

Disclosure of non-scripted video content: InDiCo and M4/AMI

Franciska de Jong ^{1,2}

¹University of Twente, Department of Computer Science,
P.O. Box 217, 7500 AE Enschede, The Netherlands
fdejong@ewi.utwente.nl

²TNO TPD, P.O. Box 155
2600 AD Delft, The Netherlands

Abstract. The paper discusses three IST projects focusing on the disclosure of video content via a combination of low-level multimodal feature analysis, content abstraction, and browsing tools. The type of content (recordings of conference lectures and meetings) can be characterized as non-scripted and is argued to generate a whole range of new research issues. Performance results are reported for some of the tools developed in InDiCo and M4.¹

1. Introduction

Since the mid nineties automated disclosure of video content has been on the agenda of several R&D projects within the European funding programmes. These projects have built on the insights and results obtained in fostering research at several laboratories and archiving institutes in Europe and elsewhere [1], [2]. As video material is a typical *multi*-media type of content, video indexing can in principle deploy and combine the analysis tools for various media. For video with a soundtrack two data channels are important, thus a first distinction is commonly made between audio and video. But to each of these two data types multiple modalities correspond. Audio may consist of speech, music and other sounds, either or not in combination. Video as it appears on a screen can also involve more than just images: often additional captions or subtitles come along, while shot images may contain textual elements as well. Most of these symbolic elements belong to the same realm as the speech parts in the audio: natural language, a medium that can help bridge the semantic gap between low-level video features and user needs, and that can also very well be exploited for the generation of time-coded indexes [3]. As for image content from the general domain no efficient and effective information semantic disclosure was nor is likely to become available soon, the deployment of speech and language processing tools played an important role in most early video retrieval projects ([4], [5]), and in the video retrieval evaluation conferences organised in the context of TREC ([6]).

¹ Thanks go to Mike Flynn and Dennis Reidsma for support in preparing this paper.

For obvious reasons (availability of test and training collections, performance expectations, commercial interest) the focus has long been on a specific type of pre-produced scripted content: broadcast news programmes. Evidently there are still a lot of problems to be solved in this domain, but due to both the boost of digital production and retrospective digitisation, an enormous growth in the number and size of digital video collections can be observed. A considerable part of this content can be characterized as non-scripted: recordings of events for which no script is available that strongly determines the behaviour of all registered people and objects. Examples are: logs of videoconferences, images captured by surveillance camera's, live reports of significant events from various domains (sports, celebrations, public ceremonies, etc.), and recordings of discussions, lectures and meetings. This paper will discuss three on-going IST-projects that have chosen recordings of non-scripted events as the target for content retrieval technology:

- project M4 (MultiModal Meeting Manager)
<http://www.m4project.org/>
- project AMI (Augmented Multi-party Interaction)
<http://www.amiproject.org/>
- project InDICO (Integrated Digital Conference)
<http://indico.sissa.it/>

Though in each of these projects a number of different functionalities determine the research agenda, they all pay attention to the disclosure and retrieval of video content at the level of fragments, partly based on the use of language and speech processing to support a form of content abstraction and/or the automated generation of metadata. As the image content is relatively static, dominated by talking heads and moving bodies, not the indexing of video frames is the primary target, but the realisation of a truly multimedia retrieval environment in which all information modalities contribute to the disclosure of a media-archive up to a high level of granularity.

One of the interesting things about the non-scripted data collections is that they open up new types of usage that may require a much more diverse range of analysis and presentation techniques. The type of collections to be discussed here for example, are not just of interest for what they are about, but they also are a valuable source for researchers interested in the study and modelling of human interaction, both in terms of verbal and non-verbal behaviour. In order to allow researchers to use a meeting corpus for this aim, the analysis should not just focus on the image or speech content, but also on interpretation of multi-modal aspects like gaze, gesture, etc. New requirements will be posed on the fission of data, and it is likely that new designs are needed to enable content abstraction for data sets with an information density that is incomparable to what news archives have to offer. Likewise the presentation and ranking of retrieval results should facilitate search tasks that are not supported by the common tools for media fusion and browsing, partly because the content can be viewed from multiple perspectives. Therefore the inseminating role non-scripted data collections may play for the field of video indexing should not be underestimated.

This paper is organized as follows. In section 2 this paper describes the aims for each of the three projects at a general level. Section 3 will discuss the characteristics of non-scripted content and section 4 will present some of the results. The paper will be concluded by a discussion of how the performance of disclosure tools for non-

scripted video content may affect future research in the field of content-based video retrieval.

2. Projects' aims

Both M4 and its follow-up project AMI focus on meeting recordings, while InDiCo aims at the disclosure of lecture recordings. The objectives vary considerably.

2.1 M4

The M4 project is concerned with the construction of a demonstration system to enable structuring, browsing and querying of an archive of automatically analyzed meeting recordings. The main focus is on a corpus of recordings of meetings that take place in a meeting room equipped with multimodal sensors (microphones, camera's) for a limited number (4) of participants, but some additional effort is put in the analysis of recordings of parliamentary sessions, with non-directed, natural behaviour and interaction.

In the case of recordings from the dedicated meeting room, data of a variety of types is generated and deployed in the analysis and disclosure of the meeting content. In addition to multiple audio channels and video streams from several cameras, there is additional information coming from interaction with PC's and an instrumented white board. Several tools for the segmentation of the audiovisual content are investigated. Experiments have been done with location-based speaker segmentation, and beamformed microphone array data. Visual processing focuses on the development of an audio-visual speaker tracker (which switches between speakers and works across cameras), face detection and tracking, and gesture recognition (e.g., pointing, standing up, sitting down.)

A meeting may be accessed by its structure (interaction patterns, dialog-acts, turn-taking, etc.), but also by what the participants say. The envisaged browsing facility that will make the content available via a media file server will initially focus on the latter, while the development for which some results are reported in section 4 is aimed at the former. To capture the structure a series of meeting actions has been defined (monologue, discussion, presentation, consensus, disagreement, ...) and models have been trained to automatically segment meetings in terms of these group actions, using audio features (such as speech activity, intonation, key words) and visual features (such as head detection). Cf. Figure 1 for a screen shot of an initial browsing prototype, displaying the structure of a discussion on favorite movies.²

2.2. AMI

AMI targets computer enhanced multi-modal interaction in the context of meetings. The project aims at substantially advancing the state-of-the-art, within important

² Development of the M4 browser was done at IDIAP, Martigny.

underpinning technologies (such as human-human communication modeling, speech recognition, computer vision, multimedia indexing and retrieval). It will also produce

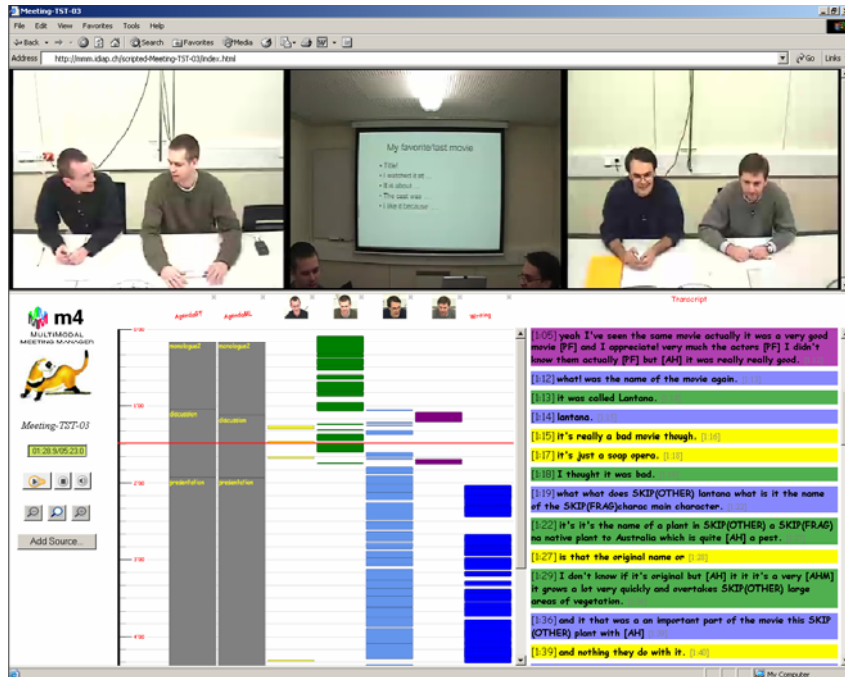


Figure 1: prototype of M4 browser

tools for off-line and on-line browsing of multi-modal meeting data, including meeting structure analysis and summarizing functions.

AMI performs the above research in the framework of a few well-defined and complementary application scenarios, involving an offline meeting browser, an online remote meeting assistant and integration with a wireless presentation system. These scenarios are being developed on the basis of smart meeting rooms and web-enhanced communication infrastructures. For the purpose of this paper the following three research areas are most relevant:

1. Multimodal low-level feature analysis, including multilingual speech signal processing (natural speech recognition, speaker tracking and segmentation) and visual input (e.g., shape tracking, gesture recognition, and handwriting recognition).
2. Integration of modalities and coordination among modalities, including (asynchronous) multi-channel processing (e.g., audio-visual tracking) and multimodal dialogue modeling.
3. Content abstraction, including multi-modal information indexing, cross-linking summarizing, and retrieval.

2.3 InDiCo

The objectives of the InDiCo project are to automate the process of managing conference content (papers, presentations, lecture recordings) for improved information sharing and exchange of the content via Internet. This is pursued by developing tools for the indexing and browsing of conference content, and to integrate these tools into an existing platform for digital publishing. Validation will be based on an experimental evaluation within the high-energy physics community at the CERN institute. Segmentation of lecture recordings into a segment-per-slide structure is a key technology. Additionally InDiCo developed domain specific speech recognition for non-native speakers and automatic clustering for the cross-linking of the content of conference speech transcripts, papers, slides and presentations, based on a memory-based learning classifier [9]. Eventually the result will be a novel navigational structure between video, slides and papers allowing users to combine the background of a paper with the compressed contents of a presentation.

3 Non-scripted content characteristics

In section 1, non-scripted content is described as recordings of events for which no script is available that strongly determines the behaviour of all participating people and objects. What this implies becomes immediately clear from the contrast with the characteristics of broadcast news programmes or movie scenes and other pre-produced and directed video recordings: what is said by whom is more or less prescribed (read speech), the position and movements of the ‘actors’ are heavily constrained, camera positions and turns follow a foreseeable pattern

Though in each of the three projects a number of different functionalities determine the research agenda, common goals are the development of tools for:

- indexing of video at the level of fragments, partly based on the use of language and speech processing
- content abstraction, to support efficient browsing through large volumes of archived content

So far there is a great overlap between research issues relevant for e.g., broadcast news archives. However, the difficulties to be solved differ because of the nature of the content and the envisaged applications. Here are two (not complete) lists of characteristics and requirements for the processing of two types of distinct non-scripted events, each corresponding with research issues that are absent or less dominant in the news domain:

Meeting recordings:

- low density of information
- gaze, gesture and movement may mark crucial events
- lack of training and evaluation material
- lack of evaluation methodology for the demonstrator functionality
- non-sequential speaker segmentation output due to overlapping speech

Conference recordings:

- ASR needed for non-natives
- lack of models for multimodal aspects of non-verbal presentation elements
- no fixed model of audience-speaker interaction
- domain models needed for highly specialised topics

Though incomplete and impressionistic, the two lists differ enough to illustrate that non-scripted multimedia content retrieval for these two domains can be viewed as two parameter instantiations, representing a more complex and a more simple case, respectively. Cf. Table 1 for a comparison of the varying parameter values for the *three* content domains and/or corresponding browse scenarios distinguished thus far.

| | Pre-produced news broadcast | Meeting with n participants | Lecture by 1 speaker, audience of n people |
|--------------------------------------------------------------------------------------------------------|-----------------------------|-------------------------------|----------------------------------------------|
| a. Number of central speakers roles | 1 or 2 | 2- n | 1 |
| b. Number of interactors per segment | 2 | 2- n | 2- n |
| c. Number of interaction moments | low | relatively high | relatively low |
| d. Number of cameras/ microphones | 1 or 2 | 1 - n | 1 or 2 |
| e. Variation in speech characteristics (pronunciation, lexicon) | little | dependent of n | dependent of n and (c) |
| f. Number of topics to be addressed | open | open | limited |
| h. Availability of data for model training and evaluation (e.g., scripts, annotated corpora, metadata) | OK | limited | limited |

Table 1: Varying parameter values for scripted and non-scripted content

The suggestion of Table 1 is that the overall complexity of analysing non-scripted content is highly related to the number of interacting speakers, the type of speech (read/non-read) and the availability of training data. The next section will present the current performance of some analysis tools for low-level features that eventually may contribute to a level of understanding for both types of non-scripted that could be called ‘semantic’.

4. Performance evaluation

This section will report on the evaluation for some of the work in M4 and InDiCo. For AMI no performance figures are available as the project started only in 2004.

4.1 M4: Location-based speaker tracking³

Automatic annotation of meetings in terms of speaker identities and their locations, which is crucial for the higher level segmentation, is achieved by segmenting the audio recordings using two independent sources of information: magnitude spectrum analysis and sound source localization. We combine the two in an HMM framework. There are three main advantages of this approach. First, it is completely unsupervised, i.e. speaker identities and number of speakers and locations are automatically inferred. Second it is threshold-free, i.e. the decisions are made without the need of a threshold value which generally requires an additional development dataset. The third advantage is that the joint segmentation improves over the speaker segmentation derived using only acoustic features. Experiments on a series of meetings recorded in the IDIAP Smart Meeting Room demonstrate the effectiveness of this approach. For more details, cf. [10].

| Clustering type | HTER | ACC |
|-----------------|------|------|
| Acoustic Only | 19.2 | 92.6 |
| Location-based | 17.3 | 94.6 |

Table 2: Speaker segmentation performance percentages

4.2. Speaker turn pattern segmentation

A corpus of meetings recorded and annotated by ICSI [7] was used for the work on speaker turn segmentation by the University of Sheffield. A 60-minute meeting has been segmented using the Bayesian Information Criterion (BIC). This segmentation was compared with manual topic segmentation. Cf. Figure 2.

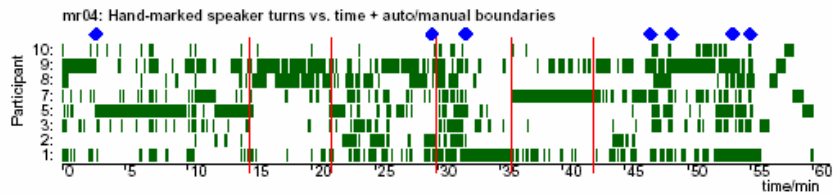


Figure 2: Each row corresponds to a different participant. The meeting ended with most participants separately reading a series of digits (56 minutes onwards). The 5 vertical lines indicate boundaries from BIC segmentation; 7 (blue) diamonds show the hand-marked topic boundaries for this meeting.

Of 36 manually-marked topic boundaries over 6 meetings, only 15 agreed with turn-based segmentation (margin 2 minutes); in addition, 16 turn-based boundaries were found that had no corresponding topic boundary. It seems turn-pattern boundaries are

³ Work carried out at IDIAP, Martigny.

not directly related to discussion topics, although they may provide an important alternative perspective on the temporal structure of meetings. For more details, cf. [8].

4.3 InDiCo: Speech recognition⁴

The project has used the SPRACHcore recognition engine, distributed by ICSI, Berkeley.⁵ Language models and vocabulary have been optimized with respect to the InDiCo development test database. The Word Error Rate (WER), which was over 80% before optimization, has been reduced to 67%, using a mixed language model based on 400 million words of North American Newspaper texts and the InDiCo document set, 45 million of pre-print texts from the high energy physics domain, from which a vocabulary of 40k words has been derived. The WER figures are likely to improve if a proper pronunciation dictionary, but already are becoming useful for certain retrieval tasks.

4.4 InDiCo: Video analysis⁶

The work on video analysis consists of two tasks: slide segmentation and slide matching. A video containing a slide presentation has to be segmented in such a way that (i) each segment contains no more than one slide, and (ii) a slide that is shown without interruption belongs to one segment only. Via slide matching the slide source can be synchronized with the video time codes. Experiments have been done for two types of slides: (i) *PowerPoint presentations*, which usually have very clear contrast, often colour and clear borders, and (ii) *Printed presentations sheets*, typical for CERN, with vague borders, no colour and illumination varying over the sheet. Results for the former are almost perfect. Results for the latter are still rather poor: on a test set of 8 hours of CERN video data, an accuracy of 65 % was obtained.

4.5 InDiCo: Linking lecture fragments and papers⁷

The cross-link utility developed at TNO to support the navigation lecture fragments to slides and from slides to papers is based on machine learning. For every page in a paper, all n-gram word sequences are extracted, which are labelled with the page number. As n-grams implement a shifting window, this produces instances like

```
Deceleration, of, antiprotons, has, been, page_6  
of, antiprotons, has, been, demonstrated, page_6  
antiprotons, has, been, demonstrated, in, page_6
```

⁴ Work carried out by TNO Human Factors, Soesterberg.

⁵ Cf. <http://www.icsi.berkeley.edu/~dpwe/projects/sprach/>

⁶ This work was performed at ISIS, University of Amsterdam.

⁷ Work carried out by TNO TPD, Delft.

A memory-based learning classifier [9] is trained on these labelled n-grams. For every slide in a presentation, n-grams are extracted along the same lines. The trained memory-based classifier is applied to every n-gram in the slide, and all classifications are gathered. Finally, majority voting is applied to the set of collected predictions, and the class with highest frequency was selected as the winning classification of the slide. The test and training data set consisted of 50 paired papers and slides from the 8th European Particle Accelerator Conference. A link was judged to be correct if (a) it was formally plausible and (b) if it was sequentially plausible, that is: not linked to a page too far from the page the previous slide was linked to. On a representative subset of 8 presentations, approximately 90% accuracy has been achieved.

5. Concluding remarks

The disclosure of recordings of non-scripted events imposes different requirements than fully directed events such as news programs. Due to the low information density of meetings, low-level audio and video features should be exploited for the recognition of high-level meeting actions and event structures that can be indexed and searched for. Metadata for conferences can be enriched by the linking of lecture recordings to collateral linguistic sources. Focus on non-scripted video content may open up new research perspectives for multimedia analysis and novel applications for multimodal information processing, but advances in this domain are highly dependent on proper training and test collections.

References

- [1] M. Maybury (ed.), "Intelligent Multimedia Information Retrieval", MIT Press, Cambridge (1997)
- [2] Content-Based Access of Image and Video Libraries. Proceedings IEEE Workshop. IEEE Computer Society, Los Alamitos, 1997.
- [3] Human language as media interlingua: Intelligent multimedia indexing. In: Proceedings of ELSNET in Wonderland. ELSNET, Utrecht (1998) 51-57.
- [4] Jong, F. de, Gauvain, J.L., Hartog, J. den, and Netter, K., Olive: Speech-based video retrieval". In Proceedings of CBMI'99. Toulouse (1999) 75-80
- [5] Jong, F. de, Gauvain, J.-L., Hiemstra, D., Netter, K., 2000. Language-Based Multimedia Information Retrieval. In: Proceedings of 6th RIAO Conference. Paris (2000) 713-722
- [6] Smeaton, A.F., W. Kraaij and P. Over, TRECVID -An Introduction, In: Proceedings of TRECVID 2003. Gaithersburg (2003)
- [7] Morgan, N., D. Baron, J. Edwards, et. al., The meeting project at ICSI. In: Proceedings HLT (2001) 246-252
- [8] Renals, S., D. Ellis, Audio Information Access from Meeting Rooms. In: Proceedings IEEE ICASSP 2003 – Hong Kong.
- [9] Daelemans, W., J. Zavrel, K. van der Sloot and A. van den Bosch, TiMBL: Tilburg Memory-Based Learner, version 5.0 (2004). Available at <http://ilk.kub.nl/papers>
- [10] Ajmera, J., G. Lathoud, and I. McCowan, Clustering And Segmenting Speakers And Their Locations In Meetings. In: Proceedings ICASSP (2004)