coded. Therefore the output layer consists in $8 \cdot N$ neurons since we have eight classes. For each of the N time frames the resulting group action is determined by the neuron with the highest activity.

## 4. Feature Extraction

This section illustrates briefly the low level algorithms that are used to provide the input for our neural-field-like system.

### 4.1. Speaker Turn Detection

The results of the speaker turn detection have been kindly provided from another partner in our international project.

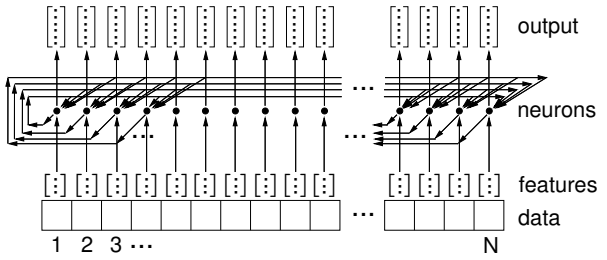


Figure 2: Architecture of the equivalent recurrent neural net (not all connections are shown)

A generic, short-term clustering algorithm is used that can track multiple objects for a low computational cost. In [6] the three-step algorithm consisting in frame-level analysis, short-term analysis and long-term analysis is presented in detail.

### 4.2. Global Motion Features

Global motion features have turned out as suitable features for gesture recognition in [14]. For the task of group action recognition some of these features were investigated. In general the global motion features are calculated as follows: In the difference image $I_d(t)$ a so called action region around a person is defined. Among other features, in this region the intensity of motion $i(t)$ of the center of mass $\mathbf{m}(t)$ is estimated. The intensity of motion is a strong indicator, whether a person is present at a specific place or not. This calculation is done for all six possible locations in the meeting room.

## 5. Group Action Segmentation

As mentioned in section 1 all features of an entire meeting are presented to the recurrent neural net in parallel. With a frame rate of five Hertz and a length of a meeting of approx. five minutes, this results in a total number of $5min \cdot 60\frac{sec}{min} \cdot 5\frac{1}{sec} = 1500$ features of at least six dimensions. Unfortunately such an amount of data is not feasible. Therefore some features have to be combined to one item. Another reason for this procedure is that we can guarantee that a recognized group action has at least the length of the merged features. This we refer to as minimum length of a group action.

We conducted several experiments with varying numbers of inputs and various time scales. First experiments with only one modality (only speaker-turns) were accomplished. Table 1 shows the results of several passes with different minimum lengths. Here the best result is achieved, when the group actions have a minimum length of twenty seconds. Then the frame error rate is only 0.333. Unfortunately there is no dependency between the number of seconds to be merged and the frame error rate. So it cannot be predicted which configuration will perform best.

| #sec. | FER (ST) | FER (GM) | FER (ST&GM) |
|---|---|---|---|
| 2 | 0.425539 | 0.589715 | 0.531016 |
| 4 | 0.424125 | 0.622475 | 0.467507 |
| 6 | 0.402729 | 0.626837 | 0.443321 |
| 8 | 0.421224 | 0.543035 | 0.497606 |
| 10 | 0.370983 | 0.524188 | 0.427644 |
| 12 | 0.402879 | 0.542997 | 0.447546 |
| 14 | 0.420272 | 0.524324 | 0.417743 |
| 16 | 0.395198 | 0.495934 | 0.468771 |
| 18 | 0.389778 | 0.555014 | 0.457162 |
| 20 | 0.333396 | 0.511352 | 0.438292 |

Table 1: Results of different time granularity with a coupling length of three neurons using only speaker-turn detection (ST), only global-motion features (GM) and both modalities (ST&GM). The first column shows the minimum length of a group action in seconds, the second column denotes the frame error rate.

Doing the same experiments using only global motion features gives a similar but slightly worse result (cf. table 1). The best frame error rate was achieved with 0.463 at a minimum length of 20 seconds and no coupling to other neurons.

In table 2 the minimum length of a group action is twenty seconds (features of 20 seconds are merged). The FER of different coupling widths is shown. As can be seen, there is also no dependency between the coupling length and the frame error rate.

One would expect that the result would increase, if more information (i.e. speaker-turns *and* global-motion features) are combined. If the columns of table 2 are compared, there is never an improvement in the frame error rate, when both modalities are used. This could have various accounts. One reason could be that the global motion features are not suitable for the task of group action recognition. Another reason may be that our architecture can not profit from the additional information but is likely to be confused by it.

Nevertheless quite promising results could be achieved. The overall best frame error rate is obtained, when only speaker-turns are used, features of twenty seconds are merged and the recurrent neural net has a coupling length of three neurons. Then the frame error rate is roughly 0.333. This result is comparable to the one that we achieved using a completely different approach in [10]. There the best results were frame error rates between 0.3180 and 0.3495, using only speaker turns, depending on which classifier was used. So this neural-field-like approach seems to be able to compete with conventional methods.

| #N | FER (ST) | FER (GM) | FER (ST&GM) |
|---|---|---|---|
| 1 | 0.393332 | 0.463366 | 0.454581 |
| 2 | 0.342446 | 0.554874 | 0.450169 |
| 3 | 0.333396 | 0.511352 | 0.438292 |
| 4 | 0.404592 | 0.532336 | 0.417735 |
| 5 | 0.360240 | 0.512298 | 0.394724 |
| 6 | 0.406739 | 0.482301 | 0.445241 |
| 7 | 0.425993 | 0.544475 | 0.458006 |
| 8 | 0.368450 | 0.480301 | 0.425083 |
| 9 | 0.396993 | 0.500507 | 0.406331 |
| 10 | 0.393496 | 0.501682 | 0.462142 |
| 11 | 0.414315 | 0.585865 | 0.437957 |
| 12 | 0.416150 | 0.511750 | 0.447949 |
| 13 | 0.401015 | 0.669767 | 0.443323 |
| 14 | 0.360258 | 0.500154 | 0.419432 |
| 15 | 0.357517 | 0.497479 | 0.436683 |
| 17 | 0.381569 | 0.464032 | 0.490743 |
| 19 | 0.381569 | 0.464032 | 0.490743 |
| 20 | 0.381569 | 0.464032 | 0.490743 |

Table 2: Results of different experiments with various coupling-length (i.e. numbers of neurons that can influence each other), using only speaker-turn detection (ST), only global-motion features (GM) and both modalities (ST&GM). The minimum length of a group action is 20 seconds. The first column shows the coupling length, the following columns denote the frame error rates.

# 6. Summary and Conclusions

In this work we presented a new architecture based on the neural field theory for the segmentation and recognition of group actions in meetings. The features are not presented sequentially but all features of a whole meeting are presented in parallel. This enables influence from the end of the meeting to the beginning and the other way round. The results were quite interesting and comparable to other approaches in this research field. Further investigations are necessary to make better use of the multi-modality. Being still at the beginning of this challenging research we are convinced that the neural-field theory bears a great potential even in pattern recognition tasks.

# References


[1] Shun-Ichi Amari. Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, 27:77–87, 1977.

[2] Michael Bett, Ralph Gross, Hua Yu, Xiaojin Zhu, Yue Pan, Jie Yang, and Alex Waibel. Multimodal meeting tracker. In *Proceedings of RIAO2000*, Paris, France, April 2000.

[3] Hannes Edelbrunner, Uwe Handmann, Christian Igel, Iris Leefken, and Werner von Seelen. Application and optimization of neural field dynamics for driver assistance. In *The IEEE 4th International Conference on Intelligent Transportation Systems (ITSC 01)*, pages 309–314. IEEE Press, 2001.

[4] Jeffrey L. Elman. Finding structure in time. *Cognitive Science*, 14(2):179–211, 1990.

[5] Christian Igel, Wolfram Erlhagen, and Dirk Jancke. Optimization of neural field models. *Neurocomputing*, 36:225–233, 2001.

[6] Guillaume Lathoud, Iain A. McCowan, and Jean-Marc Odobez. Unsupervised Location-Based Segmentation of Multi-Party Speech. In *Proceedings of the 2004 ICASSP-NIST Meeting Recognition Workshop*, Montreal, Canada, May 2004.

[7] I. McCowan, S. Bengio, D. Gatica-Perez, G. Lathoud, F. Monay, D. Moore, P. Wellner, and H. Bourlard. Modeling human interaction in meetings. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, Hong Kong, April 2003.

[8] Nelson Morgan, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Adam Janin, Thilo Pfau, Elizabeth Shriberg, and Andreas Stolcke. The meeting project at icsi. In *Proceedings of the Human Language Technology Conference*, San Diego, CA, March 2001.

[9] Stephan Reiter and Gerhard Rigoll. Segmentation and classification of meeting events using multiple classifier fusion and dynamic programming. In *IEEE Proceedings of the International Conference on Pattern Recognition (ICPR)*, pages 434–437. IEEE Computer Society, August 2004.

[10] Stephan Reiter, Sascha Schreiber, and Gerhard Rigoll. Multimodal meeting analysis by segmentation and classification of meeting events based on a higher level semantic approach. In *Proceedings of the 30th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Philadelphia, USA, March 2005.

[11] Rainer Stiefelhagen. Tracking focus of attention in meetings. In *IEEE International Conference on Multimodal Interfaces*, Pittsburgh, PA, USA, October 14–16 2002.

[12] Klaus Zechner. Automatic generation of concise summaries of spoken dialogues in unrestricted domains. In *Proceedings of the 24th ACM-SIGIR International Conference on Research and Development in Information Retrieval*, New Orleans, LA, September 2001.

[13] Dong Zhang, Daniel Gatica-Perez, Samy Bengio, Iain McCowan, and Guillaume Lathoud. Modeling individual and group actions in meetings: a two-layer hmm framework. In *the Second IEEE Workshop on Event Mining: Detection and Recognition of Events in Video, In Association with CVPR*, 2004.

[14] Martin Zobl, Frank Wallhoff, and Gerhard Rigoll. Action recognition in meeting scenarios using global motion features. In J. Ferryman, editor, *Proceedings of the Fourth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS-ICVS)*, 2003.