Automatic Meeting Segmentation Using Dynamic Bayesian Networks

Alfred Dielmann and Steve Renals, *Member, IEEE*

***Abstract*—Multiparty meetings are a ubiquitous feature of organizations, and there are considerable economic benefits that would arise from their automatic analysis and structuring. In this paper, we are concerned with the segmentation and structuring of meetings (recorded using multiple cameras and microphones) into sequences of group meeting actions such as monologue, discussion and presentation. We outline four families of multimodal features based on speaker turns, lexical transcription, prosody, and visual motion that are extracted from the raw audio and video recordings. We relate these low-level features to more complex group behaviors using a multistream modelling framework based on multistream dynamic Bayesian networks (DBNs). This results in an effective approach to the segmentation problem, resulting in an action error rate of 12.2%, compared with 43% using an approach based on hidden Markov models. Moreover, the multistream DBN developed here leaves scope for many further improvements and extensions.**

***Index Terms*—Multimodal, multistream, meeting actions.**

I. Introduction

INVOLVEMENT in meetings is a common experience in daily life, particularly in the workplace, where managers spend more than a day each week in meetings[1]. Meetings perform several functions, such as the resolution of disputes, socialization, problem solving, planning, or the review of results. Only rarely is a meeting focused on a single task; usually, groups are engaged in multiple interdependent functions on multiple concurrent projects [1].

Traditionally the minutes of a meeting are taken by someone present at the meeting. Unfortunately, this is a time-consuming job, and often fails to capture all the required information. It would be desirable to have an automatic system to enable efficient organization, search and recall of the information contained in a meeting, or a set of meetings. Such a system would be required to extract high level information such as meeting phases, meeting tasks, textual transcriptions, topic structure, and summaries [2]. These high-level descriptions can provide a multiperspective analysis of a meeting, more detailed and more objective than a hand-made minute. Moreover such an analysis could facilitate browsing over meeting series, making it possible to search for specific events [3].

In this paper, we are concerned with the automatic structuring of meetings, based on multistream meeting recordings—primarily audio and video streams captured using multiple microphones and cameras. Analysis of natural human communication based on multiple streams corresponding to recordings of different modalities is a difficult task, since acoustic recordings are corrupted by environmental noise and room reverberations; video recordings include occlusions and environmental changes; the participant interactions are highly spontaneous and usually unconstrained; there is a very wide range of topics, speakers, speaking styles and accents.

The automatic structuring of meetings is a complex task that intersects many research areas, including automatic speech recognition, gesture recognition, topic segmentation, and emotion detection. In this work we are concerned with the recognition of *group actions*, whereby a meeting is interpreted as a sequence of interactions between the participants. Our goal is to segment automatically each recorded meeting into a sequence of group meeting actions. We have used a set of five basic group meeting actions: monologues, discussions, note taking, presentations, and whiteboard-based presentations [4]. *Monologues* are focused on an individual addressing the group, which may provide an active feedback. *Discussions*, in contrast to monologues, involve two or more participants in conversation. *Presentations* are similar to monologues, except that the orator speaks from the projection screen area. Another variant of monologues are *white-board presentations*, in which the main speaker makes use of a white-board to explain concepts. Finally, *note taking* is a group action in which participants write down their own notes. These group action symbols are assumed to be mutually exclusive and non-overlapping. Moreover, the meeting action dictionary is also assumed to be exhaustive: gaps between different actions are not allowed.

To segment a meeting into a sequence of group meeting actions, we first extract features from the multimodal recordings, then construct statistical models that represent the meeting action sequence in terms of the extracted features. We have used four main categories of features: prosodic features (such as fundamental frequency), speaker turn features, lexical features (based on a word-level transcription for each speaker), and motion-based video features. This feature extraction step may be regarded as describing a meeting as a set of streams, where each stream corresponds to a particular modality. To model this *multistream* situation, we have used dynamic Bayesian network (DBN) models in which a hierarchical state space is constructed, enabling individual feature streams to be processed

Manuscript received August 23, 2005; revised May 1, 2006. This work was supported in part by the European Union IST project M4 (IST-2001-34485) and by the EU FP6 IST Integrated Project AMI (FP6-506811, publication AMI-125). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. John R. Smith.

The authors with the Center for Speech Technology Research, University of Edinburgh, Edinburgh EH8 9LW, U.K. (e-mail: a.dielmann@ed.ac.uk; s.renals@ed.ac.uk).

Digital Object Identifier 10.1109/TMM.2006.886337

[1] 3M online survey 1998 (http://www.3m.com/meetingnetwork/)

independently at a lower, subaction level, and collectively at a higher meeting action level.

The paper is structured as follows. We review related approaches to individual and group action recognition in the following Section II. More details about the multimodal meeting recordings that we have used are given in Section III. We outline the extraction of the four feature families in Section IV. Section V contains a brief introduction to DBNs and Section VI gives a more detailed description of the multistream model that we have adopted. We present a set of experiments using these models for meeting action recognition in Section VII. Finally, conclusions are drawn and some areas for future work are outlined in Section VIII.

II. Related Work

The recording and analysis of meetings has become a flourishing research area recently, with specific foci including meeting browsing, microphone array processing, speaker tracking, and person identification [5]–[7]. Several researchers have focused on the automatic recognition of actions in meetings, at both individual and group levels.

A. *Individual Action Recognition*

Automatic interpretation of human activities, and automatic recognition of individual actions in particular application domains, is an active research field. Most of the work in this area relies on a supervised approach: unseen multimodal sequences are interpreted using statistical models estimated using annotated data. Both unimodal and multimodal approaches have been used for such problems.

Hidden Markov models (HMMs) have provided a good framework for unimodal tasks, such as speech or handwriting recognition, and usually form the baseline system for multimodal situations. Starting from the assumption that incorporating more knowledge of the underlying problem into the model can improve the model's accuracy, many HMM variants have been investigated, such as hierarchical HMMs, coupled HMMs, buried HMMs and semi-Markov models. An important feature of multimodal analysis is the requirement to process multiple asynchronous and interdependent feature streams. This may be addressed through the use of models based on multiple parallel Markov chains, usually referred to as *multistream* models. Oliver *et al.* [8], for example, proposed a structured approach to the inference of typical office user activities (e.g., making a phone call, having a face to face conversation, etc.) using features derived from audio and video signals, and computer activity logs. This approach relied on a layered HMM, which is hierarchically composed of multiple HMM chains. At the lowest level, there is a signal-analysis HMM which connects low-level features to an intermediate layer, which forms the observations for a higher level HMM, and so on up to the highest level of the model. Each layer may be trained independently (with a supervised approach) and is characterized by its own temporal granularity.

Multimodal sensing has been used to improve speech-based command and control interfaces [9], such as information kiosks or video games. Here user presence and focus of attention are inferred from low level audio, video and "contextual" features using an ad-hoc developed DBN model. A custom DBN (derived from human expertise) encoded causal relations between multimodal features (mouth motion, silence detection, skin detector, face detector, etc.) and classes that need to be recognised (visible speaker, frontal view of the speaker, focus of attention).

Automatic classification of broadcast news is another relevant example of multimodal sensing. For example Snoek *et al.* [10] proposed a framework to detect TV news monologues using multiple *style detectors* based on multimodal features (frontal face detector, video optical character recogniser, speech detector and speech recogniser) and a support-vector-machine-based classifier.

Audio-video speech recognition [11] may be viewed as a particular example of multimodal human activity recognition. This is a well-defined domain and forms a good testing ground for the comparison of different approaches and models. Dupont *et al.* [12] proposed a synchronised multistream hidden Markov model, in which the audio and video streams were processed independently. Partial recognitions were integrated only at particular state space configurations (anchor points). This multistream model was implemented by considering the whole cartesian product of the two independent stream state spaces (HMMs). Therefore state durations, anchor points, and the amount of synchronism/asynchronism between the streams were all explicitly encoded into the model's state-space structure. Another multistream approach, which took advantage of a DBN-based formalism, is outlined in [13]. This approach, using words instead of subword units as anchor points, further relaxed the assumption about stream synchronisation. Moreover, in this approach, state-duration modeling and level of synchronisation between the signals were implicitly determined.

B. *Group Action Recognition*

The literature concerning group interaction analysis using multimodal features, is much less developed than that about individual action recognition. Hakeem and Shah [14] proposed a multilevel structured approach to classify visually related meeting actions and the meeting genre. Head and hand positions were estimated using a standard condensation tracking algorithm, enhanced with a small set of categorized movement attributes. Sequences of movements were mapped into actions or events by a state machine. A hierarchical set of rules was used to detect higher level meeting activity.

Howard and Jebara [15] introduced a model for multiple concurrent processes (such as the trajectories of the members of a football team), referred to as a dynamical systems tree. This DBN model consists of a structured hierarchy of aggregating parent Markov chains (aggregating-nodes), and a set of switching linear dynamical systems that are used to discretise the continuous feature space (leaf-nodes). Basu *et al.* [16] have investigated the automatic analysis of human interaction in informal settings. Multimodal features (speaker audio activities and motion based visual activities) are related to group behaviors through a coupled HMM. Direct computations using such a model, with N chains and Q states per chain, requires NQ^N parameters, making this approach intractable even for small N. Basu *et al.* approximated the model by taking into account the

Q^2 individual interactions between a chain i and neighboring chains j, instead of considering all the Q^N possible interactions between i and the remaining $N-1$ chains.

There has been some previous work using the same corpus and dictionary of meeting actions that we employ here. [2] Reiter *et al.* [17] developed an algorithm to segment meetings in terms of meeting actions, based on a minimum length constraint and dynamic programing. Using automatic speaker segmentation and other hand labeled features, this model was used to classify segments as monologues, discussions, etc. by fusing the output of different basic classification approaches (Bayesian network, Multilayer Perceptron network, and Radial Basis network). More recently, Al-Hames *et al.* [23] proposed a framework for meeting action classification based on three multimodal features: binary speech and silence segmentation, four Mel-frequency cepstral coefficients plus energy, and a visual-based global motion vector. These features were modeled using a DBN composed of three partially coupled hidden Markov chains. Experiments applying this DBN approach to artificially perturbed pre-segmented meetings offered improved accuracy compared with a baseline HMM classifier.

McCowan *et al.* [18] investigated several approaches to multimodal feature integration and meeting action recognition, investigating both participant and group actions. Both early and late integration approaches were investigated. The best results were achieved with a group-based multistream approach [12], with good results obtained using audio features alone (speaker activity and prosodic features). These results highlighted the fact that although acoustic related features outperform video derived features (such as the positions of head and hands), a multistream approach was essential to achieving good results. This work also employed the asynchronous HMM [24] to address the task of group action recognition with a model expressly designed to cope with asynchronous multimodal signals. However, the results obtained with this model did not offer an improvement over early feature integration and a simple HMM.

More recently the same feature families have been modelled using a two-level layered HMM [20]. In this hierarchical approach, features are firstly related to participant actions (such as speaking, writing, and idle) through a low-level HMM. A higher level HMM, employing the participant action probabilities and other group level features, is then used to recognise meeting actions. This framework has been adapted to the unsupervised case [25] in which meetings (or meeting series) are segmented and clustered into a set of hidden meeting actions.

Previously, we have outlined a meeting action recognition framework based on acoustic and lexical related features and a layered multistream dynamic Bayesian network model [19], [21]. This model combines the advantages of independent feature-stream processing together with a structured approach. In this paper, we provide a clear and unified view of this framework, providing further extensions both to the feature set and to the model structure.

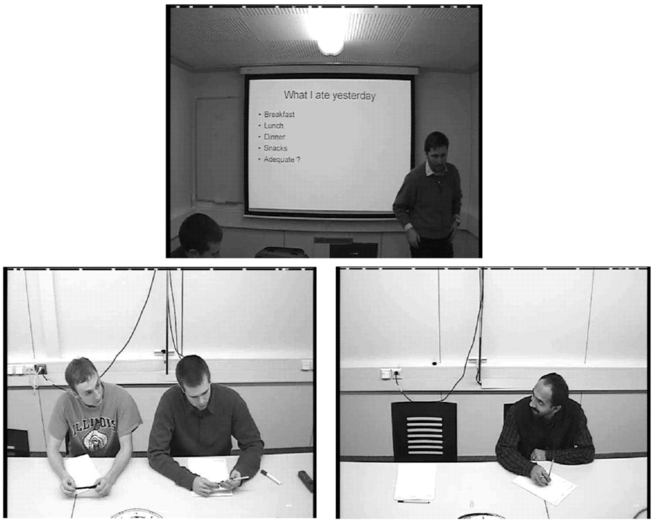

Fig. 1. Meeting scene example captured with three fixed video-cameras: whiteboard and projector screen region (top image) and two opposite sides of the table (bottom images).

III. The M4 Meeting Corpus

We performed our experiments using a corpus of 69 short meetings, recorded at IDIAP as part of the M4 project, referred to as the M4 Meeting Corpus [4][3]. Each recording in the corpus captures the interaction of four participants following an overall meeting structure that was planned in advance. The structure is defined in terms of a sequence of meeting actions from the dictionary outlined in Section I: monologue, discussion, presentation, presentation at whiteboard, and notetaking. The resultant meetings thus follow a high-level "script", but the individual participant behaviors and language are unscripted and natural. The boundaries between meeting phases tend to be smooth and spread over several seconds.

The corpus consists of more than five hours of synchronized multichannel audio-video recordings. Recordings took place in an instrumented meeting room. Each participant wore a lapel microphone, and a eight-element circular microphone array was placed on the table between participants. Note that nothing was done to prevent reverberation or to reduce environmental noise, thus offering realistic recording conditions. Orthographic (word-level) transcriptions were provided for 30 of the 69 meetings. Three fixed cameras provided visual recordings of the meeting activity (Fig. 1). Two wall-mounted cameras gave a landscape view of each side of the table (usually two people in shot). The third camera framed the projector screen and the white-board area. As for audio, the video recording conditions were unconstrained with phenomena such as object occlusions and changes in illumination.

These meeting recordings involve only audio and video, but the communicative process is spread between several modalities including speech, prosody, gestures, handwriting, and facial and body expressions. Further streams of data could be captured easily: for example, handwriting could be recorded through whiteboard capturing devices, graphic tablets, or digital pen/paper. Unfortunately this is not the case for modalities such

[2] Even sharing the same corpus and the same task, differences in the feature set, the data set subdivision and the evaluation methodology, make a direct comparisons between the present paper and [17]–[21] infeasible. A first attempt to overcome this situation, by comparing the performance of our DBN multistream model on three different feature setups (IDIAP, Munich and Edinburgh feature sets), can be found in a recent joint work [22].

[3] This corpus is publicly available from http://mmm.idiap.ch/

as gestures or facial expression, for which the use of specialized recording devices is impractical and invasive. When specialized recordings are not available, it is possible to extract multiple modalities from single streams. For example, speech could be separated from noise and other sound sources using microphone array beamforming, physical motion could be measured using image processing techniques, and may be further integrated into a gesture recogniser. Note that this is a simplified view of the problem, because a single modality corresponds to multiple different streams: for example, speech is manifested not only as a sound, but also as a lip motion. The situation is further complicated if we consider the correlations that exist between different modalities, such as speech and gestures [26].

IV. Feature Extraction

Four feature families were employed: a basic set of prosodic features, features relating to speaker turn dynamics, a lexical based monologue/dialogue discriminator, and a rough estimate of the motion of participants' head and hands.

A. Prosodic Features

A feature vector of three independent features was extracted for each participant, using audio recordings provided by individual lapel microphones. This vector consists of three entities: a smoothed estimate of fundamental frequency (F0), an estimate of the syllabic rate of speech, and energy.

The smoothed F0 is estimated in two steps: an initial F0 contour estimate using the ESPS pitch tracking algorithm[4], followed by a chain of three filters, inspired by Sonmez *et al.* [27], that denoise the initial estimate of F0. A histogram filter removes incorrect estimates arising from other undesired sound sources, followed by a median filter to smooth the F0 contour by removing spurious peaks, and a linear interpolation filter that provides a piecewise-continuous smoothed output.

The syllabic speaking rate was estimated from the acoustic signal using the algorithm *mrate*[28], which integrates the output of multiple rate of speech estimators.

The logarithm of root-mean-square energy $E_j(t)$ was evaluated for each lapel microphone channel j. $E_j(t)$, then normalized as follows [29]:

$$E_{\text{norm},j}(t) = E_j(t) - E_{\min,j} - \frac{1}{M}\sum_{k=1}^{M} E_k(t).$$

The minimum log-energy $E_{\min,j}$ can be interpreted as an estimate of the noise floor level recorded by channel j. Therefore it needs to be subtracted in order to compensate for different channel gains. The last term represents the mean log-energy averaged across all $M = 4$ channels. We are primarily interested in sounds (speech) that occur only in proximity of the channel. Considering one channel k at a time, those sounds should be considerably above the background noise (multichannel averaged energy).

In order to improve the quality of F0 and rate of speech, discretized versions of the speaker activities estimated using microphone array processing techniques (Section IV-B) were used to mask inactive lapel microphone channels. Unfortunately prosodic features could not be extracted when participants are presenting a talk or standing at the whiteboard, since the use of wired lapel microphones is feasible only when participants are close to the table. Therefore the prosodic feature set is partially incomplete and also affected by estimation errors.

Both F0 and rate of speech were normalized across the entire meeting, in order to have comparable features for different speakers. The resulting prosodic feature set thus captures variations in speaking style, highlighting specific aspects of the speech modality.

B. Speaker Turn Features

Face-to-face meetings display a complex turn-taking structure. The dynamics of this process can be extremely useful to distinguish between different meeting phases. For example, during dialogues speakers tend to alternate frequently, speaking for shorter periods.

To investigate the turn-taking process, it is necessary to detect speech activity for each participant in the meeting. This is difficult using the lapel microphone signals for two reasons. Firstly, since they are wired microphones, meeting participants only wear the lapel microphones while seated, which makes them impossible to use when someone is presenting a talk or standing at the whiteboard. Secondly the lapel microphones are omnidirectional and it is difficult to distinguish whether a signal is the speech of the participant wearing the microphone, or crosstalk from another speaker [29], [30]. Instead, we used microphone array recordings to detect speaker activity.

A microphone array can be regarded as a steerable directional microphone, but, compared with an orientable microphone, there are no moving parts. The steering direction can be imposed at any time during or after the recording session using a beamforming process. It is therefore possible to steer the virtual microphone in any direction, evaluating sound activity at a specific spatial location. There are only six spatial regions in which participants spent most of their time: the four seating regions that are individually associated with participants, the white-board and a presentation space near the projection screen. We detected continuous sound activities $L_i(t)$ in each of these six regions i, which were used as a basis for features to describe the turn-taking process. Each $L_i(t)$ is directly proportional to the probability of observing an active sound source (a meeting participant speaking or generating noise) in the spatial region i at time t, and it is zero when no activity is detected.

We constructed a 216-element feature vector to describe the turn-taking process at each time. The vector S consists of all 6^3 possible products of the six sound activity locations $L(t)$ during a time window of three frames [19]

$$S_{ijk}(t) = L_i(t) \cdot L_j(t-1) \cdot L_k(t-2) \quad \forall i,j,k \in [1,6]$$

where each vector $S_{ijk}(t)$ highlights the turn taking interaction pattern around the time t. Considering, for simplicity, a smaller turn taking matrix evaluated only on two frames

$$S_{ij}(t) = L_i(t) \cdot L_j(t-1) \quad \forall i,j \in [1,6]$$

[4] Available from http://www.speech.kth.se/snack/

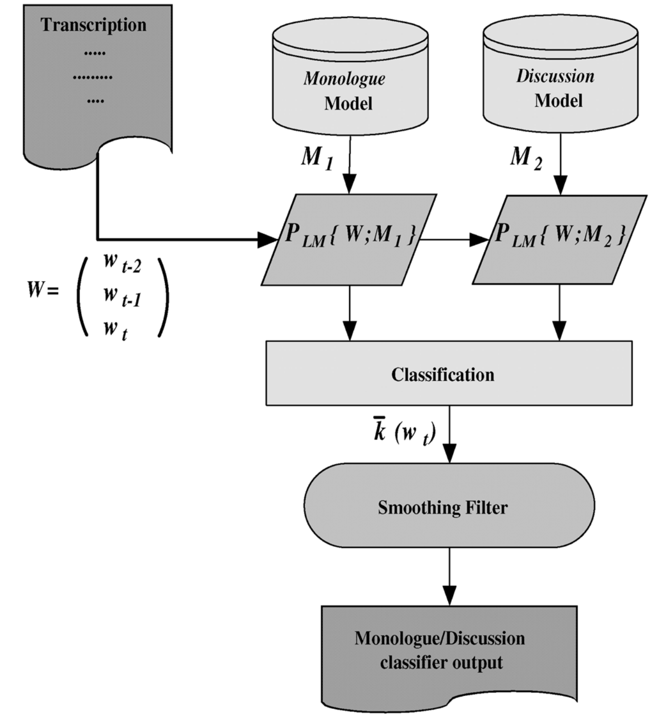

Fig. 2. Overview of the "monologue/discussion" classifier.

the diagonal elements $S_{ii}(t)$ highlight whether a speaker L_i active at time $t-1$, is still speaking at time t. The terms above the diagonal $(S_{ij}, i < j)$ are greater than zero when it is likely that L_i is speaking after L_j. Similarly $S_{ij} > 0, i > j$ implies that L_j at time $t-1$ and L_i at time t are both active. When all: S_{ii},S_{jj},S_{ij}, and S_{ji} are greater than zero, it is likely that a discussion (turn-taking alternation) between L_i and L_j is taking place. A similar discussion applies to $S_{ijk}(t)$.

Dimension reduction of S_{ijk} using principal component analysis was not effective, with reductions below 200 dimensions resulting in a degradation in performance. Thus, we used the unreduced 216-element feature vector in our experiments.

C. Lexical Features

Monologues and dialogues are characterized by different speaking styles and different language models. In particular we hypothesize that the distribution over words is different for transcripts from these two meeting phases. Using a transcript for each speaker we constructed trigram language models for each communicative context that we wish to recognize. In this work, we estimated language models for monologue and discussions only, but the idea could be extended to more elaborate domains.

The approach is illustrated in Fig. 2. Trigram language models correspond to monologues (M_1) and discussions (M_2). Those multinomial distributions over words are estimated using transcriptions from the training data set, and then used to partition unseen word sequences from the test set. Note that the language models are estimated employing all the transcribed words, irrespectively of the function they serve in the discourse. Each word w_t (together with its context w_{t-1}, w_{t-2}, if available) contained in the transcription under test is compared with both the models M_1 and $M_2 (K = 1, 2)$ and assigned to the class with the highest probability $P_{\mathrm{LM}}(w_t \,|\, w_{t-1}, w_{t-2}; M_k)$

$$\tilde{k}(w_t) = \arg\max_{k \in K} \{P_{LM}(w_t \,|\, w_{t-1}, w_{t-2}; M_k)\}$$

where $\tilde{k}(w_t)$ is the output of the classifier.

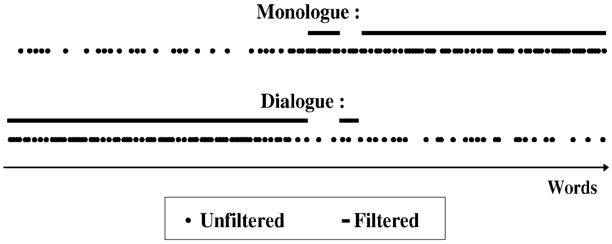

Fig. 3. Filtering of $\tilde{k}(w_t)$.

The resultant sequence of output symbols is noisy, with $\tilde{k}(w_t)$ constantly switching between the two states (small dots of Fig. 3). However, if we consider the symbol density, the output is much more stable (lines of Fig. 3). Therefore, we smooth the output by evaluating the relative frequency of $\tilde{k}(w_t)$ over a sliding window of 24 words. This window length has been arbitrarily chosen, but it seems not to be critical because values between 20 and 30 are equally acceptable. This lexically based approach is able to classify unseen word sequences as monologues or discussions with a percentage of correctly classified words of about 93%. [5] Removing the smoothing step and considering the noisy sequence $\tilde{k}(w_t)$ the classification accuracy falls to 78%. A lower bound on the class accuracy of 48% is obtainable by drawing the symbols by chance, according to the prior distribution.

D. Video Features

Meetings provide a well-defined and highly constrained environment for video and image processing. Participants spend most of the time in a few spatial locations—they move location rarely and there are relatively few physical actions. In the case of the M4 corpus, cameras are fixed, most furniture does not move and lighting conditions are partially constrained. However, participants are free to perform any action or gesture and do whatever they like. Therefore object occlusions are relatively frequent, and nothing has been done to facilitate object tracking (ie there is no "blue screen" or preassigned colors for clothing or furniture). Note that exposure settings are different for each camera. In particular, this is a critical issue for the camera oriented on the bright projection screen and dark white-board area (Fig. 1).

Although the recordings were made using high-quality equipment and good video resolutions (full frame PAL), regions of interest represent only a small fraction of the entire scene, providing a relatively low resolution. This resolution is sufficient for tasks such as tracking the head, hands, and other objects of a similar size. Close-up video recording will be required to address problems such as lip feature extraction for audio-video

[5] Average recognition using leave-one-out cross-validation strategy on 30 manually transcribed short meetings.

speech recognition or eye-gaze tracking for conversational attention prediction [31].

We assume visual information about the participants is correlated with meeting phases. For example, a speaker who is highly involved in the conversation will tend to gesticulate, and the use of a white-board involves a complex sequence of physical actions, such as standing up, walking and writing. Under that assumption we are interested in extracting a set of region based low level visual activities [16], that could improve the recognition of highly visual actions such as note taking and presentations.

Our efforts are concentrated on the two cameras oriented towards the meeting table. Each of those captures a scene with two speakers. As mentioned above, meeting participants spend most of the time sitting, and therefore only the upper body part is visible through those cameras (Fig. 1). In each scene, we analyze four areas: the head and hand regions for each of the two participants in shot. Instead of recognizing and tracking head and hand blobs [20] we have chosen a faster and more flexible approach that does not require an appearance model or any form of (re-)initialization.

Our system relies on an optical flow-based algorithm, which is used to track a fixed number ($n = 100$) of feature points. We have adopted an enhanced version of the "Kanade Lucas Tomasi" (KLT) feature tracker outlined in [32]. In particular the condition used to select the tracking feature set has been revised and extended. Given an image $F(x, y, t)$ and a region Ω, KLT relies on solving a linear system associated with the following matrix A:

$$A = \nabla F(\Omega, t)^T \cdot \nabla F(\Omega, t)$$

if λ_1 and λ_2 are the two eigenvalues of A, the system will be well conditioned if the smallest eigenvalue $\lambda_{\min}$ is large enough

$$\lambda_{\min} = \min(\lambda_1, \lambda_2) \gg 0. \tag{1}$$

Adopting this condition, Tomasi [32] states that "good features are the ones that can be tracked well", proposing therefore to track feature regions with a particularly rich texture.

Being interested in tracking skin-like regions, we have extended the feature quality metric (1) proposed by Tomasi with an additional condition over the candidate region's color

$$P(\Omega \,|\, \mathrm{Skin}) \geq P_{th} = 0.5. \tag{2}$$

It is thus feasible to evaluate the chromatic distribution of skin blobs [33], and to use that distribution to estimate the probability of a given region Ω to be skin. The chromatic space can be represented through different bases: here, we adopted the luminance and chrominance space $\{Y, Cr, Cb\}$. Skin-like colors are well clustered under the $\{Cr, Cb\}$ subspace, and a 3 component GMM was trained using unseen skin blobs. The resulting skin color model provided an easy way to estimate the skin probability (2) for each candidate region Ω. Therefore a good feature is now one that can be tracked well (1) and has a high probability to be part of a skin area (2).

Our approach to the video feature extraction process is depicted in Fig. 4. Each video stream is processed on a frame-to-frame basis, the skin probability is estimated and used to select and track 100 features. Those features are processed off-line. Feature trajectories that are too long and have a limited amount of motion are automatically removed. The next step consists of partitioning the trajectory space into four regions (head and hand areas for the two participants). Four Gaussian distributions (one centroid for each region) were estimated using the entire sequence. This rough global estimation was refined on a frame basis, by using a k-means clustering. If a trajectory T of length n is assigned to a set $K(i)$,1 $= 1, \ldots, n$ of different regions, and $\tilde{K}$ is the most frequent assignment, then the whole trajectory T is classified as part of region $\tilde{K}$.

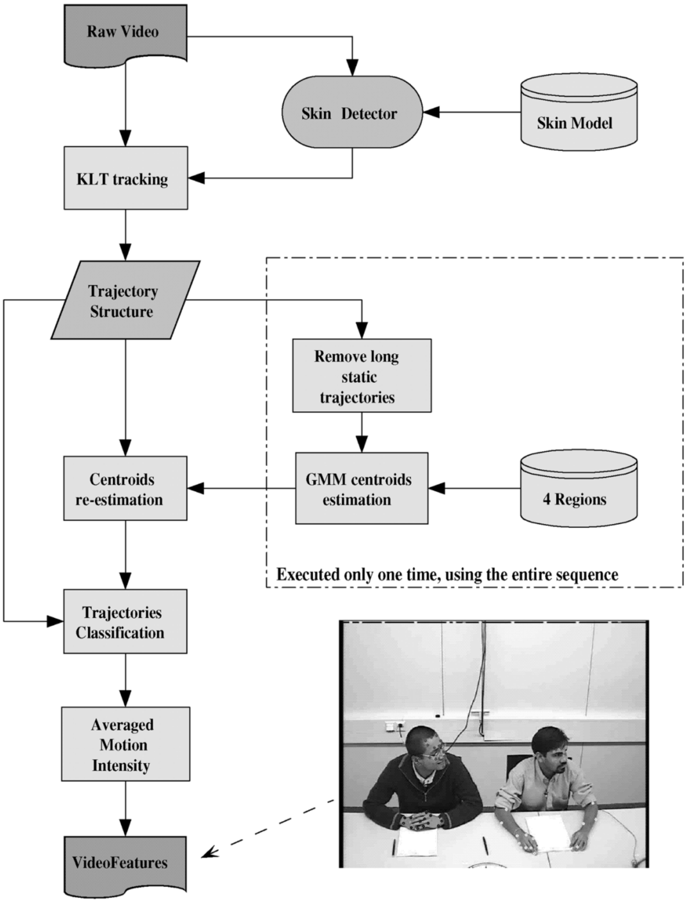

Fig. 4. Overview of the "video features" extraction process.

For each frame, and for each region, two video features were extracted: the average feature motion intensity, and the approximate motion direction. Thus, from each raw video signal, an eight-element feature vector is extracted each frame. The feature vectors from the two cameras are combined, resulting in a 16-element global video feature vector. Owing to the recording conditions of the third camera (projector screen and whiteboard area), motion vectors extracted from this source are less reliable, and were excluded from our experimental setup.

This approach exploits few assumptions about the scene structure without pretending to precisely identify head or hand blobs. Therefore, object occlusions are only partially handled. However, recovering from an unexpected event is fast and completely automatic, there is no need for manual initialization, and this technique translates well between domains. The system is able to operate in the presence of complex colored backgrounds without any performance degradation, and is able to cope with gradual illumination changes.

V. Dynamic Bayesian Networks

Bayesian networks (BNs) are directed probabilistic graphical models, in which nodes represent random variables, and directed arcs between nodes represent conditional dependences among variables. The conditional (in-)dependence relationships encoded in the graph provide a compact factorized representation for the joint probability of all the nodes. This representation may be exploited in order to reduce the computational effort required for probabilistic inference. Inference is the key step both for parameter estimation (model training) and model decoding.

For signal-processing applications, we are interested in modelling systems of time-dependent variables. DBNs are an extension of BNs to process time series. In a DBN, a local BN is instantiated for each time slice, and the complete network is formed by adding interconnections between the local networks. Each local BN describes the relations between different random variables within a time frame. The temporal dynamics are represented using additional arcs between nodes in different time frames. Hence, a DBN is a set of static BNs interconnected by some additional causal links across slices, which explicitly represent the time flow. DBNs (and graphical models in general) have several advantages over basic HMMs:

- increased flexibility in the state-space factorisation and structuring;
- capability to integrate some problem specific knowledge into the model, and therefore ability to develop potentially more discriminative models;
- improved and more parsimonious use of the parameter space;
- unified graphical-mathematical formalism.

It is possible to express simpler models such as HMMs and Kalman filters, or richer models including coupled HMMs, factorial HMMs, hierarchical HMMs, and semi-Markov models as DBNs [34]–[36]. Using this formalism common sets of tools have been developed to perform standard activities such as training model parameters, state space decoding, and sampling. In this work, we have employed the Graphical Model ToolKit (GMTK) [37].

VI. Multistream Meeting Models

This work addresses the automatic recognition of meeting group actions, in which each group action is the result of the interaction of multiple subjects over multiple communicative modalities. Several approaches to group action recognition have been proposed (Section II-B). A straightforward approach to the problem would consist of early integration of feature streams extracted from different subjects and modalities, followed by a simple HMM-based infrastructure, and such an approach has formed the baseline system used in our experiments (Section VII-A). This solution is simplistic, since two main issues are disregarded: the explicit modeling of the interaction between multiple feature families, allowing an independent tuning and a better control over each feature stream; the necessity of relaxed temporal synchronization constraints among multiple modalities and participants. Therefore, coupled HMMs, layered HMMs, and other multistream approaches are potentially better suited to this task. In particular, multistream models are highly flexible, intuitive, and lend themselves to further improvement.

Multistream approaches to group action recognition may use participant-based integration, or modality-based integration. In participant-based approaches, features from different modalities (individually extracted from each participant) are grouped together and modeled as a single stream. Thus, each stream corresponds to a participant, and the whole group behavior is inferred from the integration of single participant behaviors (substates). On the other hand, the modality-based approach focuses on modeling each communicative modality individually, grouping together behaviors associated with different participants.

Our multistream approach is based on processing different modalities independently. We assume that the group acts as a single subject and that "meeting actions" are related in the first instance to the entire group behavior. Note that features such as "speaker turns" (Section IV-B) are inherently related to the whole group rather than to individual participants. Moreover we preferred this strategy because it seems to provide better results when compared with the participant based one [18].

A third, hybrid, approach obtained by modeling each participant-based unimodal feature stream independently, could be investigated also. Unfortunately, although this approach has promising results, it requires a much larger state-space (and hence considerable compute resources) for realistic applications.

A. Multistream DBN Model

The most attractive feature of the DBN framework is its extreme flexibility in the factorization and structure the state-space. We assume that meeting actions can be interpreted as sequences of atomic units (*subactions*), much as sentences are subdivided into sequences of words. Thus, we propose a model which is structured as a hierarchy of three layers: complete meeting actions at the top, subactions in the middle and the observed feature streams at the bottom. Thus, low-level features are mapped into atomic subactions, which are themselves the building blocks of complete meeting actions.

Each feature family represents a single modality (even if extracted from multiple media). If we assume that multiple modalities are independent at a subaction level and interact only at the highest level, then the feature streams are integrated (avoiding artificially introduced forms of stream weighting) at the top level during the global meeting action recognition. Thus, this may be regarded as a multistream approach, since feature-streams are processed independently using their own subactions.

These subactions are obtained in an unsupervised way as the result of a training process. Each subaction is expected to represent a cluster of feature vectors which is associated with a particular meeting behavior and is dominated by a common underlying dynamic. There is no clear and immediate interpretation of subactions, and supervised approaches to obtain subactions could be extremely difficult and expensive.

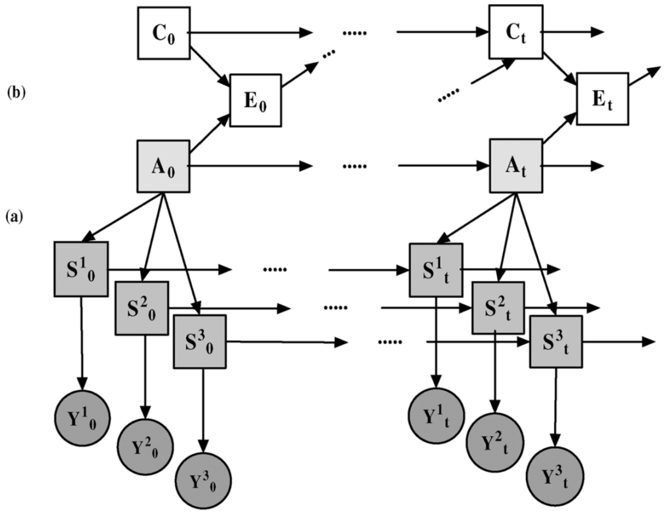

Fig. 5. Multistream DBN model (a) enhanced with a "Counter Structure" (b); square nodes represent discrete hidden variables and circles must be regarded as continuous observations.

The state-space factorization property may be exploited via both a hierarchical decomposition and a feature based subdivision. Consider a DBN, with a local BN at each time t. The resulting model [Fig. 5(a)] appears as a tree-shaped structure in which the observable features Y^F,$F=[1,N]$ are individually connected to their subaction variables S^F which are further connected to the action node A. The hidden variables S^F and A are each characterized by their own dynamics, in which each node is linked with its predecessor, forming a Markov chain.

The hierarchical relationship between A and S^F results in a structure that resembles a hierarchical HMM (HHMM) [38]. However this model is quite different, since HHMMs are characterized by a structured hierarchy of multiple Markov chains, and by a re-synchronisation mechanism which enables state transitions in higher chains only when lower HMMs have reached a "terminal state". Our model is free from this constraint, since actions A are free to change independently of the state of S^F. Similarly, the multistream approach to audio-video speech recognition [12] also relies on some re-synchronization points, referred to as anchor points. It is possible to interpret this model as a Dynamical Systems Tree [15] with three leaves, a single level of "aggregating nodes", and without the switching linear dynamical systems to couple the leaf nodes with subaction chains.

The lowest level of the model contains N continuous observable feature vectors (nodes Y^F), each of which represents a single modality that has been extracted from raw audio/video recordings. Each feature stream Y^F is then mapped into discrete substates S^F through a Gaussian mixture model with M_F components

$$P\left(Y_t^F = y \mid S_t^F = i\right) = \sum_{m=1}^{M_F} C(F,m,i) \cdot N(y;\mu_{F,m,i},\Sigma_{F,m,i}) \qquad (3)$$

where $N(y;\mu_{F,m,i},\Sigma_{F,m,i})$ is a Gaussian density with mean $\mu_{F,m,i}$ and covariance $\Sigma_{F,m,i}$, evaluated at y, and $C(F,m,i)$ is the conditional prior weight of each mixture component m associated with stream F.

Each substate node S^F, $F=[1,N]$ is part of an independent Markov chain, and each subaction node S^F is a child of the global action node A. Therefore, substate transition matrices $R_k^F(i,j)$ and an initial state distributions $\tilde{\pi}_k^F(j)$, associated with S^F, are functions of the action variable state $A_t=k$

$$R_k^F(i,j) = P\left(S_t^F = j \mid S_{t-1}^F = i, A_t = k\right) \qquad (4)$$

$$\tilde{\pi}_k^F(j) = P\left(S_1^F = j \mid A_1 = k\right) \qquad (5)$$

where $\tilde{\pi}_k^F(j)$ is the initial subaction distribution for the stream F, given an initial action $A_1=k$; and $R_k^F(i,j)$ represents the transition probability from subaction i to substate j, given that the global meeting action variable $(A_t=k)$ is in state k.

The sequence of action nodes A forms a Markov chain with multiple subaction nodes S^F as children. Therefore A can be regarded as an HMM generating N hidden discrete subaction sequences S^1, S^2, $S^3, \ldots S^N$ through $R_k^1(i,j)$, $R_k^2(i,j)$, $R_k^3(i,j), \ldots R_k^N(i,j)$, respectively. In a further analysis, $R_k^N(i,j)$ is then responsible for the modelling of the joint dynamics of N multiple streams. $P(A_1=i)=\pi(i)$ is the initial state probability vector associated with A, and $P(A_t=j \mid A_{t-1}=i)=Q(i,j)$ is the transition probability matrix. Note that the Markov chain A acts as an integration point, collating together the whole information carried by each subaction stream (representing a single feature family). Pushing the integration point to the highest level of the model in this way is referred to as "late integration". Finally, the joint distribution for a sequence of T temporal slices, considering the entire multistream model Fig. 5(a), is given by

$$\begin{aligned} &P\left(A_{1:T}, S_{1:T}^1, \ldots, S_{1:T}^N, Y_{1:T}^1, \ldots, Y_{1:T}^N\right) \\ &= P(A_1) \cdot \prod_{F=1}^{N} \left\{ P\left(S_1^F \mid A_1\right) \cdot P\left(Y_1^F \mid S_1^F\right)\right\} \\ &\quad \cdot \prod_{t=2}^{T} \left\{ P(A_t \mid A_{t-1}) \right. \\ &\quad \left. \cdot \prod_{F=1}^{N} \left\{ P\left(S_t^F \mid S_{t-1}^F, A_t\right) \cdot P\left(Y_t^F \mid S_t^F\right)\right\} \right\}. \end{aligned} \qquad (6)$$

Note that the cardinality of action nodes A is imposed by the size of the "action dictionary": $|A|=8$ in this work. The cardinalities of the subaction nodes S^F are model parameters; from some development experiments we discovered that all our feature families (except the lexical-based monologue/discussion discriminator) perform at their best when modeled with at least five subactions. Note also that the maximum allowable degree of asynchrony between the N parallel streams is directly proportional to the dimension of the product state space, $\prod_{F=1}^{N} |S^F|$.

B. Counter Structure

HMMs are characterized by a distribution in which the probability of remaining in a given state decreases as an inverse exponential [39]. This state duration distribution is not well-matched to the behavior of meeting action durations. This issue may be addressed in various ways, such as semi-Markov models, and state duplication to impose minimum duration constraints, as well as *ad hoc* solutions such as action transition penalties.

We preferred to improve the flexibility of state duration modelling, by enhancing the existing model with an additional "counter structure" Fig. 5(b). The duration of meeting actions is constrained by using a counter node C and an enabler node E. The sequence of counter nodes C forms a Markov chain, which attempts to model the expected number of recognized actions, whereby C is ideally incremented by a unit during each action transition. In this counter-structure enhanced model, action variables A are not only parents of subactions S^F, but also of the enabler nodes E. Therefore, A generates both N sequences of subactions S^F and a sequence of hidden enabler states E. Moreover, the binary enabler variables E, reach their active state 1 only in the presence of action transitions ($E_t = 1$ only if $A_t \neq A_{t-1}$ and therefore $C_t = C_{t-1} + 1$), thus providing an interface between action variables A and counter nodes C. The counter variable C can be incremented only if the enabler variable E was high $(E_{t-1} = 1)$ during the previous temporal slice $t - 1$

$$P(C_t = j \,|\, C_{t-1} = i, E_{t-1} = f) = \begin{cases} j = i + 1, & \text{if } f = 1 \\ j = i, & \text{if } f = 0 \end{cases} \tag{7}$$

where $P(C_t = j \,|\, C_{t-1} = i, E_{t-1} = f)$ represents the state transition probability for the counter variable C given global *counter structure* state during the previous frame $t - 1$. Any evolution of the enabler node E is conditioned on both the action variable A and on the counter variable C. If A is in state k and the counter C in state j, the probability to activate E is given by

$$P(E_t = f \,|\, C_t = j, A_t = k) = D_{j,k}(f) \tag{8}$$

where $D_{j,k}(f)$ represents the state transition probability associated with E. Suppose that the jth meeting action has been recognized at time t $(A_t = k)$, then the probability of encountering a new action (the $(j+1)$th) or equivalently to have E activated $(E_t = 0, E_{t+1} = 1)$ will be modelled by $D_{j,k}(f)$. Assuming that action transitions are not possible during the first time frame $t = 0$, the initial probability of E is equal to $P(E_1 = 0) = 1$ and for coherence $P(C_1 = 0) = 1$.

The complete joint distribution of the multistream model enhanced with a counter structure [Fig. 5(a) and (b) combined], computed for a sequence of T frames, is given by

$$\begin{aligned} &P\left(A_{1:T}, C_{1:T}, E_{1:T}, S^1_{1:T}, \ldots, S^N_{1:T}, Y^1_{1:T}, \ldots, Y^N_{1:T}\right) \\ &\quad = P(A_1) \cdot P(C_1) \cdot P(E_1) \cdot \prod_{F=1}^{N} \left\{ P\left(S_1^F \,|\, A_1\right) \right. \\ &\qquad \left. \cdot P\left(Y_1^F \,|\, S_1^F\right) \right\} \\ &\qquad \cdot \prod_{t=2}^{T} \left\{ P(A_t \,|\, A_{t-1}) \cdot P(C_t \,|\, C_{t-1}, E_{t-1}) \right. \\ &\qquad \cdot P(E_t \,|\, C_t, A_t) \\ &\qquad \left. \cdot \prod_{F=1}^{N} \left\{ P(S_t^F \,|\, S_{t-1}^F, A_t) \cdot P(Y_t^F \,|\, S_t^F) \right\} \right\}. \end{aligned} \tag{9}$$

Note that the use of a counter structure is not limited to the multistream model used here, but can be applied to any Markov chain.

VII. Experiments

All experiments were conducted on a subset of the publicly available meeting data corpus described in Section III. We employed a baseline HMM system and multistream DBN systems, using the feature families described in Section IV: 12 prosodic, 216 speaker turn, 1 lexical, and 16 visual features for total of 245 features. The lexical features (Section IV-C) require word-level transcriptions. However this data comprises natural speech from non-native speakers, recorded using lapel and far field microphones, which results in high automatic speech recognition (ASR) word error rates. Our experiments were therefore performed using the human transcriptions, and the reported results are for a semi-automatic system. ASR transcriptions of each speaker would be required for a fully automatic framework. Only 30 meetings were transcribed (about 150 min), which is a too small amount of data to provide separate training and test sets. We therefore performed our experiments using a leave-one-out cross-validation strategy, in which models were trained on 29 meetings and tested on the remaining one, the procedure being iterated 30 times.[6]

The task of meeting action recognition involves both segmentation and classification. Since the boundaries between meeting actions are not always precise, we have adopted an evaluation metric focused on the recognition of the correct sequence of actions and flexible about temporal boundaries [4], the action error rate (AER)

$$\text{AER} = \frac{\text{Substitutions} + \text{Deletions} + \text{Insertions}}{\text{Correct number of actions}} \cdot 100.$$

The AER is evaluated by summing the substitution, insertion, and deletion errors of each recognized sequence when aligned to its reference transcription. Note that the adopted meetings follow a predefined sequence of actions (Section III) which constitutes the ground truth for our experiments. AER is analogous to the word error-rate metric used in speech recognition, and like word error-rate, is usually more severe than the frame-based accuracy.

A. Baseline Systems

A baseline system to relate low-level features with high-level meeting actions was developed using an ergodic HMM. Six systems were developed, one trained on each of the four feature sets individually, one trained combining nonvisual features only, and a sixth using all four feature sets combined together. Since the four feature sets previously outlined were extracted in different contexts, they have different sampling rates. In order to share the same sampling frequency, all of them were down-sampled, to a common sampling rate of 2 Hz. The word-level-based time scale of lexical features was converted using the word-time boundaries provided by transcriptions. Although the feature families shared the same sampling frequency after this process, it is not

[6] Compared with the experimental setup in [19], here we used a different subset of the M4 meeting corpus, a more robust experimental methodology (cross-validation), and a smaller parameter space ($|S^F| = 5$ instead of 7).

TABLE I
COMPARISON BETWEEN MEETING ACTION RECOGNITION RATE (% CORRECT) AND (SUBSTITUTION, INSERTION, DELETION, AND OVERALL) ERROR RATES ACHIEVED USING FOUR FEATURE CONFIGURATIONS AND A SIMPLE HMM MODEL

Feature	Corr.	Sub.	Del.	Ins.	AER
Speaker turn features alone	65.4	16.7	17.9	20.5	55.1
Lexical features alone	58.3	23.7	17.9	7.1	48.7
Prosodic features alone	50.0	21.8	28.2	9.6	59.6
Turn, lex. and pros. features	**71.5**	**10.3**	**19.2**	**14.7**	**44.2**
Video features alone	48.1	21.8	30.1	7.1	59.0
All 4 feature families	**71.2**	**10.3**	**18.6**	**14.7**	**43.6**

the case that they show similar temporal behaviors: each feature set has its own time scale and level of asynchrony.

Tests on a development set (without the lexical information) indicated that an 11-state ergodic HMM was well suited to this data. Table I shows the action error rates for each feature set. It can be seen that speaker turns provide the highest percentage of correctly recognized actions, followed by lexical features and prosodic features. Lexical features are most useful for discriminating between discussion and monologue, and the video-related features help most to discriminate between highly visual actions (note taking, presentation, and presentation at the whiteboard). Note that monologue and discussions represent the 66% of the corpus, with the other actions comprising only 34%. All the results shown in Table I are thus affected by this action distribution: speaker turn and lexical feature results are enhanced and video features weakened. The integration of visual features (last line of Table I) into the baseline system composed by speaker turn, lexical and prosodic features (fourth line of Table I) resulted in a small improvement in the overall recognition rate.

B. Multistream Model

We compared experimentally the accuracy of the baseline HMM system with the multistream DBN model (Section VI-A), and the multistream model enhanced with a counter structure (Section VI-B). The results of these experiments are reported in Table II. The multistream models were trained using three independent feature streams. Note that prosodic and lexical features were early integrated into a single 13-dimensional feature vector Y^3, and that the state-space has been limited to only five subactions per stream ($|S^1| = |S^2| = |S^3| = 5$). The multistream model shows a decisive improvement over this baseline system: the recognition rate (% correct) is increased by 17.9%, and together with a significant drop in the number of insertions, this results in a substantially reduced AER of 13.5%. Further small improvements were provided by the addition of a counter structure. This halved the number of insertions (at the cost of a small increase in the number of deletions), indicating an increased state duration, resulting in a further improvement in AER (12.2%), the best results achieved on this task.

Model training is about three times slower than real-time on a 3-GHz P4 processor, and feature decoding/recognition is two times faster than real-time. However, the memory requirements of Viterbi decoding were large, with about 1.5 Gb required for decoding a system that used five substates per stream.

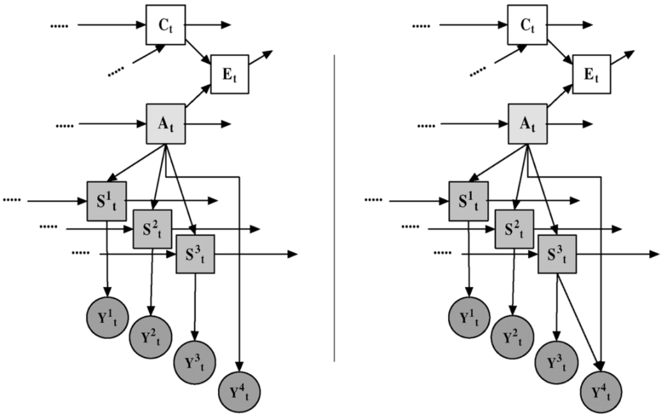

Fig. 6. Model A (left), model B (right); both the models are enhanced with a "Counter Structure".

TABLE II
AERs (%) FOR: A SIMPLE HMM, A 3-STREAMS DBN MODEL, AND A 3-STREAMS COUNTER-ENHANCED VERSION; LOWER AERs INDICATE BETTER PERFORMANCES

Model	Corr.	Sub.	Del.	Ins.	AER
HMM	71.2	10.3	18.6	14.7	43.6
multistream	89.1	3.2	7.7	2.6	13.5
multistream + counter	89.1	2.6	8.3	1.3	12.2

C. Extended Multistream Models

The binary lexical features are able to discriminate between monologue and discussion with a good accuracy. Since these two categories are a subset of the action dictionary, there is no reason why they need to be integrated with prosodic features and then modeled by an intermediate Markov chain (subaction S^F). Hence, we have investigated an extended model (model A in Fig. 6) in which observable lexical features Y^4 are direct parents of the top level action chain (nodes A). The whole joint distribution after unrolling the model for T temporal frames is given by a slightly modified version of (6) or (9). The number F of independent streams was set to $F = 3$, and $P(A_t \mid A_{t-1})$ was replaced by $P(A_t \mid A_{t-1}, Y^4)$. Note that speaker turns, prosodic features, and motion data are modeled as usual using three independent subaction Markov chains F with the following cardinalities: $|S^1| = 5$, $|S^2| = 5$, and $|S^3| = 5$. As can be seen in Table III, the AERs obtained using this model are poorer than the standard multistream approach discussed below, supporting the need of a dedicated intermediate level (subaction nodes S^F) for lexical feature processing.

In order to further address this issue, we investigated a second hybrid model (model B on the right side of Fig. 6) based on the multistream approach (Fig. 5). The lexical feature data stream was modeled in conjunction with prosodic data using a substate chain S^3, that was directly related to action nodes A. Similar to model A, the joint probability distribution could be obtained from (9) by replacing $P(A_t \mid A_{t-1})$ with $P(A_t \mid A_{t-1}, Y^4)$. Note that Y^3 is the prosodic feature vector (as for the previous experiment) and Y^4 contains only the binary lexical feature. The experimental results achieved with this model are reported in the last two rows of the Table III: the extended model B has

TABLE III
AERs (%) FOR TWO EXTENDED VERSIONS OF THE MULTISTREAM MODEL

Model	Corr.	Sub.	Del.	Ins.	AER
model A	87.2	4.5	8.3	4.5	17.3
model A + counter	87.8	2.6	9.6	2.6	14.7
model B	89.7	2.6	7.7	1.9	12.2
model B + counter	90.4	2.6	7.1	2.6	12.2

a lower AER compared with A, but the counter structure does not seems to improve the AER for model B.

Unfortunately, the meeting corpus adopted for these experiments is very small, and it is not possible to discriminate between the standard multistream model and model B. These two models offer similar accuracy, and the addition of a direct dependency of the highest level Markov chain on a low-level feature stream, did not compromise the performance. In order to validate and improve the accuracy of these results, we are looking forward to repeating all these experiments on a much larger and more realistic corpus [40]. Furthermore, it is of interest to investigate multistream models enhanced with a more complicated dependency structure between the subactions and the feature vectors.

VIII. SUMMARY AND CONCLUSION

In this work, we have addressed the problem of automatically segmenting a meeting into a sequence of group meeting actions taken from a dictionary of events such as monologue, discussion, and presentation. We performed our experiments using a publicly available corpus of meetings recorded using multiple cameras and microphones. This corpus has some limitations, including the short duration of each meeting (5 min per meeting, on average), the fact that only 30 meetings (150 min) were fully annotated, and the somewhat artificial content of the meeting agenda and topics. Despite these limitations, the corpus does feature natural and spontaneous interactions between participants, and provides a good basis for investigations in multimodal processing and event recognition in multiparty meetings.

The multiperspective audio/video recordings were processed by extracting relevant multimodal features, followed by statistical modeling. Four feature families were extracted from these recordings, representing speaker turn dynamics, prosodic and lexical information, and participant motion (head/hand/body movements). In order to relate these low-level feature streams with high-level meeting actions, a DBN multistream model was adopted. Using this multistream framework, it is possible to process each feature stream independently at a lower level of the model, and to collect together partial results at the upper stage of the model, thus offering a hierarchical approach to the integration of multiple feature streams.

The capability to incorporate some knowledge of the problem into the model structure is one of the principal features of the DBN framework, resulting in a more parsimonious model compared with simple HMMs. Moreover, the use of a multistream approach shows some advantages over merging all the feature families into a single feature vector (early integration).

- The integration point in which knowledge from different feature streams is collected together, may be delayed to a later stage of the processing (*late integration*).
- The independent feature processing increases the flexibility in modeling the interdependences between different modalities, allowing the model to encompass complex statistical dependences, lack of synchronism, and multiple time scales.

These advantages have resulted in a significant increase in accuracy when the DBN multistream models are used in place of an HMM for the meeting action recognition task, resulting in an action error rate of 12.2%.

The systems that we developed could be improved in various ways, such as through the use of a reliable cross-talk detector, reliable ASR to obtain the word transcriptions, and improvements to the visual features by using an appearance model for the head area and by estimation of head poses (useful to predict speaker addressing). Custom visual features specialized on the projector/whiteboard area are another potentially valuable extension for the present framework.

Many aspects of the multistream DBN framework outlined above (e.g., multiple time scale models) have not been exploited in the current work and there is much scope for exploration and improvement. Furthermore it is our intention to validate this framework, as soon as possible, on a more realist multimodal meeting corpus [40] that we are currently collecting, that is characterized by real, fully unconstrained meetings.

REFERENCES

[1] J. E. McGrath, “Time, interaction and performance (TIP) - a theory of groups,” *Small Group Res.*, vol. 22, no. 2, pp. 116–129, May 1991.

[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 *Proc. IEEE ICASSP*, May 2001.

[3] R. Kazman, R. Al Halimi, W. Hunt, and M. Mantei, “Four paradigms for indexing video conferences,” *IEEE Multimedia*, vol. 3, no. 1, pp. 63–73, Spring 1996.

[4] I. McCowan, S. Bengio, D. Gatica-Perez, G. Lathoud, F. Monay, D. Moore, P. Wellner, and H. Bourlard, “Modelling human interaction in meetings,” in *Proc. IEEE ICASSP*, 2003.

[5] T. Schultz, A. Waibel, M. Bett, F. Metze, Y. Pan, K. Ries, T. Schaaf, H. Soltau, M. Westphal, H. Yu, and K. Zechner, “The ISL meeting room system,” in *Proc. Workshop Hands-Free Speech Communication*, Apr. 2001.

[6] D. Lee, B. Erol, and J. Graham, “Portable meeting recorder,” *ACM Multimedia*, Dec. 2002.

[7] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, “The ICSI meeting corpus,” in *Proc. IEEE ICASSP*, Apr. 2003.

[8] N. Oliver and E. Horvitz, “Selective perception policies for guiding sensing and computation in multimodal systems: A comparative analysis,” *ACM ICMI*, Nov. 2003.

[9] A. Garg, V. Pavlovic, and J. M. Rehg, “Boosted learning in dynamic Bayesian networks for multimodal speaker detection,” *Proc. IEEE*, vol. 91, no. 9, pp. 1355–1369, Sep. 2003.

[10] C. G. Snoek, M. Worring, and A. G. Hauptmann, “Detection of TV news monologues by style analysis,” in *Proc. IEEE ICME*, Taipei, Taiwan, R.O.C., Jun. 2004.

[11] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, “Recent advances in the automatic recognition of audiovisual speech,” *Proc. IEEE*, vol. 91, no. 9, pp. 1306–1326, Sep. 2003.

[12] S. Dupont and J. Luettin, “Audio-visual speech modeling for continuous speech recognition,” *Proc. IEEE Trans. Multimedia*, vol. 2, no. 3, Sep. 2000.

[13] Y. Zhang, Q. Diao, S. Huang, and W. Hu, “DBN based multi-stream models for speech,” in *Proc. IEEE ICASSP*, 2003.

[14] A. Hakeem and M. Shah, "Ontology and taxonomy collaborated framework for meeting classification," in *Proc. Int. Conf. Pattern Recognition*, Aug. 2004, pp. 263–268.

[15] A. Howard and T. Jebara, "Dynamical systems trees," *Uncertainty in Artific. Intell.*, Jul. 2004.

[16] S. Basu, T. Choudhury, B. Clarkson, and A. Pentland, "Towards measuring human interactions in conversational settings," in *Proc. IEEE Workshop on Cues in Communication at CVPR*, Dec. 2001.

[17] S. Reiter and G. Rigoll, "Segmentation and classification of meeting events using multiple classifier fusion and dynamic programming," in *Proc. IEEE ICPR*, Aug. 2004.

[18] I. McCowan, D. Gatica-Perez, S. Bengio, and G. Lathoud, "Automatic analysis of multimodal group actions in meetings," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 305–317, Mar. 2005.

[19] A. Dielmann and S. Renals, "Dynamic Bayesian networks for meeting structuring," in *Proc. IEEE ICASSP*, May 2004, pp. 629–632.

[20] D. Zhang, D. Gatica-Perez, S. Bengio, I. McCowan, and G. Lathoud, "Modeling individual and group actions in meetings: A two-layer HMM framework," in *Proc. IEEE CVPR, Workshop on Event Mining in Video*, Jul. 2004.

[21] A. Dielmann and S. Renals, "Multi-stream segmentation of meetings," in *Proc. IEEE Workshop on Multimedia Signal Processing*, Sep. 2004, pp. 167–170.

[22] M. Al-Hames, A. Dielmann, D. Gatica-Perez, S. Reiter, S. Renals, G. Rigoll, and D. Zhang, "Multimodal integration for meeting group action segmentation and recognition," in *Proc. Multimodal Interaction and Related Machine Learning Algorithms Workshop (MLMI-05)*, 2006, pp. 52–63.

[23] M. Al-Hames and G. Rigoll, "A multi-modal mixed-state dynamic bayesian network for robust meeting event recognition from disturbed data," in *Proc. IEEE ICME*, Amsterdam, The Netherlands, Jul. 2005, pp. 45–48.

[24] S. Bengio, "An asynchronous Hidden Markov Model for audio-visual speech recognition," in *Advances in Neural Information Processing Systems, NIPS 15*. Cambridge, MA: MIT Press, 2003.

[25] D. Zhang, D. Gatica-Perez, S. Bengio, I. McCowan, and G. Lathoud, "Multimodal group action clustering in meetings," in *Proc. ACM Multimedia, Workshop on Video Surveillance and Sensor Networks*, Oct. 2004.

[26] D. McNeill and S. D. Duncan*, Language and Gesture*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[27] K. Sonmez, E. Shriberg, L. Heck, and M. Weintraub, "Modelling dynamic prosodic variation for speaker verification," in *Proc. ICSLP*, 1998, vol. 7, no. 920, pp. 3189–3192.

[28] N. Morgan and E. Fosler-Lussier, "Combining multiple estimators of speaking rate," in *Proc. IEEE ICASSP*, 1998, pp. 729–732.

[29] T. Pfau, D. Ellis, and A. Stolcke, "Multispeaker speech activity detection for the ICSI meeting recorder," in *Proc. IEEE ASRU Workshop*, Dec. 2001.

[30] S. J. Wrigley, G. J. Brown, V. Wan, and S. Renals, "Speech and crosstalk detection in multi-channel audio," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 84–91, Jan. 2005.

[31] R. Vertegaal, R. Slagter, G. van der Veer, and A. Nijholt, "Eye gaze patterns in conversations: There is more to conversational agents than meets the eyes," in *Proc. ACM SIGCHI Conf. Human Factors in Computing Systems*, 2001, pp. 301–308.

[32] J. Shi and C. Tomasi, "Good feature to track," in *Proc. IEEE CVPR*, Seattle, WA, 1994.

[33] J. Yang, W. Lu, and A. Waibel, "Skin-color modeling and adaptation," in *Proc. Asian Conf. Computer Vision*, 1998, vol. 2, pp. 687–694.

[34] P. Smyth, D. Heckerman, and M. I. Jordan, "Probabilistic independence networks for hidden Markov probability models," *Neural Comput.*, vol. 9, no. 2, pp. 227–269, 1997.

[35] K. P. Murphy, "Dynamic Bayesian Networks: Representation, Inference and Learning," Ph.D. dissertation, Comput. Sci. Div., Univ. California, Berkeley, Jul. 2002.

[36] J. Bilmes, "Graphical models and automatic speech recognition," in *Mathemat. Found. Speech and Lang. Process.*, 2003.

[37] J. Bilmes and G. Zweig, "The Graphical Model ToolKit: An open source software system for speech and time-series processing," in *Proc. IEEE ICASSP*, Jun. 2002.

[38] S. Fine, Y. Singer, and N. Tishby, "The hierarchical Hidden Markov Model: Analysis and applications," *Mach. Learn.*, vol. 32, no. 1, pp. 41–62, 1998.

[39] L. R. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[40] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The AMI meeting corpus: A pre-announcement," in *Proc. Multimodal Interaction and Related Machine Learning Algorithms Workshop (MLMI-05)*, 2006, pp. 28–39.

Alfred Dielmann received the Laurea degree in electronic engineering from the University of Cagliari, Cagliari, Italy, in 2002. He is currently pursuing the Ph.D. degree at the Centre for Speech Technology Research, School of Informatics, University of Edinburgh, Edinburgh, U.K.

From 2002–2003, he was a graduate Research Assistant at the Speech and Hearing Research Group, Computer Science Department, University of Sheffield, Sheffield, U.K. Since October 2003, he has been a Research Associate at the Centre for Speech Technology Research, School of Informatics, University of Edinburgh. His research interests concern multimodal signal processing and machine learning, in particular probabilistic graphical models for multiparty interaction modelling and natural language processing.

Steve Renals (M'91) received the B.Sc. degree in chemistry from the University of Sheffield, Sheffield, U.K., in 1986, the M.Sc. degree in artificial intelligence in 1987 and the Ph.D. degree in speech recognition and neural networks in 1990, both from the University of Edinburgh, Edinburgh, U.K.

He is a Professor in the School of Informatics, University of Edinburgh, where he is the Director of the Centre for Speech Technology Research. From 1991 to 1992, he was a Postdoctoral Fellow at the International Computer Science Institute, Berkeley, CA, and was then an EPSRC Postdoctoral Fellow in Information Engineering at the University of Cambridge, Cambridge, U.K. (1992–1994). From 1994 to 2003, he was a Lecturer and Reader at the University of Sheffield, moving to the University of Edinburgh in 2003. His research interests are in the area of signal-based approaches to human communication, in particular spoken language processing and machine learning approaches to modelling multimodal data. He has over 100 publications in these areas.

Dr. Renals is an Associate Editor of IEEE SIGNAL PROCESSING LETTERS and has been a member of the Technical Committee on Machine Learning and Signal Processing.