MODELING HUMAN INTERACTION IN MEETINGS

Iain McCowan, Samy Bengio, Daniel Gatica-Perez, Guillaume Lathoud, Florent Monay, Darren Moore, Pierre Wellner, Hervé Bourlard

Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP)
P. O. Box 592, CH-1920 Martigny, Switzerland
{mccowan, bengio, gatica, lathoud, monay, moore, wellner, bourlard}@idiap.ch

ABSTRACT

This paper investigates the recognition of group actions in meetings by modeling the joint behaviour of participants. Many meeting actions, such as presentations, discussions and consensus, are characterised by similar or complementary behaviour across participants. Recognising these meaningful actions is an important step towards the goal of providing effective browsing and summarisation of processed meetings. In this work, a corpus of meetings was collected in a room equipped with a number of microphones and cameras. The corpus was labeled in terms of a predefined set of meeting actions characterised by global behaviour. In experiments, audio and visual features for each participant are extracted from the raw data and the interaction of participants is modeled using HMM-based approaches. Initial results on the corpus demonstrate the ability of the system to recognise the set of meeting actions.

1. INTRODUCTION

When people hold meetings, they do so to communicate and develop information. The IDIAP smart meeting room project is investigating how information from meetings can be captured, stored, structured, queried, and browsed using multimodal sensors, analysis, and user interfaces. The aim is to provide techniques that will help people quickly obtain required information from a meeting archive without having to listen and view entire recordings. This will assist both people who have missed a meeting, as well as those who attended but need to recall certain details.

A number of groups are researching the application of speech and video processing techniques to the meeting domain. The meeting project at ICSI [1], for example, has focused primarily on the challenging problem of producing text transcriptions of speech in meetings. Work at CMU includes speech transcription and summarisation, development of a meeting browser [2], and also the use of cameras to track the focus of attention in meetings [3]. The Microsoft distributed meeting system supports live broadcast of audio and video meeting data and includes a recorded meeting browser [4].

Meetings constitute natural and important cases of people interaction, occur in reasonably constrained, yet challenging conditions, and can be described by a relatively well-defined dictionary of relevant actions. Presentations, discussions, monologues, consensus and disagreements in meetings are actions in which people play and exchange similar, opposite, or complementary roles, each one possibly being played by more than one individual. Furthermore, these actions are inherently semantic and can be used as queries in a retrieval system, or to give structure for browsing.

In this paper, we investigate a probabilistic approach to segmenting meetings by modeling the interaction between participants. The individual behaviour of participants is monitored using a set of features from both the audio and visual modalities. Different sequence models are then trained to recognise high-level events (*meeting actions*) within meetings, such as presentations, general discussion, consensus and note-taking. For experimentation, a corpus of meetings was recorded across multiple audio and visual channels. To facilitate the research, these meetings were loosely scripted in terms of the type and schedule of actions, but otherwise the content is natural.

The paper is organised as follows. Section 2 discusses the recognition of multi-modal actions in meetings by modeling the joint behaviour of participants. Section 3 then describes the collection of a meeting corpus in the IDIAP smart meeting room. Finally, experiments to segment the corpus in terms of meeting actions are presented in Section 4.

2. MULTI-MODAL RECOGNITION OF GROUP ACTIONS IN MEETINGS

This section gives an overview of the proposed approach for recognising meeting actions, and describes the different sequence models that will be investigated in experiments.

2.1. Overview

There is growing interest in computer vision and multimedia signal processing for understanding the behaviour of interacting people, for actions that are defined by playing both similar and complementary roles (e.g. a handshake, a dancing couple, or a children's game) [5], [6], [7], [8]. While most of the work for recognition of interactions has been directed towards visual surveillance in outdoor [7] and office scenarios [6], the analysis of people interaction constitutes a richer research domain.

Group interaction recognition can be approached probabilistically with models that handle multiple information streams and capture consistent data relationships. Within this framework, interaction recognition can be addressed from at least two angles. The first one attempts to recognise actions of individuals *independently*, and fuse all responses at a higher level for further recognition of the interaction. While usually more tractable, models based on this assumption somehow overlook the starting point: the behaviour of an individual during an interaction *is constrained* by the behaviour of the others, i.e., it is not completely independent.

Modeling such constraints amounts to modeling the interactions. The second approach aims at recognising group actions directly, integrating all observations into a unique probabilistic model, and assuming that the constraints can be jointly learned from data. Detection/segmentation/tracking are needed tasks, but recognition of personal actions is skipped altogether. In fact, when interactions *are* the actions, individual behaviour might become less crucial, as long as the group as a whole provides enough evidence about the performed action. This potentially increases robustness to imperfect feature extraction and measurement processes.

2.2. Sequence Models

In order to model the temporal behaviour of a meeting using features extracted from multiple audio and video channels, we propose to use statistical generative models based on Hidden Markov Models (HMMs). HMMs have been used with success for numerous speech and handwritten recognition tasks.

The success of HMMs for these tasks is based on a careful design of sub-models corresponding to language units (phonemes, words, letters). In the case of meetings, we decided to decompose each meeting in units such as monologue or presentation, which we call *meeting actions*. Hence, as for speech recognition systems, given a set of feature sequences representing meetings for which we know the corresponding labeling (but not necesarily the precise alignment), we can train HMMs using the classical embedded training method based on EM, in order to maximize the likelihood of the data. Afterwards, when extracted features of a new meeting are given to the HMM system, the corresponding sequence of meeting actions may be obtained by simply applying the Viterbi decoding algorithm.

A more complex option is the *multi-stream* approach [9]: knowing that the features describing a meeting represent in fact different entities acting during the meeting, we could first model each entity separately with a specific HMM, and then recombine them during decoding using various recombination schemes. Multi-stream models are typically employed with separate streams for audio and visual features in multi-modal tasks, or for different frequency sub-bands in speech recognition. In modeling group interactions however, the streams could instead represent the individual participants.

3. MEETING DATA COLLECTION

The IDIAP Smart Meeting Room is a 8.2m×3.6m×2.4m rectangular room containing a 4.8m×1.2m rectangular meeting table. The room has been equipped with fully synchronised multichannel audio and video recording facilities. For audio acquisition, twenty four high quality miniature lapel microphones are simultaneously recorded at 48kHz with 24-bit resolution. The microphones are identical and are used both as close-talking lapel microphones attached to meeting participants, and in table-top microphone arrays. For video acquisition, three closed-circuit television cameras output PAL quality video signals, which are recorded onto separate MiniDV cassettes using three "video walkman" digital video tape recorders. Each camera is fitted with an adjustable wide-angle lens with a $38^\circ - 80^\circ$ field of view. Full details of the hardware setup are presented in [10].

A "scripted meeting" approach was taken to collect the required audio-visual data for the meeting action recognition experiments. A set of legal meeting actions was defined as,

- Monologue (one participant speaks continuously without interruption)
- Monologue with note-taking (all other participants take notes during the monologue)
- Presentation (one participant at front of room makes a presentation using the projector screen)
- Presentation with note-taking
- White-board (one participant at front of room talks and makes notes on the white-board)
- White-board with note-taking
- Consensus (all participants express consensus)
- Disagreement (all participants express disagreement)
- Note-taking (all participants write notes)
- Discussion (all participants engage in a discussion)

An ergodic Markov model was then used to generate meeting scripts. Each meeting action corresponded to a state in the Markov model with the self-loop transition probabilities governing the relative duration of each action. The transition probabilities were tuned by hand to ensure that the generated action sequences and durations were realistic. On average, each meeting contained 5 actions and was constrained to begin with a monologue and to end with either a consensus, disagreement, or discussion. After generation of each meeting script, the action durations were normalised using a random time (in minutes) drawn from a $N(5, 0.25)$ distribution, in order to constrain the total time to be approximately five minutes.

Two disjoint sets of eight meeting participants each were drawn from the (international) research staff population at IDIAP. For each set, thirty 4-person meeting scripts were generated as described above. The four participants for each meeting were chosen at random from the set of eight people. Every scripted meeting action involving a single participant (monologues, presentations, and whiteboards) was then allocated at random to one of the four participants, giving a total set of 28 potentially distinct meeting actions. Each meeting script was assigned a topic at random (eg. my favourite movie). A dedicated timekeeper (off-camera) monitored the scripted action durations during meeting recording, and made silent gestures to indicate transitions between actions in the script.

The meeting room configuration for the recordings is illustrated in Figure 1. Two cameras each acquired a front-on view of two participants including the table region used for note-taking. The third camera looked over the top of the participants towards the white-board and projector screen. The seating positions of participants were allocated randomly, with the constraint that participants who presented or used the white-board sat in one of the two seats closest to the front of the room (the latter was not exploited during analysis). All participants wore lapel microphones, and an eight-element circular equi-spaced microphone array of 20cm diameter was centrally located on the meeting table.

A total of 60 meeting recordings have been collected (30 recordings $\times$ 2 participant sets), resulting in approximately 5 hours of multi-channel, audio-visual meeting data. Each recording consists of three video channels, and twelve audio channels. While the experiments in this paper investigate the task of segmentation in terms of meeting actions, this corpus is suitable for a number of other audio, visual and multi-modal processing tasks, such as speaker turn detection, topic segmentation, and gaze tracking. To facilitate further research in such directions, the current database

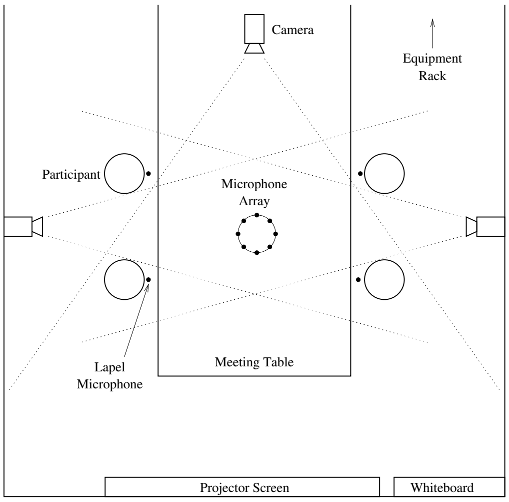

Figure 1: Meeting recording configuration

will be expanded to contain 100 meetings, and is being made available for public distribution [11].

4. EXPERIMENTS

This section presents experiments to recognise meeting actions occuring in the corpus. Due to limited data at the time of testing, the list of actions to be recognised was restricted to monologue by position (4), presentation, whiteboard, discussion, disagreement, consensus and note-taking, giving a vocabulary size of 10 actions. No distinction was thus made between presentations, whiteboards and monologues with or without note-taking. To recognise the meeting actions, a number of different audio and visual features were extracted from the raw data and modeled using HMMs.

4.1. Feature Extraction

The total feature set consists of 19 audio-visual features, which were extracted at a frame rate of 5 Hz. Audio features were extracted to measure the speech activity of each participant, along with the occurrence of a set of positive and negative keywords. The speech activity was measured for 6 predefined locations (participants' seats, whiteboard and presentation screen) from the microphone array signals using the SRP-PHAT measure described in [12]. Two keyword-based features were also calculated for each participant, indicating to the occurence of a list of positive words (e.g. yes, agree, yeah, etc), and a list of negative words (e.g. no, disagree, don't, etc). The final set of audio features consisted of 14 features (speech activity of the 6 locations, and 2x4 keyword streams).

Visual features were extracted using standard methods. For the cameras looking at people at the table, GMM models of skin/-background colors in RGB space were used to extract head blobs, Skin/background pixel classification and morphological postprocessing were performed inside image regions enclosing typical head locations. For each person, the detected head blob was represented by the vertical position of its centroid (normalized by the average centroid computed over the meeting duration). For the wide-view camera, moving blobs were detected by background substraction and represented by their (quantised) horizontal position. The final set of visual features consists of 5 features (1 for each seated head location, plus one from the whiteboard/screen camera).

Model	Action Error Rate
Early Integration HMM	20.0%
Multi-Stream HMM	44.5%
Average of each stream	80.0%

Table 1: Action Error Rates (in percent, lower is better) on the test set with various HMM architectures modeling meeting actions.

4.2. Results and Discussion

Preliminary experiments were performed using the set of artificial meetings recently recorded at IDIAP. For these experiments, there were up to 30 available meetings from the first group of people, which were used as the training set and 29 available meetings from the second group of people, which were used as the test set.

Using a simple leave-one-out cross-validation technique on the training set only, we selected various hyper-parameters of the different HMM models that were tried: the number of states per word, the number of Gaussians per state, and the minimum relative variance allowed per Gaussian. Initialization of the models was done from the known approximate alignment, using Kmeans to train each word model separately. The Viterbi algorithm was then used to train the systems via the embedded training approach. Finally decoding was also performed using the Viterbi algorithm. The only constraint coded in the grammar was to forbid self-loops between actions.

Since the objective was to model the general behaviour of a meeting and not the behaviour of individuals, we compared two approaches:

- an *early integration* approach, where all the features from all participants were merged into a single stream of data, and a single HMM system was trained;
- a simple *multi-stream* approach, where each participant was first modeled separately with a specific HMM trained on his own features, then a single decoding pass was performed on all HMM models simultaneously, merging the likelihoods of each stream at each time step and for each state by simply multiplying them to obtain a unified likelihood.

The difference between the two approaches can be seen as modelling or not the correlation between participants at the state level. If important relations exist between participants, then the early integration approach should perform better, while if this is not the case or if noise is present in one or more streams, then the multi-stream approach should be better.

Table 1 shows the results obtained for both approaches in terms of action error rate (equivalent to word error rate in speech recognition). In addition, the average performance of the HMMs modeling each individual separately is shown. The early integration model yields significantly better performance than the multi-stream approach, which in turn gives a large improvement over the individual HMMs. These results confirm the importance of modeling the correlation and interactions between participants.

Further analysis of the results shows that all events were well recognized except for consensus and disagreements, which were

Model	Action Error Rate
Early Integration HMM	5.7%
Multi-Stream HMM	33.8%
Average of each stream	79.8%

Table 2: Action Error Rates (in percent, lower is better) on the test set, with consensus and disagreement removed from the lexicon.

typically misclassified as discussion. Prior to data collection, it was supposed that consensus and disagreement would be key points that could be characterised by the co-occurence of positive or negative keywords across participants, and possibly also by head movements. Having collected the data and done initial experiments, we have found that on the basis of the selected features, and even to human observers, consensus and disagreement are difficult to distinguish from discussions. During discussion people regularly said words like ‘yes’, ’okay’ or ‘no’, but with little semantic meaning (back-channels), and it was rare for them to make head movement to indicate agreement/disagreement. To improve recognition of these actions in the current framework, we could investigate other audio-visual features or better define the meeting language model. Table 2 gives the results where consensus and disagreement were removed from the lexicon (all occurrences were relabeled in the ground-truth as discussion, and all models retrained).

Although the results presented here are preliminary, they are quite promising, showing that it is indeed possible to model the general behaviour of meetings using statistical models. It is expected that more work on the extraction of other discriminant audio-visual features, coupled with investigation of other sequence models and collection of more training meetings, should improve the performance. Further work will also aim at defining other important meeting actions to be recognised.

5. CONCLUSIONS

This paper has presented an approach to recognising meeting actions by modeling the interactions of participants. In the system, recognition of group actions is the goal, rather than recognition of the individual behaviour of each participant. Audio-visual features were extracted from a multi-modal meeting corpus and HMMs were trained for a set of meeting actions characterised by group behaviour, including presentations, whiteboards, discussions, monologues, consensus, disagreement and note-taking.

Two modeling approaches were investigated : early integration, where the features from all participants were combined in a single HMM, and a multi-stream approach, where participants were modeled in different streams. The early integration approach demonstrated best results, as it better models the correlation between participants. An action error rate of 20.0% was achieved, and this improved to 5.7% when consensus and disagreement were removed from the lexicon (being relabeled as discussion).

While the results presented here are preliminary, they demonstrate the ability to recognise meeting actions by modeling the joint behaviour of participants. Segmentation in terms of these meaningful actions is an important step towards the goal of providing effective summarisation of processed meetings. Ongoing work involves the collection of a more significant meeting corpus, as well as the definition of other meeting actions and more audio-visual features. Other sequence models, such as asynchronous HMMs, will also be investigated to exploit both the correlation and potential asynchronicity between participants.

6. ACKNOWLEDGEMENTS

The authors thank the Swiss National Science Foundation for supporting this work through the National Centre of Competence in Research (NCCR) on ”Interactive Multimodal Information Management (IM2)”. The work was also funded by the European project ”M4: MultiModal Meeting Manager”, through the Swiss Federal Office for Education and Science (OFES). We also thank our colleagues at IDIAP for their assistance during the data collection.

7. REFERENCES

[1] 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 *Proc. of the Human Language Technology Conference*, San Diego, CA, March 2001.

[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 *Proceedings of ICASSP-2001*, Salt Lake City, UT, May 2001.

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

[4] R. Cutler, Y. Rui, A. Gupta, JJ Cadiz, I. Tashev, L. He, A. Colburn, Z. Zhang, Z. Liu, and S. Silverberg. Distributed meetings: A meeting capture and broadcasting system. In *Proceedings of ACM Multimedia Conference*, 2002.

[5] N. Johnson, A. Galata, and D. Hogg. The acquisition and use of interaction behaviour models. In *Proc. IEEE Int. Conference on Computer Vision and Pattern Recognition*, June 1998.

[6] T. Jebara and A. Pentland. Action reaction learning: Automatic visual analysis and synthesis of interactive behaviour. In *Proc. International Conference on Vision Systems*, January 1999.

[7] N. Oliver, B. Rosario, and A. Pentland. A bayesian computer vision system for modeling human interactions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8), August 2000.

[8] M. Isard and J. MacCormick. Bramble: A bayesian multi-blob tracker. In *Proc. IEEE Int. Conference on Computer Vision*, Vancouver, July 2001.

[9] A. Morris, A. Hagen, H. Glotin, and H. Bourlard. Multi-stream adaptive evidence combination for noise robust ASR. *Speech Communication*, 2001.

[10] D. Moore. The IDIAP smart meeting room. *IDIAP Communication 02-07*, 2002.

[11] IDIAP data distribution. http://rhonedata.idiap.ch/.

[12] J. DiBiase, H. Silverman, and M. Brandstein. Robust localization in reverberant rooms. In M. Brandstein and D. Ward, editors, *Microphone Arrays*, chapter 8, pages 157–180. Springer, 2001.