ON AUTOMATIC ANNOTATION OF MEETING DATABASES

Daniel Gatica-Perez, Iain McCowan, Mark Barnard, Samy Bengio, Hervé Bourlard

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

ABSTRACT

In this paper, we discuss meetings as an application domain for multimedia content analysis. Meeting databases are a rich data source suitable for a variety of audio, visual and multi-modal tasks, including speech recognition, people and action recognition, and information retrieval. We specifically focus on the task of semantic annotation of audio-visual (AV) events, where annotation consists of assigning labels (event names) to the data. In order to develop an automatic annotation system in a principled manner, it is essential to have a well-defined task, a standard corpus and an objective performance measure. In this work we address each of these issues to automatically annotate events based on participant interactions.

1. INTRODUCTION

Multimedia content analysis addresses, among many others, the task of automatically annotating audio-visual material with labels relevant to browsing and retrieval [14, 2]. Annotation is a rich domain; here we focus on single labels that name semantic entities. The labels could consist of relatively low-level events or concepts, such as words, identities and specific objects, as well as higher-level semantic concepts, such as a weather report within news, a goal within a football match, or the genre of a documentary.

Automatic video annotation has mostly been focused on a small number of application domains, including broadcast news, sports videos, and documentaries. The data in these applications is highly produced, and thus has a strong structure due to shot cuts that segment a video into a coherent 'story'. For this reason, most approaches to date have used shots as the basis for event segmentation and classification. In the more general case, however, produced content cannot be assured, and so this reliance on shots as the fundamental unit for processing and recognition is a limiting assumption. Meetings [9, 13, 3] provide a counter-example in which the input media naturally consists of raw AV streams.

In order to progress from low-level to more semantic annotations, statistical models are commonly used to infer high-level events from lower-level AV features. In particular, Hidden Markov models (HMMs) are sequence models that for the task of video annotation have been used to model the content of broadcast material such as news [4], documentaries [6], and sports [15].

Two distinct approaches to such event-based semantic annotation exist. Perhaps the most common approach is to consider a set of events that occur sporadically within the data stream. Commonly, systems following this approach classify each segment (usually a shot) according to the presence/absence of each event using decision thresholding. A second approach is to consider that the data stream consists of a continuous sequence of events, and in this case continuous decoding strategies can be employed, alleviating the need for a pre-segmentation. In this paper we investigate such an approach, in which meetings are decomposed into a sequence of *meeting actions*, such as monologues, discussions, and presentations. The annotation task is then clearly defined as recognising the correct sequence of meeting actions.

As well as having a well-defined annotation task, there is a need for standardised measures by which the quality of the annotations can be assessed. One response to this need is the NIST TREC video track project [12]. The metrics used in the NIST TREC 2001 evaluation were *recall* and *precision*, which relate mainly to retrieval and two-class classification problems. However, to assess the quality of video annotations involving more than two classes, performance measures are still non-standard. In this paper, we advocate the use of the word error rate commonly used in the speech recognition domain [10]. If the video annotation task is defined as the recognition of a continuous sequence of events (as discussed above), and when shot-cut boundaries are not present (or are irrelevant to semantic events), the word (or event) error rate presents a natural and effective metric for system performance.

The paper is organised as follows. Section 2 discusses semantic annotation of meetings in the context of multimedia content analysis. Section 3 describes our approach in detail, including event definition, the performance evaluation protocol, and discusses both its applicability to other domains and its limitations. Section 4 provides some final remarks.

2. MEETINGS AS MULTIMEDIA DATA

Meetings depict people interaction, and occur in reasonably constrained yet challenging conditions. As a source of multimedia information, meetings consist of unedited streams of audio and video, captured with multiple cameras (covering participants and workspace areas, including white-boards or projector screens) and microphones. A possible production model for real-time communication could merge AV streams focused on the current speaker(s) [3], possibly using motorised cameras. However, in many real settings (including the one described in this paper) cameras are fixed and AV data are archived in raw form.

In this setting, the typical concepts of shots and scenes are absent. Cameras and microphones continuously capture people engaged in discussions, gesticulating, and making/listening to presentations. The continuous nature of the data renders methods for discovering syntactic rules (commonly studied in multimedia content analysis) of relatively little relevance.

However, meetings are strongly structured data in semantic

This work was carried out in the framework of the Swiss National Center of Competence in Research on Interactive Multimodal Information Management (IM)2, and the European project M4.

terms. Events at different semantic levels, ranging from low-level actions and gestures (entering or leaving the room, standing to make a presentation, raising a hand) to high level actions (discussing, doing monologues, making presentations) to very high-level notions (planning, negotiating, making decisions) are all common in meetings, and induce semantic hierarchies in the data. In many cases, these events represent meaningful annotations and could be directly used as queries in a retrieval system.

As multimedia data, meetings have common features with other data types generated by "looking at people", like surveillance and instructional videos (raw data, multiple cameras, with an obvious difference in the number and quality of the audio sources), and also share characteristics with some highly produced content, like news and interview programs (where speakers play a leading role, and the audio track represents a very strong cue). As a result, many of the problems in analysing meetings are shared by other domains.

2.1. The relevance of analysing meetings

Meetings have been extensively studied by social psychologists for over fifty years, with the general purpose of understanding group dynamics and communication [1, 8]. However, the automatic analysis of meetings constitutes an emerging field [9, 13, 3].

Regarding automatic annotation, the amount and relevance of semantic labels that can be potentially extracted from meetings are considerable, both from what is said and from what is done. At the individual meeting level, annotations could improve collaborative work by helping people quickly retrieve information from a meeting archive without having to listen/view entire recordings. At the database level, important high-level trends could be discovered and attached to the database as labels, useful for instance for organisational management tasks.

Paraphrasing [2], meeting databases have three clear highlights in terms of research relevance and applicability. In the first place, meetings -as raw data- are suitable for the automatic generation of metadata not available from production. Although identities of participants -and to some extent the basic semantic structure if an agenda was available- could be potentially extracted by automatic means at the time of acquisition, people statements and actions are natural and cannot be generated at production time without human intervention. In the second place, off-line annotation of meetings is a task for which humans are not good/fast at generating. In the third place, meetings occur regularly and are often generated in large numbers. As the number of meetings increases, however, their individual value tends to decrease, mainly in terms of novelty. Analysing a meeting database would increase the value of the raw data if annotations related to management tasks (like the progress of a project over a period of time) could be produced.

Needless to say, the research problems at hand are ambitious. It is also clear that, even though many of the very high-level events cannot be recognised by present state-of-the-art means (computer vision, speech processing, data fusion, language modeling), the current technology (as witnessed by other domains in multimedia content analysis) should allow for the labeling of low- and some high-level semantic events.

2.2. Semantic annotations in meetings

Although far from perfect, identifying meeting participants, transcribing what they say, and partially inferring what they do, are becoming feasible. However, given the large list of potential events, what specific events should we (or can we) annotate? We can briefly state four broad categories, each generating annotations of distinct (but possibly complementary) nature and complexity.

Speech transcription-based annotation. The analysis of Automatic Speech Recognition (ASR) transcriptions by language modeling and text retrieval techniques is expected to generate the highest-level concepts for annotation, varying from specific key-word detection and recognition of participant identities, to topic and subtopic detection, etc. [13, 9].

Audio-based annotation. When rooms are equipped with microphone arrays, the approximate location of participants can be robustly inferred from the audio streams [5], so location-based events can be identified, including monologues, turn-taking, or presentations. However, the amount of (non-speech) events that can be extracted from audio-only is limited.

Video-based annotation. Recognition of people and some of their actions can currently be addressed by computer vision algorithms to perform person identification in individual meetings (face detection/recognition) and across meetings (face clustering), gesture recognition, facial expression recognition, etc. Video-based annotation faces two main challenges : robustness and usefulness. Many of the low-level semantic labels that can be generated (usually related to recognition of sparse, low-level actions performed by individuals) do not constitute annotations directly useful for indexing or retrieval (nobody needs to query a system looking for people standing up from their seats, or pointing to the white-board). In other words, the problem of mapping low-level gestures to semantically meaningful concepts remains open, as in all other multimedia content analysis domains. These events, however, can be the building blocks towards recognising high-level semantic events. Note that the definition of high-level events admits multiple dictionaries and different levels of semantic granularity.

Multi-modal annotation implies the development of principled frameworks for the integration of multiple data streams of different nature and frame rate (audio, text from speech transcripts, and video in this paper) to detect events. In meetings, events are inherently multimodal, but the involved modalities have complex relations (they might be asynchronous, and contain significantly distinct amounts of relevant information related to the event). The general goal is to combine low-level features and events provided by the individual modalities into high-level event recognisers.

3. ANNOTATING MEETINGS

In this section, we consider meetings as continuous sequences of AV events with natural transitions. If a list of possible events is defined, the task of annotation then consists of finding the sequence of events that constitutes a particular meeting. Given such a definition, the video annotation task becomes analogous to that of speech recognition, and so a similar training, decoding and assessment methodology can be employed.

3.1. Definition of events

Many different sets of events could be defined for the task of meeting annotation. In [7], we proposed a list of events characterised by group behaviour of meeting participants. This list included monologues (by participant), discussions, consensus, disagreement, presentations, white-boards and note-taking. These are all natural actions in which participants play and exchange similar or complementary roles. As these events are based on group interactions, we refer to them as *meeting actions*.

The definition of such a lexicon of meeting actions is interesting from a research perspective, as recognition of group interactions could be approached from at least two distinct angles. In a first case, the actions of individual participants could be recognised, and then these responses fused at a higher level to recognise the interaction. Such an approach, however, overlooks the fact that the behaviour of individuals in meetings is somehow constrained by the behaviour of the other participants. A second approach (taken here) is to model the interactions directly, by integrating all observations into a unique probabilistic model and learning the constraints from the data. If the group as a whole provides enough evidence for the performed action, recognition of personal actions could be bypassed altogether, potentially increasing robustness to imperfect feature extraction and measurement processes.

From the retrieval viewpoint, defining the events based on group actions has the benefit of attaching a single semantic annotation to all audio-visual streams. In contrast, annotations based on individual actions would result in different annotations for each camera or microphone. Also, individual actions would tend to be sparse in nature, while the above list of meeting actions can be treated as a continuous sequence.

3.2. Methodology

To annotate meetings as a sequence of events, we use statistical generative models based on HMMs [10]. HMMs have been successfully used to recognise speech, visual and audio-visual sequences. When the video annotation problem can be posed as recognising a continuous sequence of events, techniques and assessment metrics can be borrowed from these other tasks.

To use HMMs to annotate meetings, we require an event lexicon (as described above) and feature vectors appropriate for measuring the defined events. Given a training sequence of feature vectors with the corresponding labelling (but not necessarily the precise alignment), HMMs can be trained using the classical embedded training method based on Expectation-Maximisation (EM). Recognition then simply involves application of the Viterbi decoding algorithm to find the most likely sequence of states which can be translated into the corresponding sequence of meeting actions.

3.3. Meeting database

A corpus of meetings was recorded in the IDIAP smart meeting room [7]. Meetings were recorded using 3 cameras and 12 microphones, with all channels fully synchronised. Currently the database contains 60 meetings (30 train, 30 test), where each meeting consists of 4 participants and lasts approximately 5 minutes. The meetings are loosely scripted in terms of the type and schedule of the high-level actions, but otherwise the content is natural. The corpus is fully described in [7] and is publicly available [16].

3.4. Performance evaluation

Speech recognition is often quoted as an example of a processing domain where research has been greatly aided by the use of standard performance metrics, facilitating comparisons between different systems. While performance measures for retrieval have been largely standardised, methods of assessing the accuracy of video annotations are still largely system dependent [4, 15]. A major benefit of posing the annotation problem as described above, is that standard performance metrics, such as the word error rate used in speech recognition, may be employed. This was acknowledged long ago in computer vision for gesture recognition [11].

The above methodology for annotating meetings was applied in [7] using our meeting corpus. A feature vector of 19 audio-visual features was extracted from the input channels at a rate of 5 Hz. From 2 cameras looking at people at the table, Gaussian Mixture models (GMMs) of skin/background colours in RGB space were used to extract head blobs. Skin/background pixel classification and morphological post-processing were performed inside image regions enclosing typical head locations. For each person, the detected head blob was represented by the vertical position of its (normalised) centroid. From a wide-view camera capturing the presentation screen and white-board area, moving blobs were detected by background subtraction and represented by their (quantised) horizontal position. Audio features were extracted to measure the speech activity of different locations, as well as the occurrence of a set of positive and negative keywords. These features were used to train HMMs (left-to-right topology, GMM emission pdfs, with number of states per event, and mixture components per state chosen by cross-validation) using the train set.

The system performance was assessed on the test set in terms of the *action error rate*, which is equivalent to the word error rate in speech recognition. The word (event or action) error rate is an appropriate metric where finding the correct sequence of annotations is more important than precisely determining their temporal boundaries. This is often the case when the annotation labels are high-level semantic concepts. The word error rate is calculated as $100 \times$ the ratio between the number of substitution, insertion and deletion errors, and the correct number of words. Due to the inclusion of insertion and deletions in the error rate calculation, it is a more severe measure than classification accuracy. Video annotation systems are commonly designed as 'shot classifiers', in which case insertions and deletions do not occur, however the use of shots as a fundamental unit is often not appropriate, and the word error rate is a measure with more general application. The overall action error rate achieved in these experiments was 20.0% [7].

In addition to a standard performance measure, it is also necessary to analyse the results to determine common sources of errors. A useful analysis tool for a small vocabulary recognition task is the *confusion matrix*, which shows the distribution of recognised events according to events in the ground-truth (note that this differs from the standard multi-class confusion matrix, due to the lack of hard boundaries, consequently including deletions and insertions). The confusion matrix for the above task is shown in Table 1. Analysis of the confusion matrix is particularly useful in this case, as it shows that neither consensus and disagreement are recognised correctly, instead being commonly confused with discussion or deleted. These are examples of events for which the features are clearly not discriminative enough; the issue requires further research. This observation is discussed in detail in [7], and it is shown that if consensus and disagreement are removed from the lexicon by relabelling them as discussions for both training and testing purposes, then the action error rate decreases to 5.7%.

3.5. Applicability to other video annotation domains

The above methodology for meeting video annotation is applicable to other domains where inherent structure exists such that the video can be considered as a continuous sequence of events. For example, in [4], such an approach is taken to annotate televised news broadcasts in terms of content classes. Sports videos and

	mono1	mono2	mono3	mono4	white	note	cons	disc	pres	disa	DEL
mono1	10										1
mono2		9							1		
mono3			17								
mono4				10							1
white					18						
note						6					
cons								6			9
disc								45			
pres					1				12		1
disa								1			7
INS		1		1				1			

Table 1: Confusion matrix of recognised meeting actions, including monologues (mono1-4), white-boards (white), note-taking (note), consensus (cons), discussions (disc), presentations (pres) and disagreements (disa). Zero values are represented as empty cells. Columns and rows show desired and obtained labels, respectively.

documentaries are other domains where such structure may exist.

While such an approach has been investigated across different annotation domains, the method of reporting results is still non-standard. Different methods of assessing event segment boundaries are used, and classification accuracies differ based on an assumed level of segmentation (frames, shots, scenes, etc). As discussed above, where the annotation labels are high-level semantic concepts (e.g., presentations, discussions, interviews, shots at goal), often the concept of precise boundaries between segments has little relevance. Also, as multiple events can occur within a video shot, or conversely a single event could span multiple shots, shots are not always an appropriate unit for classification. In such a context, the 'word' error rate is a meaningful performance measure that could be adopted across different video annotation systems recognising such a continuous sequence of semantic events.

3.6. Limitations

This methodology for video annotation has a number of limitations. First, it excludes the co-occurrence of multiple events at a given time. Second, it cannot explicitly handle the case when events occur sporadically, and not as a continuous sequence. In some cases, the first limitation could be addressed by employing a hierarchical recognition scheme, in which recognised events are decomposed into a further sequence of sub-events. As a simple example of this in the context of meetings, we could handle the occurrence of note-taking during presentations by first recognising presentations, and then recognising this as a sequence of segments with or without note-taking. Clearly such a hierarchical system has limited application, and a need exists for a more general methodology allowing the joint occurrence of multiple events.

The second limitation could be addressed by introducing a 'silence' or 'garbage' event model to match periods where no explicit events occur. This is analogous to the approach taken in speech keyword spotting systems. In such an approach, however, selection of an appropriate garbage model is often a non-trivial task.

4. CONCLUSIONS

This paper has discussed meetings as a source of data for multimedia content analysis, specifically focusing on the task of automatic audio-visual event annotation. A methodology for treating meetings as a continuous sequence of events was proposed, leading to a well-defined annotation task and clear performance evaluation. As a case study, a system annotating a database of meetings as a sequence of meeting actions (monologues, presentations, discussions, white-boards, note-taking, consensus and disagreement) was presented and assessed in terms of the word (action) error rate. The advantages of our methodology, and its applicability to other types of multimedia data were discussed, along with limitations of the approach. In conclusion, we propose to the research community the use of this corpus [16], in general, and the particular task and evaluation measure used in this article (and [7]).

5. REFERENCES

[1] R.F. Bales. *Interaction Process Analysis: a method for the study of small groups*. Addison-Wesley, 1951.

[2] S.-F. Chang. The Holy Grail of Content-based Media Analysis. *IEEE Multimedia Magazine*, 9(2), Apr. 2002.

[3] R. Cutler, Y. Rui, A. Gupta, JJ Cadiz, I. Tashev, L. He, A. Colburn, Z. Zhang, Z. Liu, and S. Silverberg. Distributed Meetings: a Meeting Capture and Broadcasting System. In *Proc. ACM MM Conf.*, 2002.

[4] S. Eickeler and S. Müller. Content-based Video Indexing of TV Broadcast News using HMMs. In *Proc. ICASSP*, Phoenix, 1999.

[5] G. Lathoud and I. McCowan. Location-based Speaker Segmentation. In *Proc. ICASSP*, Hong Kong, 2003.

[6] T. Liu and J. R. Kender. A Hidden Markov Model Approach to the Structure of Documentaries. In *IEEE Work. on CBAIVL*, 2000.

[7] I. McCowan, S. Bengio, D. Gatica-Perez, G. Lathoud, F. Monay, D. Moore, P. Wellner, and H. Bourlard. Modeling Human Interaction in Meetings. In *Proc. ICASSP*, Hong Kong, 2003.

[8] J.E. McGrath. *Groups: Interaction and Performance*. Pr.-Hall, 1984.

[9] N. Morgan, D. Baron, J. Edwards, D. Ellis, D. Gelbart, A. Janin, T. Pfau, E. Shriberg, and A. Stolcke. The Meeting Project at ICSI. In *Proc. Human Lang. Techn. Conf.*, San Diego, March 2001.

[10] L. Rabiner and B.H. Juang. Fundamentals of Speech Recognition. Pr. Hall, 1993.

[11] T. Starner and A. Pentland. Visual Recognition of American Sign Language using HMMs. In *Proc. Int. Work. on Auto. Face and Gesture Recognition*, Zurich, 1995.

[12] E. M. Voorhees and D. K. Harman, editors. *NIST Special Publication 500-250: The Tenth Text REtrieval Conference (TREC 2001).*

[13] 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. ICASSP*, Salt Lake, 2001.

[14] Y. Wang, Z. Liu, and J. Huang. Multimedia Content Analysis using Audio and Visual Clues. *IEEE SP Magazine*, 17(6), Nov. 2000.

[15] L. Xie, S.-F. Chang, A. Divakaran, and H. Sun. Structure Analysis of Soccer Video with HMMs. In *Proc. ICASSP*, 2002.

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