

Recognition of Multi-modal Group Actions in Meetings

Iain McCowan, Daniel Gatica-Perez, Samy Bengio and Guillaume Lathoud



Outline

1. Why research meetings?
2. Our approach.
3. Experiments.
4. Research challenges.
5. Summary.

Why research meetings?

- A rich application domain for research
 - Content:
 - raw, multi-sensor (audio, video, text) data,
 - large volume with low individual value.
 - Metadata:
 - not available from production,
 - people are not good/fast at generating,
 - laborious and expensive if done-by-hand,
 - would increase value of raw data,
 - e.g. text on whiteboard, discussion segments, decision points.
 - Many relevant tasks:
 - truly multi-modal application,
 - intra- and inter-meeting tasks,
 - open problems in many existing fields (including machine learning),
 - potential for new research tasks,
 - applications, e.g. browsing, retrieval, management tasks.

Why research meetings?

- The social psychology perspective
 - important case study of human interaction
 - studied for over fifty years (Bales 1951, McGrath 1984)
 - the basis of group behaviour
 - phases of communication, control, decision, evaluation, reintegration,...
 - patterns of dominance/influence
 - structured, observational approaches
 - categorization of group behavior (coding system)
 - analyst observes the group (overtly or not)
 - the multimodal nature of human communication
 - mechanisms and significance of turn-taking patterns
 - significant info in non-verbal cues
 - modality disambiguation
- computational frameworks for meeting analysis → relevant for social scientists!

Outline

1. Why research meetings?
2. **Our approach:**
 - (a) The multi-modal nature of communication in meetings.
 - (b) The group nature of communication in meetings.
 - (c) Recognition of multi-modal group actions.
3. Experiments.
4. Research challenges.
5. Summary.

Multi-modal Nature

		Meeting progress				Turn-taking			Content/meaning					Atmosphere			
		Align agenda	Task assignments	Decision points	Who was there	What was presented	Backchannel rec.	Side conversation rec.	Floor detection	Transcription	Topic identification	Multimodal dialog acts	Named entity rec.	Coreference resolution	Emotion marking	Disagreements	Consensus
AUDIO	Keyword spotting		•	•								•				•	
	Speech recognition	•	•	•		•	•		•	•	•	•	•		•	•	•
	Speaker identification	•			•				•								
	Syntax	•					•				•	•					
	Disfluency						•							•	•		
	Prosody			•			•	•		•	•		•	•	•	•	•
	Overlap detection			•				•	•	•					•	•	
	Localization				•	•											
	Gun Shot Detection			•											•		
VIDEO	Gaze tracking					•	•	•			•	•			•	•	
	Face recognition	•			•												
	Facial expression rec			•			•	•	•	•				•	•	•	•
	Action/gesture recognition			•			•	•	•		•	•		•	•		
	Posture recognition						•		•		•			•	•	•	•
	Person tracking				•	•								•			
	Lip tracking	•				•	•	•		•	•						
WHITEBOARD	Handwriting recognition	•	•	•		•				•		•	•				
	Diagram analysis									•		•					
PC INTERACTION		•	•		•	•											
PROJECTOR		•				•				•	•		•	•			
CROSS-MODAL	AV tracking				•	•	•			•		•		•			
	AV person identification	•			•										•		
	AV speech recognition	•	•	•		•	•	•	•	•	•	•	•		•	•	•

Group Nature

- To date, most research in automatic meeting analysis has focussed on extracting information from individual participants, e.g. speech, gaze, identity, etc.
- However, meetings are fundamentally different from single user scenarios, due to the presence of multiple interacting people.
- In studying group interactions, the behaviour of each individual is somehow constrained by the behaviour of the others.
 - examples studied by vision researchers include handshakes and dancing couples.
- This has implications for traditional tasks, such as speech recognition, as well as allowing scope to define new group-level tasks.
 - One of the ultimate goals of meeting analysis is to summarise a meeting as a series of high-level agenda items, such as decisions, action points, topics of agreement or disagreement, etc.
 - Such actions are the cumulative result of group interactions.

Recognition of Multi-modal Group Actions

- Motivated by the **multi-modal and group nature** of actions in meetings, our approach consists of
 1. Defining a lexicon of group actions.
 2. Observing relevant multi-modal individual actions, such as gaze, pitch, location, motion.
 3. Recognising the actions using sequence models that capture the interactions between people across modalities.

Lexicon of Group Actions

- We define a meeting as a continuous sequence of mutually exclusive actions taken from an exhaustive set of N meeting actions,

$$V = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_N\}. \quad (1)$$

- Different lexica (coding systems) can be defined to give alternate views of a meeting, e.g.
 - Focus-of-Attention

$$V = \{\textit{'standing presentation'}, \textit{'seated presentation'}, \textit{'whiteboard'}, \textit{'seated monologue'}, \textit{'notes'}, \textit{'dictating'}, \textit{'unfocussed'}\}. \quad (2)$$

- Turn-taking

$$V = \{\textit{'monologue'}, \textit{'dialogue'}, \textit{'group discussion'}, \textit{'floored discussion'}, \textit{'group silence'}\}. \quad (3)$$

Lexicon of Group Actions

- Different lexica can be defined to give alternate views, e.g.

- Interest level

$$V = \{ \textit{‘engaged’}, \textit{‘neutral’}, \textit{‘disengaged’} \}. \quad (4)$$

- Mood

$$V = \{ \textit{‘positive’}, \textit{‘neutral’}, \textit{‘negative’} \}. \quad (5)$$

- Task based (one day perhaps...)

$$V = \{ \textit{‘brainstorming’}, \textit{‘decision making’}, \textit{‘information sharing’}, \dots \}. \quad (6)$$

- These are all actions that characterise the behaviour of the group as a whole, in which participants exhibit similar or complementary behaviour.

Outline

1. Why research meetings?
2. Our approach.
3. Experiments:
 - (a) Lexicon of group actions.
 - (b) Data collection.
 - (c) Observations.
 - (d) HMM variants for modelling multi-model group actions.
 - (e) Results.
4. Research challenges.
5. Summary.

Lexicon of Group Actions

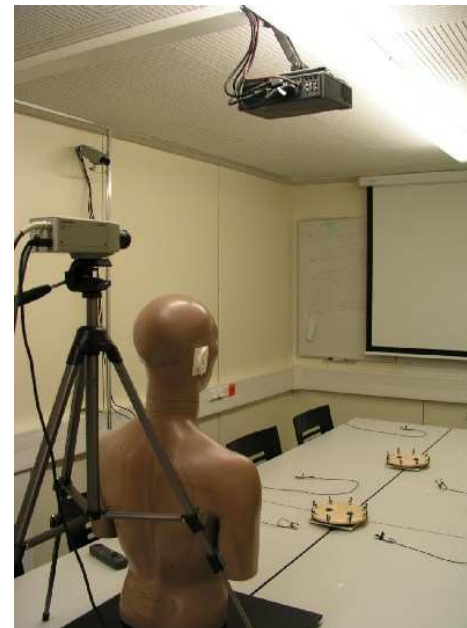
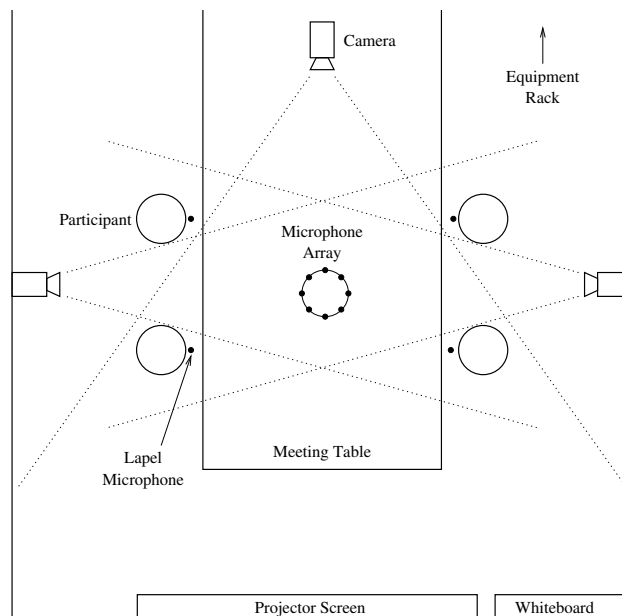
- Action lexicon:

$$V = \{ \text{'monologue1'}, \text{'monologue2'}, \text{'monologue3'}, \text{'monologue4'}, \\ \text{'discussion'}, \text{'presentation'}, \text{'whiteboard'}, \text{'note-taking'} \}. \quad (7)$$

- Essentially turn-taking actions:
 - In social psychology, analysis of turn-taking patterns has been used to give insight into issues such as inter-personal trust, cognitive load in interactions, and patterns of dominance and influence.

Data Collection

- A corpus of 60 (30 train, 30 test), 5-minute, 4-person, meetings was recorded.
- The test and train sets each have a total of approximately 140 actions.
- To facilitate data collection and annotation, each meeting was scripted as a sequence of the actions.
- Apart from this, participant behaviour is unconstrained and natural.



Observations

- 39-dimension observation vectors extracted from 12 audio and 3 visual channels, at 5 Hz.

Feature	Modality		Participants	
	Audio	Visual	Individual	Other
seat speech activity	✓		✓	
white-board speech activity	✓			✓
presentation speech activity	✓			✓
speech pitch	✓		✓	
speech energy	✓		✓	
speaking rate	✓		✓	
head blob vertical centroid		✓	✓	
hand blob horizontal centroid		✓	✓	
hand blob eccentricity		✓	✓	
hand blob angle		✓	✓	
combined motion		✓	✓	
white-board/presentation blob		✓		✓

HMM Variants

- We will adopt a statistical framework similar to the one used for speech recognition:
 - For each kind of action \mathbf{v}_j , model the distribution $p(\mathbf{O}|\theta_j)$, where the sequence of observations \mathbf{O} corresponds to the action \mathbf{v}_j .
 - Maximise the likelihood of L training observation sequences

$$\theta_j^* = \arg \max_{\theta_j} \prod_{l=1}^L p(\mathbf{O}_l|\theta_j). \quad (8)$$

- Use models in the family of Hidden Markov Models (HMMs).
- Use decoding techniques in order to obtain the most probable sequence of actions given a new sequence of observations.

HMM Variants

- The multi-modal and multi-person nature of meetings suggests the need to model multiple interacting sub-processes, or streams.
- Different HMM variants exist to handle multi-stream observations depending on varying assumptions of stream interdependencies, such as
 - **feature-level** correlation,
 - **state-level** correlation, and
 - whether streams evolve **synchronously or asynchronously**.
- Depending on several assumptions, we might
 - model all data into one stream
 - model each modality (audio/video) in separate streams
 - model each individual in separate streams.

HMM Variants

Early integration HMM : the features of all the streams are concatenated into a single vector and modeled using a standard HMM. Such a model assumes :

- feature-level correlation between streams is important, and
- streams evolve frame synchronously.
- efficient training and decoding algorithms: $O(TN^2)$.

State-level Multi-stream HMM : model each stream separately, and combine likelihoods at the state level during decoding.

- during decoding, the emission probability of the combined observations is estimated as the weighted product of each stream emission probability
- such a model assumes independence between features in different streams
- however, during decoding, streams evolve frame synchronously.
- efficient training and decoding algorithms: $O(TN^2)$.

HMM Variants

Asynchronous HMM : Similar to Pair HMMs, allows for possible asynchrony between streams by stretching them.

- models the joint probability of all streams during training and decoding
- only one state variable, but M streams of data
- handles feature-level correlation between streams
- also handles possible (feature-level) asynchrony between streams.
- efficient training and decoding algorithms for the case of two streams: $O(TN^2k)$, if the stretching is limited to k frames.
- however it becomes quickly intractable for more than two streams.

Results

- Results in terms of **Action Error Rate** (similar to word error rate)

$$AER = \frac{Subs + Del + Ins}{Total\ Events} \times 100$$

where

Subst : is the number of substituted actions

Del : is the number of deleted actions (with respect to target)

Ins : is the number of added actions (with respect to target).

Total Events : is the total number of target actions.

- Results varied according to random initialisation in EM training, and so are presented as the **mean and standard deviation over 10 runs**.
- Basic HMM structure: left-to-right topology with minimum durations.

Results

Model	Action Error Rate	Std Dev
Audio-Only	7.0	0.8
Visual-Only	50.9	1.7
Individual Participants	40.2	2.5
Early Integration	8.8	1.7
Participant Multi-stream	11.7	0.7
Audio-Visual Multi-stream	5.5	0.5
Audio-Visual Asynchronous	8.3	0.1

1. Audio is predominant for these actions, however visual has some useful information.
2. Audio-visual multi-stream (optimal) 'better' than early integration.
3. Important to model correlation (interaction) between participants.
4. No particular asynchronous effect between modalities for these actions.

Results

	mono1	mono2	mono3	mono4	white	note	disc	pres	DEL
mono1	11								
mono2		10							
mono3			17						
mono4				11					
white					18				
note						5			1
disc							51		1
pres					1			12	1
INS		1				1			

- Confusion matrix of best system.
- Note low number of examples of each action
 - Cannot distinguish top four systems within 95% confidence interval.
 - Need more data to draw statistically significant conclusions.

Outline

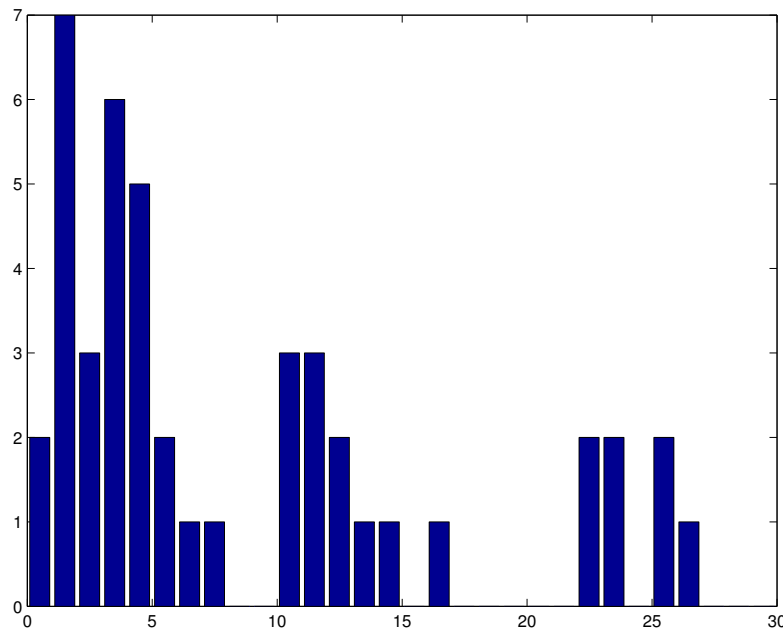
1. Why research meetings?
2. Our approach.
3. Experiments.
4. **Research challenges.**
 - (a) Feature extraction.
 - (b) Modelling asynchrony.
 - (c) Annotation and assessment methodologies.
 - (d) Other potential challenges.
5. Summary.

Feature Extraction

- Spontaneous, natural behaviour
 - Robust people detection
 - Focus-of-attention (gaze, head orientation, body postures)
 - Emotion: facial expressions, acoustic features
 - Other body language cues
 - Adding language: keywords, dialog acts, ...
- Feature selection
 - Features are currently hand-coded
 - Optimal features for a given task?
- Non-intrusive sensors
 - Non-ideal image resolution
 - Multiple views
 - Real-time processing

Modelling Asynchrony

- Intuitively, we expect asynchrony between individuals to be significant, as people react to each other's actions.
- Empirically, we see for the current task that asynchrony of the order of 5-10 seconds between participants is important.



Histogram of delays (in seconds) between different participant streams.

Annotation and Assessment

- Given the subjective nature of some actions and their soft temporal boundaries, there is a requirement to propose appropriate methodologies for annotation and assessment.
- For annotation, there is much to be learned from structured observational approaches to social psychology, including:
 - Constructing well-formed coding schemes (lexica),
 - Event-based or interval-based annotation schemes,
 - The use of multiple annotators, and measures of their agreement.
- For assessment, there is a need to adapt existing measures for this task:
 - Temporal segmentation needs to be assessed, but with flexibility around segment boundaries.
 - The assessment measure must cater for a multi-class problem.
 - Action (word) error rate has some benefits, but only considers the order in which actions occur, without regard for temporal alignment.

Other Potential Challenges

1. Computational complexity:
 - Models that cater for multiple asynchronous streams currently suffer from inefficient training and decoding algorithms for more than 2 streams.
2. Mixed data types:
 - for instance, using discrete and continuous data together, such as recognising disagreements using speech words, pitch information and visual motion.

Summary

- Automatic meeting analysis presents a range of challenging and worthwhile research tasks.
- One framework for this research is to view meetings as a sequence of multi-modal group actions.
- While encouraging progress has been made, this framework poses a number of interesting research challenges.
- Ongoing work:
 - Improving feature extraction processes.
 - Recognition of group interest level.
 - Data collection:
 - Current corpus publicly available at <http://mmm.idiap.ch>.
 - Collection of a new meeting corpus underway (100+ hours of varied, natural, small group meetings).



Recognition of Group Interest Level

- Meetings annotated in terms of the group interest level:
 - 2 annotators per meeting, using 5-point scale.
 - interval scheme used to score 15 second intervals.
 - annotations normalised, merged (mean), then converted to 2 levels (neutral/high) by thresholding.
- Observations include speech activity, speaking rate, head and hand movement.
- Early integration HMM:
 - $\approx 75\%$ accuracy at frame level, allowing flexibility in temporal boundaries (± 2.5 seconds).