Towards Computer Understanding of Human Interactions

Iain McCowan, Daniel Gatica-Perez, Samy Bengio, Darren Moore, Hervé Bourlard

IDIAP Research Institute,
P.O. Box 592, CH-1920, Martigny, Switzerland
{mccowan,gatica,bengio,moore,bourlard}@idiap.ch,
http://www.idiap.ch/

Abstract. People meet in order to interact - disseminating information, making decisions, and creating new ideas. Automatic analysis of meetings is therefore important from two points of view: extracting the information they contain, and understanding human interaction processes. Based on this view, this article presents an approach in which relevant information content of a meeting is identified from a variety of audio and visual sensor inputs and statistical models of interacting people. We present a framework for computer observation and understanding of interacting people, and discuss particular tasks within this framework, issues in the meeting context, and particular algorithms that we have adopted. We also comment on current developments and the future challenges in automatic meeting analysis.[1]

1 Introduction

The domain of human-computer interaction aims to help humans interact more naturally with computers. A related emerging domain of research instead views the computer as a tool to assist or understand human interactions : putting computers in the human interaction loop [1]. Humans naturally interact with other humans, communicating and generating valuable information. The most natural interface for entering this information into a computing system would therefore be for the computer to extract it directly from observing the human interactions.

The automatic analysis of human interaction is a rich research area. There is growing interest in the automatic understanding of group behaviour, where the interactions are defined by individuals playing and exchanging both similar and complementary roles (e.g. a handshake, a dancing couple, or a children's game) [2–6]. Most of the previous work has relied on visual information and statistical models, and studied three specific scenarios: surveillance in outdoor scenes [5, 6],

[1] This article is an updated version of one that originally appeared in *Proceedings of the European Symposium on Ambient Intelligence*, Springer Lecture Notes in Computer Science, November 2003.

workplaces [3, 4], and indoor group entertainment [2]. Beyond the use of visual information, dialogue modelling [7, 8] analyses the structure of interactions in conversations.

While it has only recently become an application domain for computing research, observation of human interactions is not a new field of study - it has been actively researched for over fifty years by a branch of social psychologists [9–11]. For example, research has analysed turn-taking patterns in group discussions [12–14], giving insight into issues such as interpersonal trust, cognitive load in interactions, and patterns of dominance and influence [11]. Research has also shown that interactions are fundamentally multimodal, with participants coordinating speaking turns using a variety of cues, such as gaze, speech back-channels, changes in posture, etc. [12, 13, 15]. In general, visual information can help disambiguate audio information [16], and when the modalities are discrepant, participants appear to be more influenced by visual than by audio cues [11, 17].

Motivated therefore by a desire to move towards more natural human-machine interfaces, and building upon findings of social psychologists regarding the mechanisms and significance of human interactions, this article presents an observational framework for computer understanding of human interactions, focussing on small group meetings as a particular instance.

Meetings contain many complex interactions between people, and so automatic meeting analysis presents a challenging case study. Speech is the predominant modality for communication in meetings, and speech-based processing techniques, including speech recognition, speaker identification, topic detection, and dialogue modelling, are being actively researched in the meeting context [18, 8, 19, 20]. Visual processing, such as tracking people and their focus of attention, has also been examined in [1, 21]. Beyond this work, a place for analysis of text, gestures, and facial expressions, as well as many other audio, visual and multimodal processing tasks can be identified within the meeting scenario. While important advances have been made, to date most approaches to automatic meeting analysis have been limited to the application of known technologies to extract information from individual participants (e.g. speech, gaze, identity, etc). Intuitively, the true information of meetings is created from interactions between participants, and true understanding of meetings can only emerge from considering their group nature.

The remainder of this article is organised as follows. Section 2 describes a multi-sensor meeting room that we have installed to enable our research. A framework for computer understanding of human interactions is outlined in Section 3, along with some specific issues and algorithms related to the meeting context. Finally, some perspective on future directions in automatic meeting analysis is given in Section 4, followed by concluding remarks in Section 5.

2 A Multi-Sensor Meeting Room

As mentioned above, interactions between people in meetings are generally multimodal in nature. While the *audio* modality is the most obvious source of information in discussions, studies have shown that significant information is conveyed in the *visual* modality, through expressions, gaze, gestures and posture [12, 13, 15]. In meetings, the *textual* modality is also important, with presentation slides, whiteboard activity, and shared paper documents providing detailed information.

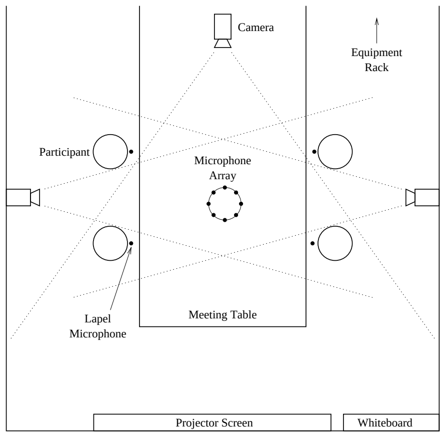

Fig. 1. Meeting recording configuration

To facilitate research into automatic meeting analysis, a meeting room at IDIAP has been equipped with multi-media acquisition facilities for recording meetings with up to 4 participants. Audio information is captured from both headset and lapel microphones on each participant, a tabletop microphone array, and a binaural manikin. Video information is collected using seven cameras. Four cameras are positioned in the centre of the meeting table, providing close-up facial views of each participant with sufficient resolution for tasks such as face identification and audio-visual speech recognition. The three remaining cameras acquire wider angle frontal views of the participants and a view of the entire meeting room scene. Unique presentation slides are captured at native VGA resolutions from the monitoring output of a data projector, whiteboard activity is recorded using transmitting pens and a receiver attached to a standard whiteboard, and participants' notes are acquired using a digital pen capture system. The acquisition of all modalities is completely synchronised and all data streams are accurately time-stamped.

Meeting recording efforts at IDIAP have occurred at various stages in the evolution of the meeting room acquisition capabilities. An initial audio-visual corpus of approximately sixty, five-minute, four-person scripted meetings was acquired using three wide-angle cameras, per-participant lapel microphones and a microphone array. Subsequent recordings focussed on the recording of less constrained and more naturally occurring meeting scenarios and used the same A/V sensor configuration together with slide and whiteboard capture capabilities. The meeting room configuration used for these recordings is illustrated in Figure 1. The resulting meeting recordings have been annotated to differing degrees, and all raw and meta- data is available for online public distribution through a MultiModal Media file server at `mmm.idiap.ch`. A new round of meeting recordings has been recently launched using the full multimodal acquisition capabilities. This round of recordings (in conjunction with recordings from two partner sites) aims to collect 100 hours of annotated meeting data to satisfy the multimodal meeting data needs of the AMI research consortium[2].

3 Multimodal Processing

We propose a framework for computer understanding of human interactions that involves the following basic steps in a processing loop :

```
1. locate and track participants
2. for each located participant
   (a) enhance their audio and visual streams
   (b) identify them
   (c) recognise their individual actions
3. recognise group actions
```

The first step is necessary to determine the number and location of participants. For each person present, we then extract a dedicated enhanced audio and visual stream by focussing on their tracked location. Audio-visual (speech and face) speaker identification techniques can then be applied to determine who the participant is. Individual actions, such as speech activity, gestures or speech words may also be measured or recognised from the audio and visual streams. The ultimate goal of this analysis is then to be able to recognise actions belonging to the group as a whole, by modelling the interactions of the individuals.

Specific issues and algorithms for implementing a number of these steps for the case of meeting analysis are presented in the following sub-sections. A primary focus of our research is the multimodal nature of human interactions in meetings, and this is reflected in the choice of tasks we have included. Naturally, there are many other processing tasks involved in understanding meetings, such as speech recognition and dialogue modelling, that are not covered here.

[2] `http://www.amiproject.org/`

3.1 Audio-Visual Speaker Tracking

The problem in the global view Locating and tracking speakers represents an important first step towards automatic understanding of human interactions. As mentioned previously, speaker turn patterns convey a rich amount of information about the behaviour of a group and its individual members [10, 13]. Furthermore, experimental evidence has highlighted the role that non-verbal behaviour (gaze, facial expressions, and body postures) plays in interactions [13]. Recognising such rich multimodal behaviour first requires reliable localisation and tracking of people.

Challenges in the meeting context The separate use of audio and video as cues for tracking are classic problems in signal processing and computer vision. However, sound and visual information are jointly generated when people speak, and provide complementary advantages. While initialisation and recovery from failures can be addressed with audio, precise object localisation is better suited to visual processing.

Long-term, reliable tracking of multiple people in meetings is challenging. Meeting rooms pose a number of issues for audio processing, such as reverberation and multiple concurrent speakers, as well as for visual processing, including clutter and variations of illumination. However, the main challenge arises from the behaviour of multiple participants resulting in changes of appearance and pose for each person, and considerable (self)-occlusion. At the same time, meetings in a multi-sensor room present some advantages that ease the location and tracking tasks. Actions usually unfold in specific areas (meeting table, whiteboard, and projector screen), which constrains the group dynamics in the physical space. In addition, the availability of multiple cameras with overlapping fields of view can be exploited to build more reliable person models, and deal with the occlusion problems.

Our approach We are developing principled methods for speaker tracking, fusing information coming from multiple microphones and uncalibrated cameras [22], based on *Sequential Monte Carlo* (SMC) methods, also known as *particle filters* (PFs) [23]. For a state-space model, a PF recursively approximates the conditional distribution of states given observations using a dynamical model and random sampling by (i) generating candidate configurations from the dynamics (*prediction*), and (ii) measuring their likelihood (*updating*), in a process that amounts to random search in a configuration space.

The state-space formulation is general. As an option, it can be defined over only one person, implying that the tracker should lock onto the current speaker at each instant. More generally, the state-space could be defined over all the people present in the scene. In this joint representation, both the location and the speaking status of each participant should be tracked all the time.

Our work is guided by inherent features of AV data, taking advantage of the fact that data fusion can be introduced in both stages of the PF algorithm. First, audio is a strong cue to model discontinuities that clearly violate usual assumptions in dynamics (including speaker turns across cameras), and (re)initialisation. Its use for prediction thus brings benefits to modelling real situations. Second, audio can be inaccurate at times, but provides a good initial localisation guess that can be enhanced by visual information. Third, although audio might be imprecise, and visual calibration can be erroneous due to distortion in wide-angle cameras, the joint occurrence of AV information in the constrained physical space in meetings tends to be more consistent, and can be learned from data.

Our methodology exploits the complementary features of the AV modalities. In the first place, we use a 2-D approach in which human heads are visually represented by their silhouette in the image plane, and modelled as elements of a *shape-space*, allowing for the description of a head template and a set of valid geometric transformations (motion). In the second place, we employ *mixed-state* space representations, where in addition to the continuous subspace that represents head motion, we also include discrete components. In a multi-camera setup, a discrete variable can indicate the specific camera plane in which a speaker is present, thus helping define a generative model for camera switching. For a multi-object state space, discrete variables are additionally used to indicate the speaking/non-speaking status of each participant. In the third place, we asymmetrically handle audio and video in the PF formulation. Audio localisation information in 3-D space is first estimated by an algorithm that reliably detects speaker changes with low latency, while maintaining good estimation accuracy. Audio and skin-color blob information are then used for prediction, and introduced in the PF via *importance sampling*, a technique which guides the search process of the PF towards regions of the state space likely to contain the true configurations. Additionally, audio, color, and shape information are jointly used to compute the likelihood of candidate configurations. Finally, we use an AV calibration procedure to relate audio estimates in 3-D and visual information in 2-D. The procedure uses easily generated training data, and does not require precise geometric calibration of cameras and microphones [22].

When applied to the single-object state-space, the particle filtering framework results in a method that can initialise and track a moving speaker, and switch between multiple people across cameras with low delay, while tolerating visual clutter. An example for the setup of Figure 1 is shown in Figure 2, for a two-minute sequence, using 500 particles. Given a ground-truth of speaker segments, which consists of the camera index and the approximate speaker's head centroid in the corresponding image plane for each speaker segment, Table 1 shows that the percentage of error on the estimated camera indices ϵ_k is quite small for the close-view cameras, but larger for the wide-view case. Additionally, the median localisation error in the image plane $\epsilon_{(T^x,T^y)}$ (in pixels) remains within a few pixels, and is smaller than the error obtained using only the audio modality, thus justifying a multimodal approach. Other AV tracking examples for single- and multi-camera set-ups can be found at `www.idiap.ch/~gatica`.

An example of the joint multi-object tracking system is shown in Fig. 3 for the case of non-overlapped views, using 1000 particles. The four participants

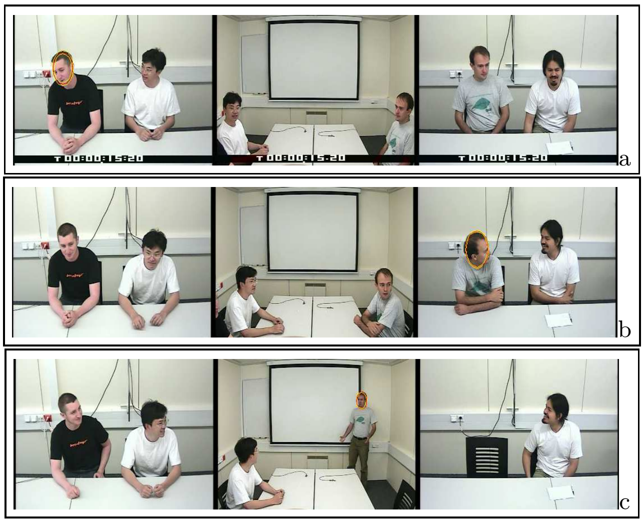

Fig. 2. Single-object speaker tracker in the meeting room. The tracker locks onto one speaker.

error type	modality	cam_1	cam_2	cam_3	global
$\epsilon_k(\times 10^{-2})$	AV	1.91	0.31	25.00	11.27
$\epsilon_{(T^x,T^y)}$	AV	1.88	1.69	0.40	1.00
	A	11.39	11.86	10.60	11.20

Table 1. Single-object AV speaker tracking results. For full details of techniques and experimental conditions, see [22].

are simultaneously tracked, and their speaking status is inferred at each time. In practice, the multi-object tracker significantly requires more computational resources given the joint object representation. Refinements of the approach, and the evaluation of the algorithms are part of current work.

Open problems Although the current approaches are useful in their current form, there is much room for improvement. In the following we identify three specific lines of research. We will extend the multi-object tracker to a multi-camera scenario with overlapping fields of view, which involves the consistent labelling of tracked objects across cameras. In the second place, a joint state-space representation for multi-object tracking significantly increases the dimensionality of the state space, which calls for efficient inference mechanisms in the resulting statistical model. We have made some progress in this direction [24]. The third line of research is the joint formulation of tracking and recognition. We are conducting research on head trackers that simultaneously estimate head orientation (a simple form of recognition), which is in turn a strong cue for detection of focus of attention, and useful for higher-level recognisers [25].

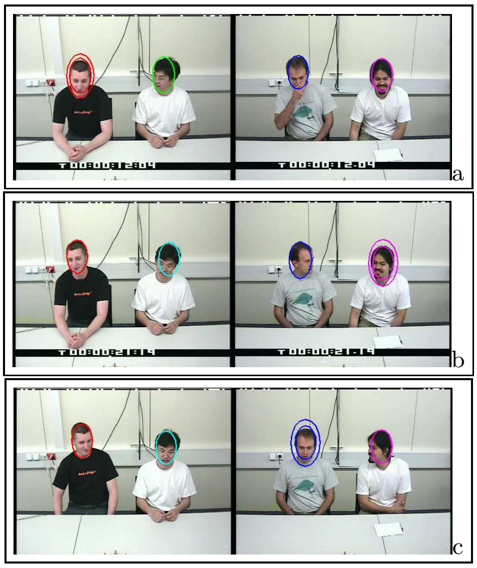

Fig. 3. Multi-object speaker tracker in the meeting room. The speaking status is inferred for each participant, a speaker is shown with a double ellipse.

3.2 Speech Segmentation and Enhancement using Microphone Arrays

The problem in the global view Having located and tracked each person, it is next necessary to acquire an enhanced dedicated audio channel of their speech. Speech is the predominant communication modality, and thus a rich source of information, in many human interactions.

Most state-of-the-art speech and speaker recognition systems rely on close-talking head-set microphones for speech acquisition, as they naturally provide a higher signal-to-noise ratio (SNR) than single distant microphones. This mode of acquisition may be acceptable for applications such as dictation, however as technology heads towards more pervasive applications, less constraining solutions are required. *Microphone arrays* present a promising alternative to close-talking microphones, as they allow for signal-independent enhancement, localisation and tracking of speakers, and non-intrusive hands-free operation. For these reasons, microphone arrays are being increasingly used for speech acquisition in such applications [26, 27].

Challenges in the meeting context Meetings present a number of interesting challenges for microphone array research. A primary issue is the design of the *array geometry* : how many microphones should be used, and where should they be placed in the room? Naturally a geometry giving high spatial resolution uniformly across a room is desirable for best performance and lowest constraint on the users, however this requires prohibitively large numbers of microphones, and complex installation [28]. For these reasons, more practical solutions with smaller numbers of microphones need to be researched to address computational and economical considerations.

A second challenge in the meeting context is the natural occurrence of overlapping speech. In [29] it was identified that around 10-15% of words, or 50% of speech segments, in a meeting contain a degree of overlapping speech. These overlapped segments are problematic for speaker segmentation, and speech and speaker recognition. For instance, an absolute increase in word error rate of between 15-30% has been observed on overlap speech segments using close-talking microphones [29, 8].

Our approach While it is clear that a large microphone array with many elements would give the best spatial selectivity for localisation and enhancement, for microphone arrays to be employed in practical applications, hardware cost (microphones, processing and memory requirements) must be reduced. For this reason, we focus on the use of small microphone arrays, which can be a viable solution when assumptions can be made about the absolute and relative locations of participants.

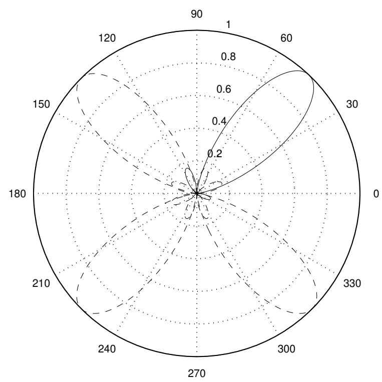

Fig. 4. Microphone array directivity patterns at 1000 Hz (speaker 1 direction in bold)

As shown in Figure 1, the particular array geometry we have chosen is an 8-element circular array (of radius 10cm) placed at the centre of the meeting table. This geometry and placement was selected based on the assumption that a meeting generally consists of small groups of people seated and talking face to face in well-defined regions. Each array is designed to cater for a small group of up to 4 people. In larger meetings, multiple (potentially interacting) small array modules are positioned along the table, where each module is responsible for the people in its local region. The circular geometry was selected as it gives uniform spatial selectivity between people sitting around it, leading to good general performance in separating overlapping speech. This is important for meetings where background noise is generally low, and so overlapping speech is the primary noise source. To illustrate, Figure 4 shows the theoretical *directivity pattern* (array gain as a function of direction) for the array at 1000 Hz for 4 speakers separated by 90 degrees. Having the array on the table also means it is placed in close proximity to participants, leading to naturally high signal levels compared to background noise caused by distant sources.

Given accurate tracking of the speaker locations in the room, the next task is to determine segments of continuous speech from a given speaker location. *Speaker segmentation* in meetings is problematic for traditional techniques based on simple energy or spectral features, as a significant amount of cross-talk from other speakers exists even on close-talking microphones [30, 31]. In [32, 33] we presented a *location-based segmentation* technique that is capable of providing a smooth speech/silence segmentation for a specified location in a room. As it is based on speech location features from the microphone array, rather than standard spectral features, this location-based segmentation has the important benefit of being able to accurately handle multiple concurrent speakers (identifying which locations are active at any given time). In [34], this segmentation algorithm was integrated with automatic speaker tracking and tested on a set of 17 short (5 minute) meetings recorded in the room described in Section 2. Results of these experiments are summarised in Table 2. Results are in terms of the common precision (PRC), recall (RCL) and F measures ($F = \frac{2 \times PRC \times RCL}{PRC + RCL}$). The location-based technique is compared to a baseline energy-based approach using lapel microphones. The results show that, while the location-based approach yields comparable overall segmentation accuracy, it achieves a significant improvement during periods of overlapping speech (recall increasing from 66% to 85%, precision from 47% to 55%). Full experimental details and discussion can be found in [34].

Metric	Location-based	Lapel baseline
PRC	79.7 (55.4)	84.3 (46.6)
RCL	94.6 (84.8)	93.3 (66.4)
F	86.5 (67.0)	88.6 (54.7)

Table 2. Segmentation results on 17 meetings. The location-based approach uses distant microphones only. Values are percentages, results on overlaps only are indicated in brackets. Table reproduced from [34].

Once the location of the speakers is known along with their speech activity segmentation, we can then apply microphone array *beamforming* techniques to enhance their speech, attenuating background noise and conflicting speech sources. Beamforming consists of filtering and combining the individual microphone signals in such a way as to enhance signals coming from a particular location. For beamforming filters, we adopt standard *superdirective* filters, which are calculated to maximise the array gain for the desired direction [35]. In addition, we apply a *Wiener post-filter* to the beamformer output to further reduce the broadband noise energy. The post-filter is estimated from the auto- and cross-spectral densities of the microphone array inputs, and is formulated assuming a diffuse background noise field [36]. This post-filter leads to significant improvements in terms of SNR and speech recognition performance in office background noise [36].

To assess the effectiveness of the beamformer in improving eventual speech recognition performance, a multi-microphone corpus was recorded for experimentation and public distribution. The database was collected by outputting utterances from the Numbers corpus (telephone quality speech, 30-word vocabulary) on one or more loudspeakers, and recording the resulting sound field using a microphone array and various lapel and table-top microphones. The goal of this work was to compare relative speech recognition performance using different microphone configurations in various noise situations, and thus a small vocabulary corpus was considered appropriate. Initial results on this corpus (MONC: Multi-channel Overlapping Numbers Corpus, available from the Center for Spoken Language Understanding at OGI) were presented in [37], and are reproduced in Table 3. These results show that the array processing significantly improves over a single distant microphone (centre), and also over a close-talking lapel microphone in situations where there is significant overlapping speech between speakers.

Simultaneous Speakers	Lapel	Centre	Array
1	7.01	10.06	7.00
2	24.43	57.56	19.31
3	35.25	73.55	26.64

Table 3. Word error rate results for speech recorded on a close-talking lapel microphone, a microphone placed in the centre of the meeting table, and the output of a microphone array beamformer. For full details of techniques and experimental conditions, see [37]

Open problems While microphone array speech processing techniques are already relatively mature, a number of open issues remain in this context. One of these is the need to focus on algorithms that handle multiple concurrent, moving, speakers. While work cited in this paper indicates progress in this direction, there remains a need for testing of multi-speaker localisation, tracking and beamforming in real applications, such as large vocabulary speech recognition in meetings. Another interesting research direction is the use of multiple interacting small microphone array modules to cover arbitrary areas, instead of using a single larger array.

3.3 Audio-Visual Person Identification

The Problem in the Global View Identifying participants is important for understanding human interactions. When prior knowledge about the participants is available (such as their preferred way of communicating, topics of interests, levels of language, relative hierarchical levels in a given context, etc), knowing the participants' identities would imply knowing this prior information, which could in turn be used to better tune the algorithms used to analyse the interaction. Fortunately, *biometric authentication* [38], which is the general problem of authenticating or identifying a person using his or her behavioural and physiological characteristics such as the face or the voice, is a growing research domain which has already shown useful results, especially when using more than one of these characteristics, as we propose to do here.

Challenges in the Meeting Context In order to perform AV identification during a meeting, we need to extract reliably the basic modalities. For the face, we require a face localisation algorithm that is robust to the kind of images available from a video stream (relatively low-quality and low-resolution), robust to the participants' varying head poses, and able to cope with more than one face per image. This could be done using our AV tracking system described in Section 3.1. For the voice, taking into account that several microphones are available in the meeting room, the first challenge is to separate all audio sources and attribute each speech segment to its corresponding participant. Again, this could be done using our speaker segmentation and enhancement techniques, described in Section 3.2. Afterward, classical face and speaker verification algorithms could be applied, followed by a fusion step, which provides robustness to the failure of one or the other modality. Finally, an identification procedure could be applied.

Our Approach Our identification system is based on an AV biometric verification system. Assuming that we are able to obtain reliable speech segments and localised faces from the meeting raw data, we can then apply our state-of-the-art verification system, which is based on a *speaker verification* system, a *face verification* system, and a *fusion* module.

Our speaker verification system first starts by extracting useful features from the raw speech data: we extract 16 Mel scale Frequency Cepstral Coefficient (MFCC) features every 10 ms, as well as their first temporal derivative, plus the first derivative of the log energy of the signal. Then, a silence detector based on an unsupervised 2-Gaussian system is used to remove all silence frames. Finally, the verification system itself is based on the modelling of one Gaussian Mixture Model (GMM) for each individual, adapted using *Maximum A Posteriori* (MAP) techniques from a *World Model* trained by *Expectation-Maximisation* on a large set of prior data. The score for a given access is obtained as the logarithm of the ratio between the likelihood of the data given the individual model and the likelihood given the world model. This system obtains state-of-the-art performance on several benchmark verification databases [39].

Our face verification system is comprised of two main parts: an automatic face locator and a local feature probabilistic classifier. To locate faces, a fast cascade of boosted Haar-like features is applied to the integral image to detect potential faces [40], followed by post-processing using a Multi-Layer Perceptron [41] to provide the final localized face. The probabilistic classifier uses DCTmod2 features [42] and models faces using pseudo-2D Hidden Markov Models (HMMs) [43]. In DCTmod2 feature extraction, each given face is analyzed on a block by block basis; from each block a subset of Discrete Cosine Transform (DCT) coefficients is obtained; coefficients which are most affected by illumination direction changes are replaced with their respective horizontal and vertical deltas, computed as differences between coefficients from neighbouring blocks. For the pseudo-2D HMM topology, we use a top-to-bottom main HMM with each state being modeled by a left-to-right HMM. Parameters for each client model are obtained via Maximum *a Posteriori* (MAP) adaptation of a generic face HMM; the generic face HMM is in turn trained using the Expectation Maximization algorithm, on a large generic dataset. As for the speech system, a score for a given face is found by taking the logarithm of the ratio between the likelihood of the face belonging to the true client and the likelihood of the face belonging to the impostor model.

Our fusion algorithm is based on Multi-layer Perceptrons (experiments with Support Vector Machines give similar performances). The fusion model takes as input the log likelihood scores coming from both the face and the speaker verification systems, and combines them non-linearly in order to obtain a unified and more robust overall score. Optionally, confidence values could also be computed on both the voice and face scores, which then enhance the quality of the fusion model [44].

Finally, in order to identify the correct individual, the whole verification system is run over all previously stored individual models, and the model corresponding to the highest obtained score over a pre-defined threshold (in order to account for unknown individuals) identifies the target individual.

While we currently do not have results in the context of meetings, we did apply them on several benchmark databases and always obtained state-of-the-art performance. For instance, Table 4 shows the performance of our models on the difficult but realistic audio-visual BANCA database [45], using protocol P of the English subset, and measured in terms of *half total error rate* (HTER), which is the average of the rates of false acceptances and false rejections.

Voice	Face	Fusion
4.7%	20.9%	2.8%

Table 4. Verification performance on the English subset of the BANCA database, protocol P, in terms of HTER (the lower, the better).

We can see from this Table that speaker verification is in general more robust than face verification, and that fusing both of them still increases the overall performance. We note that this face verification system ranked first in a recent international competition on this corpus [46].

Open Problems Assuming that speaker segmentation and face tracking have given perfect segmentation, for a given meeting, we will have potentially several minutes of speech and face data per individual. In general, a classical verification system only requires a few face images and less than one minute of speech data to attain acceptable performance. However, the environment is unconstrained, the meeting data may be noisy for different reasons - the individual may not always look at the camera and speak loudly and intelligibly. In this case, rather than using all available data to identify a person, a better solution could be to be more strict on the selection of faces and speaker segments in order to keep only the best *candidates* for identification. Hence, we should try to remove highly noisy or overlapping speech segments, badly tracked face images and faces that are not in a good frontal pose and good lighting condition.

3.4 Group Action Recognition

The problem in the global view The ultimate goal of automatic analysis of human interactions is to recognise the group actions. As discussed previously, the true information of meetings is created from interactions between participants playing and exchanging roles. In this view, an important goal of automatic meeting analysis is the segmentation of meetings into high-level agenda items which reflect the action of the group as a whole, rather than just the behaviour of individuals (e.g. discussions and presentations, or even higher level notions, like planning, negotiating, and making decisions).

Challenges in the meeting context Recognition of group actions in meetings entails several important problems for which no satisfactory solutions currently exist. These include (1) devising tractable multi-stream sequence models, where each stream could arise from either a modality (AV) or a participant; (2) modelling asynchronicity between participants' behaviour; (3) extracting features for recognition that are robust to variations in human characteristics and behaviour; (4) designing sequence models that can integrate language features (e.g. keywords or dialog acts) with non-verbal features (e.g. emotion as captured from audio and video); and (5) developing models for recognition of actions that are part of a hierarchy.

One potentially simplifying advantage to recognise group actions in meetings is that participants usually have some influence on each other's behaviour. For example, a dominant speaker grabbing the floor often makes the other participants go silent, and a presentation will draw most participants' attention in the same direction. The recognition of some group actions can be therefore benefit from the occurrence of these multiple similar individual behaviours.

Our approach We have addressed meeting group action recognition as the recognition of a continuous, non-overlapping, sequence of lexical entries, analogous to observational approaches in social psychology for analysis of group interaction [10], and to speech or continuous gesture recognition [47, 48]. Continuous recognition generates action-based meeting segmentations that can be directly used for browsing. Furthermore, the definition of multiple lexica would provide alternative semantic views of a meeting. Note that in reality, most group actions are characterised by soft (natural) transitions, and specifying their boundaries beyond a certain level of precision has little meaning.

In particular, we have modelled meeting actions based on a set of multimodal turn-taking events. Speaking turns are mainly characterised by audio information, but significant information is also present in non-verbal cues like gaze and posture changes [13], which can also help disambiguate audio information [16]. The specific actions include monologues (one participant speaks continuously without interruption), discussions (all participants engage in a discussion), presentations (one participant at front of room makes a presentation using the projector screen), white-boards (one participant at front of room talks and uses the white-board), and group note-taking (all participants write notes).

In a first approach [49], we used a number of Hidden Markov Model (HMM) variants to recognise the group actions by direct modelling of low-level features. The models investigated included early integration HMMs [47], multi-stream HMMs [50], asynchronous HMMs [51], and coupled HMMs [52]. Features were extracted from both audio and visual modalities, and included speech activity, pitch, energy, speaking rate, and head and hand location and motion features. For experiments, we used the meeting corpus described in Section 2. Meetings followed a loose script to ensure an adequate amount of examples of all actions, and to facilitate annotation for training and testing, but otherwise the individual and group behaviour is natural.

A detailed account of the experiments and results can be found in [49], but we repeat the summarised findings here:

1. The best system achieves an action error rate (equivalent to word error rate in ASR) of 8.9%.
2. There is benefit in a multi-modal approach to modelling group actions in meetings.
3. It is important to model the correlation between the behaviour of different participants.
4. There is no significant asynchrony between audio and visual modalities for these actions (at least within the resolution of the investigated frame rate).
5. There is evidence of asynchrony between participants acting within the group actions.

These findings appeal to the intuition that individuals act in a group through both audio and visual cues which can have a causal effect on the behaviour of other group members.

More recently, a two-layer HMM framework was proposed in [53]. The first layer HMM (individual-level) recognises a small set of individual actions for each

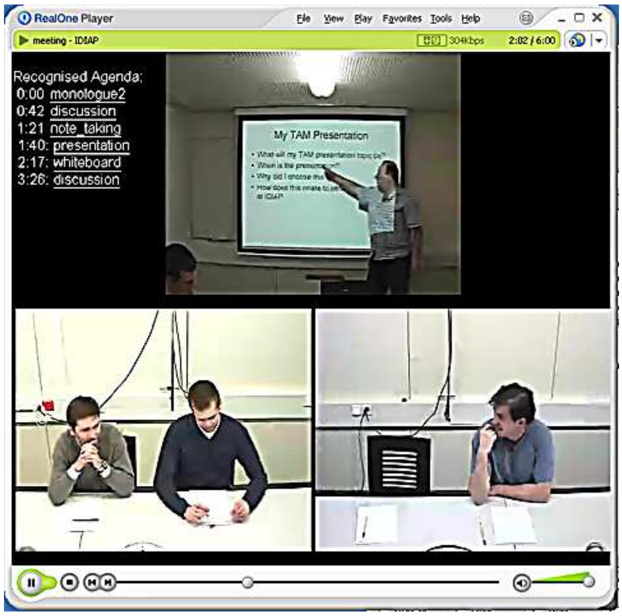

Fig. 5. Simple meeting browser interface, showing recognised meeting actions.

participant (speaking, writing, idle) using the same set of low-level audio-visual features described above. The results of these first layer HMMs are concatenated and modelled by a second layer HMM (group-level), which then attempts to recognise the group actions. For an augmented set of group actions (discussion, monologue, monologue + note-taking, note-taking, presentation, presentation + note-taking, white-board and white-board + note-taking), the two-layer system achieved an action error rate of only 15.1%, compared with a 23.7% error rate on the same task using the best single-layer HMM system (equivalent to those proposed in [49]: the higher error rate is due to the increased lexicon size). Full experimental details can be found in [53].

An example of the application of the action recognition results for meeting browsing is shown in Figure 5.

Open problems The experience gained from our results confirms the importance of modelling the interactions between individuals, as well as the advantage of a multimodal approach for recognition. We believe there is much scope for work towards the recognition of different sets of high-level meeting actions, including other multimodal turn-taking events, actions based on participants' mood or level of interest, and multimodal actions motivated by traditional dialogue acts. To achieve this goal, ongoing and future work will investigate richer feature sets, and appropriate models for the interactions of participants. Another task will be to incorporate prior information in the recognition system, based on the participant identities and models of their personal behaviour. We also plan to collect a larger meeting corpus, and work on the development of more flexible assessment methodologies.

4 Future Directions

From the framework outlined in the beginning of Section 3, while much room clearly remains for new techniques and improvements on existing ones, we can see that steps `1-2(c)` are reasonably well understood by the state-of-the-art. In contrast, we are far from making similar claims regarding step `3`, recognition of group actions.

The first major goal in computer understanding of group actions, is to clearly identify lexica of such actions that may be recognised. A simple lexicon based on multimodal turn-taking events was discussed in Section 3.4, however there is a need to progress towards recognition of higher level concepts, such as decisions, planning, and disagreements. In this regard, the social psychology literature represents an important source of information for studies on the tasks and processes that arise from human interactions, as was discussed in [49].

Having identified relevant group actions, a further research task is then to select appropriate features for these actions to be recognised. At this moment, features are intuitively selected by hand, which has obvious limitations. Approaches for feature selection could arise from two areas. The first one is human. We require a deeper understanding of human behaviour. Existing work in psychology could provide cues for feature selection towards, for example, multimodal recognition of emotion [54]. The second one is computational. Developments in machine learning applied to problems in vision and signal processing point to various directions [40].

Finally, to recognise the group actions, there is a need to propose models capable of representing the interactions between individuals in a group (see e.g. [55, 5, 49]). Some particular issues are the need to model multiple data streams, asynchronicity between streams, hierarchies of data and events (e.g. building on [53]), as well as features of different nature (e.g. discrete or continuous).

5 Conclusion

This article has discussed a framework for computer understanding of human interactions. A variety of multimodal sensors are used to observe a group and extract useful information from their interactions. By processing the sensor inputs, participants are located, tracked, and identified, and their individual actions recognised. Finally, the actions of the group as a whole may be recognised by modelling the interactions of the individuals.

While initial work in this direction has already shown promising progress and yielded useful results, it is clear that many research challenges remain if we are to advance towards true computer understanding of human interactions.

6 Acknowledgements

The authors would like to acknowledge our colleagues at IDIAP involved in the research described in this article, in particular Guillaume Lathoud, Dong Zhang, Norman Poh, Johnny Mariéthoz, Sebastien Marcel, Conrad Sanderson, Olivier Masson, Pierre Wellner, Mark Barnard, Kevin Smith, Sileye Ba, Jean Marc Odobez and Florent Monay.

This work was supported by the Swiss National Science Foundation 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).

References

1. Waibel, A., Schultz, T., Bett, M., Malkin, R., Rogina, I., Stiefelhagen, R., Yang, J.: SMaRT:the Smart Meeting Room Task at ISL. In: Proc. IEEE ICASSP 2003. (2003)
2. Bobick, A., Intille, S., Davis, J., Baird, F., Pinhanez, C., Campbell, L., Ivanov, Y., Schutte, A., Wilson, A.: The KidsRoom: A Perceptually-Based Interactive and Immersive Story Environment. PRESENCE: Teleoperators and Virtual Environments **8** (1999)
3. Johnson, N., Galata, A., Hogg, D.: The acquisition and use of interaction behaviour models. In: Proc. IEEE Int. Conference on Computer Vision and Pattern Recognition. (1998)
4. Jebara, T., Pentland, A.: Action reaction learning: Automatic visual analysis and synthesis of interactive behaviour. In: Proc. International Conference on Vision Systems. (1999)
5. Oliver, N., Rosario, B., Pentland, A.: A bayesian computer vision system for modeling human interactions. IEEE Transactions on Pattern Analysis and Machine Intelligence **22** (2000)
6. Hongeng, S., Nevatia, R.: Multi-agent event recognition. In: Proc. IEEE Int. Conference on Computer Vision, Vancouver (2001)
7. Carletta, J., Isard, A., Isard, S., Kowtko, J., Doherty-Sneddon, G., Anderson, A.: The coding of dialogue structure in a corpus. In Andernach, J., van de Burgt, S., van der Hoeven, G., eds.: Proceedings of the Twente Workshop on Language Technology: Corpus-based approaches to dialogue modelling. Universiteit Twente (1995)
8. Morgan, N., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Janin, A., Pfau, T., Shriberg, E., Stolcke, A.: The meeting project at ICSI. In: Proc. of the Human Language Technology Conference, San Diego, CA (2001)
9. Bales, R.F.: Interaction Process Analysis: A method for the study of small groups. Addison-Wesley (1951)
10. McGrath, J.E.: Groups: Interaction and Performance. Prentice-Hall (1984)
11. McGrath, J., Kravitz, D.: Group research. Annual Review of Psychology **33** (1982) 195–230
12. Padilha, E., Carletta, J.C.: A simulation of small group discussion. In: EDILOG. (2002)

13. Parker, K.C.H.: Speaking turns in small group interaction: A context-sensitive event sequence model. Journal of Personality and Social Psychology **54** (1988) 965–971
14. Fay, N., Garrod, S., Carletta, J.: Group discussion as interactive dialogue or serial monologue: The influence of group size. Psychological Science **11** (2000) 487–492
15. Novick, D., Hansen, B., Ward, K.: Coordinating turn-taking with gaze. In: Proceedings of the 1996 International Conference on Spoken Language Processing (ICSLP-96). (1996)
16. Krauss, R., Garlock, C., Bricker, P., McMahon, L.: The role of audible and visible back-channel responses in interpersonal communication. Journal of Personality and Social Psychology **35** (1977) 523–529
17. DePaulo, B., Rosenthal, R., Eisenstat, R., Rogers, P., Finkelstein, S.: Decoding discrepant nonverbal cues. Journal of Personality and Social Psychology **36** (1978) 313–323
18. Kubala, F.: Rough'n'ready: a meeting recorder and browser. ACM Computing Surveys **31** (1999)
19. Waibel, A., Bett, M., Metze, F., Ries, K., Schaaf, T., Schultz, T., Soltau, H., Yu, H., Zechner, K.: Advances in automatic meeting record creation and access. In: Proc. IEEE ICASSP, Salt Lake City, UT (2001)
20. Renals, S., Ellis, D.: Audio information access from meeting rooms. In: Proc. IEEE ICASSP 2003. (2003)
21. Cutler, R., Rui, Y., Gupta, A., Cadiz, J., Tashev, I., He, L., Colburn, A., Zhang, Z., Liu, Z., Silverberg, S.: Distributed meetings: A meeting capture and broadcasting system. In: Proc. ACM Multimedia Conference. (2002)
22. Gatica-Perez, D., Lathoud, G., McCowan, I., Odobez, J.M.: A mixed-state i-particle filter for multi-camera speaker tracking. In: Proceedings of WOMTEC. (2003)
23. Doucet, A., de Freitas, N., Gordon, N.: Sequential Monte Carlo Methods in Practice. Springer-Verlag (2001)
24. Smith, K., Gatica-Perez, D.: Order matters: a distributed sampling method for multi-object tracking. In: IDIAP Research Report IDIAP-RR-04-25, Martigny (2004)
25. Ba, S., Odobez, J.M.: A probabilistic framework for joint head tracking and pose estimation. In: Proc. ICPR, Cambridge (2004)
26. Cutler, R.: The distributed meetings system. In: Proceedings of IEEE ICASSP 2003. (2003)
27. Stanford, V., Garofolo, J., , Michel, M.: The nist smart space and meeting room projects: Signals, acquisition, annotation, and metrics. In: Proceedings of IEEE ICASSP 2003. (2003)
28. Silverman, H., Patterson, W., Flanagan, J., Rabinkin, D.: A digital processing system for source location and sound capture by large microphone arrays. In: Proceedings of ICASSP 97. (1997)
29. Shriberg, E., Stolcke, A., Baron, D.: Observations on overlap: findings and implications for automatic processing of multi-party conversation. In: Proceedings of Eurospeech 2001. Volume 2. (2001) 1359–1362
30. Pfau, T., Ellis, D., Stolcke, A.: Multispeaker speech activity detection for the ICSI meeting recorder. In: Proceedings of ASRU-01. (2001)
31. Kemp, T., Schmidt, M., Westphal, M., Waibel, A.: Strategies for automatic segmentation of audio data. In: Proceedings of ICASSP-2000. (2000)
32. Lathoud, G., McCowan, I.: Location based speaker segmentation. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing. (2003)

33. Lathoud, G., McCowan, I., Moore, D.: Segmenting multiple concurrent speakers using microphone arrays. In: Proceedings of Eurospeech 2003. (2003)
34. Lathoud, G., Odobez, J.M., McCowan, I.: Unsupervised location-based segmentation of multi-party speech. In: Proceedings of the 2004 ICASSP-NIST Meeting Recognition Workshop. (2004)
35. Bitzer, J., Simmer, K.U.: Superdirective microphone arrays. In Brandstein, M., Ward, D., eds.: Microphone Arrays. Springer (2001) 19–38
36. McCowan, I., Bourlard, H.: Microphone array post-filter based on noise field coherence. To appear in IEEE Transactions on Speech and Audio Processing (2003)
37. Moore, D., McCowan, I.: Microphone array speech recognition: Experiments on overlapping speech in meetings. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing. (2003)
38. Jain, A., Bolle, R., Pankanti, S.: Biometrics: Person Identification in Networked Society. Kluwer Publications (1999)
39. Mariéthoz, J., Bengio, S.: A comparative study of adaptation methods for speaker verification. In: Proceedings of the International Conference on Spoken Language Processing, ICSLP. (2002)
40. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proc. IEEE Int. Conf. on Computer Vision (CVPR), Kawaii (2001)
41. Rowley, H., Baluja, S., Kanade, T.: Neural network-based face detection. IEEE Tran. Pattern Analysis and Machine Intelligence **20(1)** (1998) 23–38
42. Sanderson, C., Paliwal, K.: Fast features for face authentication under illumination direction changes. Pattern Recognition Letters **24** (2003) 2409–2419
43. Cardinaux, F., Sanderson, C., Bengio, S.: Face verification using adapted generative models. In: Proc. Int. Conf. Automatic Face and Gesture Recognition (AFGR), Seoul, Korea. (2004)
44. Bengio, S., Marcel, C., Marcel, S., Mariéthoz, J.: Confidence measures for multimodal identity verification. Information Fusion **3** (2002) 267–276
45. Bailly-Baillière, E., Bengio, S., Bimbot, F., Hamouz, M., Kittler, J., Mariéthoz, J., Matas, J., Messer, K., Popovici, V., Porée, F., Ruiz, B., Thiran, J.P.: The BANCA database and evaluation protocol. In: 4th International Conference on Audio- and Video-Based Biometric Person Authentication, AVBPA, Springer-Verlag (2003) 625–638
46. Messer, K., Kittler, J., Sadeghi, M., Hamouz, M., Kostyn, A., Marcel, S., Bengio, S., Cardinaux, F., Sanderson, C., Poh, N., Rodriguez, Y., Kryszczuk, K., Czyz, J., Vandendorpe, L., Ng, J., Cheung, H., Tang, B.: Face authentication competition on the BANCA database. In: International Conference on Biometric Authentication, ICBA. (2004)
47. Rabiner, L.R., Juang, B.H.: Fundamentals of Speech Recognition. Prentice-Hall (1993)
48. Starner, T., Pentland, A.: Visual recognition of american sign language using HMMs. In: Proc. Int. Work. on Auto. Face and Gesture Recognition, Zurich (1995)
49. McCowan, I., Gatica-Perez, D., Bengio, S., Lathoud, G.: Automatic analysis of multimodal group actions in meetings. Technical Report RR 03-27, IDIAP (2003)
50. Dupont, S., Luettin, J.: Audio-visual speech modeling for continuous speech recognition. IEEE Transactions on Multimedia **2** (2000) 141–151
51. Bengio, S.: An asynchronous hidden markov model for audio-visual speech recognition. In Becker, S., Thrun, S., Obermayer, K., eds.: Advances in Neural Information Processing Systems, NIPS 15, MIT Press (2003)
52. Brand, M.: Coupled hidden markov models for modeling interacting processes. TR 405, MIT Media Lab Vision and Modeling (1996)

53. Zhang, D., Gatica-Perez, D., Bengio, S., McCowan, I., Lathoud, G.: Modeling individual and group actions in meetings: a two-layer hmm framework. In: Proc. IEEE CVPR Workshop on Event Mining, Washington, DC (2004)
54. De Gelder, B., Vroomen, J.: The perception of emotions by ear and by eye. Cognition and Emotion **14** (2002) 289–311
55. Basu, S., Choudhury, T., Clarkson, B., Pentland, A.: Learning human interactions with the influence model. Technical Report 539, MIT Media Laboratory (2001)