A COMBINED LSTM-RNN - HMM - APPROACH FOR MEETING EVENT SEGMENTATION AND RECOGNITION

Stephan Reiter, Björn Schuller, and Gerhard Rigoll

Institute for Human-Machine-Communication
Technische Universität München
Arcisstr. 21, 80290 Munich, Germany
email: {reiter, schuller, rigoll}@ei.tum.de

ABSTRACT

Automatic segmentation and classification of recorded meetings provides a basis that enables effective browsing and querying in a meeting archive. Yet, robustness of today's approaches is often not reliable enough. We therefore strive to improve on this task by introduction of a tandem approach combining the discriminative abilities of recurrent neural nets and warping capabilities of hidden markov models. Thereby long short-term memory cells are used for audio-visual frame analysis within the neural net. These help to overcome typical long time lags. Extensive test runs on the public M4 Scripted Meeting Corpus show great performance applying our suggested novel approach.

1. INTRODUCTION

Automatic analysis of meetings has the potential to greatly reduce time and costs compared to human annotation. However, adequate robustness is yet to meet. Numerous research activities are therefore concerned with the development of reliable meeting recorder and browser systems: In the meeting project at ICSI [6], e.g., 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 [11] and the multi-modal tracking of people throughout the meeting [1, 9]. 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, so called meeting events as discussion, monologue or presentation. A common approach is to present the features in a sequential order as done in [5, 7, 8, 12]. Thereby various standard techniques for pattern recognition like Hidden Markov Models (HMM), Bayesian Networks, Multilayer Perceptrons (MLP) and Support Vector Machines (SVM) are used. However we propose a new approach by combining two sequential approaches and their inherent strengths: Long Short-Term Memory Recurrent Neural Nets (LSTM-RNN) and HMM.

The paper is organized as follows: Section 2 describes the database. In Section 3 the used features are described. Section 4 then gives an overview of the system and in Section 5 the results are presented.

2. MEETING CORPUS

Within our research we use the publicly available M4 Scripted Meeting Corpus, described in [5]. It consists of fully scripted meetings recorded in a Smart Meeting Room at IDIAP, which is equipped with fully synchronized multichannel audio and video recording facilities. Each of the recorded participants had a close-talk lapel microphone attached to his clothes. An additional microphone array was mounted on top of one center meeting table. Video signals were recorded onto separate digital video tape recorders by three television video cameras, providing PAL quality.

Each recorded meeting consists of a set of predefined group actions in a fixed order defined in an according 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 the room talks and makes notes on the white board)
- Presentation (one participant at front of the room presents using the only projector screen)

The database comprises a total of 53 scripted meetings with two disjoint sets of participants. A fixed training set makes use of 30 videos, while the remaining 23 are used throughout evaluation. In each meeting there were four par-

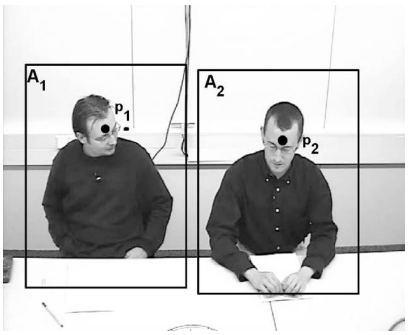

Fig. 1. Video frame marked with action regions and center of the head provided by the tracking algorithm.

ticipants at six possible positions: four seats plus whiteboard and presentation board.

3. FEATURE EXTRACTION

We use a multi-modal feature extraction mechanism that exploits the speaker activity, derived from the audio streams, and the individual gestures that are performed by each participant, and detected in the video streams.

3.1. Speaker Turn Detection

The results of the speaker turn detection have been kindly provided by another partner as mentioned in Section 7. A generic, short-term clustering algorithm is used that can track multiple objects at a low computational cost. In [4] the applied three-step algorithm consisting in frame-level, short-term and long-term analysis is presented in detail.

3.2. Individual Gesture Recognition

In order to recognize gestures of individual persons, we define an action region in which the gesture is exspected. It is determined by the head position of the participant. Next global motion features are calculated from subsequent images in action regions surrounding a participant (see fig. 1). The features in particular are center of mass $\mathbf{m}(t) = [m_x(t), m_y(t)]$, change of center $\Delta\mathbf{m}(t) = [\Delta m_x(t), \Delta m_y(t)]$, variance of motion $\sigma(t) = [\sigma_x(t), \sigma_y(t)]$, and intensity of motion $i(t)$. The resulting 7-dimensional feature stream is segmented within a manually assisted Bayesian Information Criterion framework (cf. [10]). Afterwards these segments are fed forward to a HMM based recognizer which has been trained on roughly 1,000 gestures consisting of writing, pointing, standing up, sitting down, nodding, and headshaking. For more detail refer to [13].

3.3. Multimodal feature vector

From both types of features the most adequate for our task are chosen. The selected elements from both audio and video streams are coupled to a multimodal feature vector of seven dimensions. In particular the entries comprise the amount of talking on the six possible positions and additionally the writing gesture summed up of all participants. Since the feature vector is derived from the signal stream by a 10 seconds windowing with an overlap of nine seconds the frame rate is $1/s$. The ground truth was derived from the agenda given in the database.

4. SYSTEM OVERVIEW

We propose a two stage system for the segmentation and recognition of meetings into meeting events, similar to the tandem approach used for speech recognition, introduced in [2]. As an alteration we use LSTM-RNN instead of MLP in order to profit of their capabilities to handle long time lags. In detail feature vectors derived of single audiovisual frames are classified by the LSTM-RNN. These results are fed forward as posteriors to continuous Gaussian mixture HMM to provide a segmentation via the Viterbi algorithm.

4.1. Long Short-Term Memory Recurrent Neural Net

Recurrent neural networks trained by back-propagation through time and other established methods have the big drawback that they cannot store information over a longer period of time. Bridging such longer lags is demanding as error signals tend to either blow up or vanish. So events lying back in time are likely to be forgotten. To overcome this problem Hochreiter and Schmidhuber introduced the above named Long Short-Term Memory (LSTM) cells [3]. Thereby the hidden cells of a conventional recurrent neural net are replaced by memory blocks, which contain of one or more memory cells (see fig. 2). At the center of a cell a simple linear unit with a single self-recurrent connection is found having its weight set to 1.0. This connection preserves the cell's current state throughout one time step. The output of one cell $y^{c_j}(t)$ is computed as follows:

$$y^{c_j}(t) = y^{out_j}(t) h(s_{c_j}(t)) \tag{1}$$

where the internal state $s_{c_j}(t)$ is

$$\begin{aligned} s_{c_j}(0) &= 0 \\ s_{c_j}(t) &= s_{c_j}(t-1) + y^{in_j}(t) g(net_{c_j}(t)) \end{aligned} \tag{2}$$

This architecture has the advantegous effect that salient events can be remembered over arbitrarily long periods of time. Several cells can be combined to blocks, which share the input and output gate.

The arcitechture of our LSTM-RNN consists in three layer: an input, one hidden, and an output layer. The input

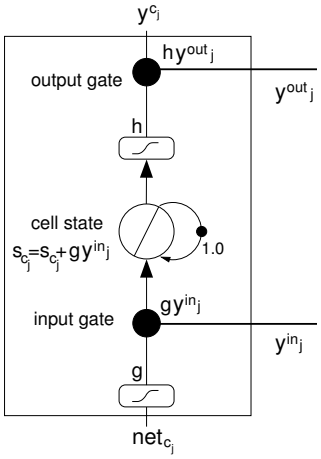

Fig. 2. A memory cell of a LSTM-RNN.

layer has seven nodes, according to the dimensionality of the feature vector. For the hidden layer we use four blocks with four LSTM cells each. In the output layer eight nodes are used corresponding to the number of classes to be discriminated. The weight of each output node is taken as probability of a frame belonging to a specific class.

4.2. HMM framework

The outputs of the LSTM-RNN directly allude to a frame's class. However, the starting and end points of the events to be recognized is unknown. Therefore we apply a HMM framework on top of the results of the neural-filed-like system to get a segmentation.

According to the tandem approach as described in [2] we use the raw output of the LSTM-RNN as the feature stream for the HMM, i.e. there is no explicit assignment of a class for each frame. Because of the great instability of the features, we apply a smoothing mechanism, namely a moving average filter with different window-types and window-lengths. Each class to be recognized is modeled by a continuous system with various numbers of states and a Gaussian mixture output. The segmentation is then done by a Viterbi-decoder.

5. EXPERIMENTS

Prior to presentation of results we describe measures used throughout evaluation. Results thereof are provided for three different meeting event classification approaches: LSTM-RNN stand-alone, and the combined approach with HMM.

5.1. Performance measures

We use the *frame error rate* (FER) and the *Accuracy* as measures to evaluate the results of the meeting event recognition.

States	Accuracy
8	74.31 %
14	82.57 %
18	87.16 %
22	86.24 %
30	78.90 %

Table 1. Results of our tandem approach using different numbers of states of the HMM

The FER is used to evaluate the results of the neural-field-like classifier. It is defined as one minus the ratio between the number of correctly recognized frames and the number of total frames: $FER = (1 - \frac{correctframes}{totalframes}) \times 100\%$. For the evaluation of the segmentation performance we use the commonly accepted accuracy measurement defined as one minus the sum of insertion (Ins), deletion (Del), and substitution (Subs) errors, divided by the total number of events in the ground truth: $Accuracy = (1 - \frac{Subs+Del+Ins}{TotalEvents}) \times 100\%$.

5.2. Performance results

For test and training we use the split defined in section 2. First elaborate tests using only LSTM-RNN are conducted varying the number of hidden LSTM cells. To get comparable results to a standard HMM approach, we apply predefined boundaries of the ground truth to the results and get the meeting event in the specific segment by majority vote. Applying these boundaries gives an accuracy of 96.33%. Short comparative tests with standard multilayer-perceptrons failed completely on this task. Therefore we did not follow up the classic tandem approach any further but restrained our effort on the suggested LSTM-RNN HMM system.

Tests with a stand-alone HMM and automatic segmentation show a maximum accuracy of 83.49%. A significant improvement can be achieved by applying the suggested tandem approach. Using the raw results of the LSTM-RNN without any further preprocessing and a HMM with 18 states we achieve a maximum accuracy of 87.16% (cf. table 1). Compared to a stand-alone HMM we obtain an improvement of 3.67%.

A further improvement of the segmentation performance can be achieved by filtering the output of the LSTM-RNN using a moving average filter. Table 2 shows a comparison of the results with different filter widths using a HMM with 18 states. The best results can be achieved by applying a moving average filter of 5 frames. Then the accuracy of the whole system increases by 5.5% absolute in respect of the system without filtering and reaches 92.66%. In comparison with the hand segmented result the difference is only 3.66% which is a quite satisfactory result. Compared to the stand-alone HMM approach our method outperforms the latter by 9.17%. In table 3 all important results are clearly summarized.

Filter width	Accuracy
1 (no filtering)	87.16 %
5	92.66 %
10	90.83 %
15	88.07 %
20	85.32 %

Table 2. Comparison of different widths of the moving average filter using our tandem approach.

System	Accuracy
LSTM-RNN + ground truth	96.33 %
HMM	83.49 %
LSTM-RNN + HMM	87.16 %
LSTM-RNN + filter + HMM	92.66 %

Table 3. Accuracy using different system setups. (LSTM-RNN = Long Short-Term Memory Recurrent Neural Net, HMM = Hidden Markov Models)

6. SUMMARY AND CONCLUSION

In this work we presented an approach for the segmentation of recorded meetings into group actions. Combining recurrent neural nets and HMM results in a highly discrimative system with warping capabilites. By incorporating LSTM cells into recurrent neural nets even long time series can be modeled. Conventional methods were outperformed by our suggested tandem approach. Post processing of intermediate results further increases the recognition results. The accuracy reached a maximum of 92.66%. In our future research we aim at incorporation of Dynamic Bayesian Networks for their ability to representing complex stochastic processes.

7. ACKNOWLEDGMENTS

This work was partly supported by the EU 6th FWP IST Integrated Project AMI (FP6-506811, publication AMI-137). We also thank IDIAP for kindly providing results of the speaker turn detection.

8. REFERENCES

[1] M. Bett, R. Gross, H. Yu, X. Zhu, Y. Pan, J. Yang, and A. Waibel. Multimodal meeting tracker. In *Proceedings of RIAO2000*, Paris, France, April 2000.

[2] H. Hermansky, D. Ellis, and S. Sharma. Tandem connectionist feature extraction for conventional hmm systems. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Istanbul, June 2000.

[3] S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 1997.

[4] G. Lathoud, I. A. McCowan, and J.-M. Odobez. Unsupervised Location-Based Segmentation of Multi-Party Speech. In *Proceedings of the 2004 ICASSP-NIST Meeting Recognition Workshop*, Montreal, Canada, May 2004.

[5] 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.

[6] N. Morgan, D. Baron, J. Edwards, D. Ellis, D. Gelbart, A. Janin, T. Pfau, E. Shriberg, and A. Stolcke. The meeting project at icsi. In *Proceedings of the Human Language Technology Conference*, San Diego, CA, March 2001.

[7] S. Reiter and G. 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.

[8] S. Reiter and G. 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.

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

[10] A. Tritschler and R. A. Gopinath. Improved speaker segmentation and segments clustering using the bayesian information criterion. In *Proceedings EUROSPEECH '99*, 1999.

[11] K. 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.

[12] D. Zhang, D. Gatica-Perez, S. Begio, I. McCowan, and G. Lathoud. Modeling individual and group actions in meetings: a two-layer hmm framework. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition, Workshop on Event Mining in Video (CVPR-EVENT)*, Washington DC, July 2004.

[13] M. Zobl, F. Wallhoff, and G. 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.