Meeting Event Recognition using a Parallel Recurrent Neural Net Approach

Stephan Reiter and Gerhard Rigoll

Technische Universität München
Institute for Human-Machine Communication
80290 Munich, Germany

Extended Abstract

In this paper we present a novel architecture for the task of meeting event recognition in meetings that was inspired by the neural field theory. These group actions provide a basis that enables effective browsing and querying in a meeting archive.

For our research we used the public available meeting corpus that is described in [3]. This corpus consists in special scripted meetings that were recorded in the IDIAP Smart Meeting Room. 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), and 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 whiteboard and presentation board.

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]. 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 - and the other way round - which should increase the robustness of the classification task.

The typical equation for a neural field is denoted in eq. 1.

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

An advanced investigation of eq. 1 (especially discretization and relying only on stable points) reveals a certain similarity to recurrent neural networks. Assuming that there is no time dependency and choosing proper functions for the

From: MLMI 2005, Edinburgh, UK

activity function, we can write the resulting equation of the recurrent neural net that is easy to learn using one of the well known algorithms:

$$a(k) = \sum W_r a(k-1) + h + s(k) \tag{2}$$

where W_r is the weight matrix, $s(k)$ is the input and h is the bias.

The input of each neuron consists of features that result from a speaker-turn detection algorithm by [2] and of global-motion features that have turned out to be suitable for gesture recognition in [5]. Depending on whether we use an unimodal or a multi-modal approach the dimensionality is six or twelve. 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. The detected meeting event in a specified time window is then derived by majority voting. For a more complete description of the model please refer to [4].

As mentioned all features of an entire meeting are presented to the recurrent neural net in parallel. First experiments with only one modality (only speaker-turns or global-motion-features resp.) were accomplished. As result we achieved a recognition error rate of 0.4748 and 0.4393 respectively.

In combining audio and visual features the results could be improved significantly. The recognition error rate of the meeting events decreased by 18.64% absolute (39.7% relative) to 0.2864.

	audio	video	audio+video
recognition error	47.48%	43.93%	28.64%

The results show that by providing a multi-modal input the results could be improved. Further investigations have to demonstrate the ability of the proposed architecture to be suitable for other multi-modal recognition purposes.

References

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

[2] 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.

[3] 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. IDIAP-RR 02-59.

[4] Stephan Reiter and Gerhard Rigoll. A neural-field-like approach for modeling human group actions in meetings. In *Proceedings of the International Conference on Multimedia and Expo (ICME)*, Amsterdam, The Netherlands, July 2005.

[5] 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.