Chapter 5

Document-centric and multimodal meeting assistants*

In this chapter, we will show how human-computer interaction (HCI) can benefit to meeting support technology, by reviewing the evolution of HCI research from a smart meeting minutes application, with document-centric meeting browsers, towards more user-centric assistance tools for meetings. The chapter will exemplify these trends with research performed at the University of Fribourg, within the IM2 NCCR.

The chapter first focuses on a smart meeting minutes application, which consists in recording meetings and analyzing them in order to automatically produce minutes. More specifically, we show the importance of analyzing documents that are discussed or shown during meetings. Multimodal document alignment techniques link documents with other types of media such as meeting transcripts, audio, and video recordings. This opens the possibility of developing document-centric meeting browsers that use documents as indexes towards meeting parts and to the associated audio-video records. For instance, clicking on a part of a document will play the audio-video sequences of the meeting in which this document part was discussed or projected.

Further on, additional explorations are discussed, including mainly egocentric and cross-meeting browsing of large archives of multimedia meeting data through keywords, links and personal cues. A multimodal toolkit is presented, which facilitates the development of multimodal user interfaces (using voice, gestures, eyes movements, etc., to interact with machines) that can be used during meetings as online assistants to improve teamwork. For instance, a tabletop application was developed with this toolkit to facilitate brainstorming in groups.

*This chapter was written by Denis Lalanne.

Finally, we present the Communication Board (CBoard): a user-centered application that applies the previous technologies, along with other multimodal processing methods presented in this book. The CBoard facilitates remote collaboration, and displays the emotional state of meeting participants to augment teamwork performance.

Figure 5.1: A meeting room was equipped with camera/microphone pairs for up to 8 persons and several cameras for capturing documents that were projected or visible on the table. Data capture was synchronized thanks to a distributed architecture. On the right, a mosaic of the captured video streams replayed with SMIL.

5.1 The Smart Meeting Minutes application

The first major requirement to develop meeting support technology is to setup an infrastructure to record meetings. Two meeting rooms were created: one in the Idiap Research Institute, and one at the University of Fribourg, as presented in the introduction to this book. The second meeting room aimed at recording meetings where documents are often discussed or in the visual focus (projected on a screen or visible on the table). Thus, the so-called 'document-centric meeting room', was tailored to capture all the phenomena related to documents (see Figure 5.1).

The meeting room was equipped with 10 camera/microphone pairs for each participant, two overview cameras, one for the projection screen capture, and one for capturing documents on the table). The very first meting room used lightweight equipment (PCs with Firewire webcams), cheap and non-intrusive. Camera and microphone pairs' synchronization was guaranteed on a per-computer basis. Due to the volume of the data acquired on each camera, several PCs, synchronized and controlled by a master PC, were used for the acquisition. The master PC had a user-friendly interface to start, pause and stop meeting recordings, to configure post-processing

such as compression (for streaming and archiving) and to control file transfers to a server. This capture application was part of the Organizer tool which permitted to specify the participants' names and positions, which camera/microphone should be used, etc. Furthermore, the Organizer tool assisted users in the preparation and archiving of a meeting. This included services for registering meeting participants, and gathering documents and related information. At the end of a meeting, a web-based meeting browser was a automatically generated based on the available annotations, using SMIL technology to synchronously play multimedia streams. About 40 meetings were recorded in this room. Another room was created a few years later with high resolution cameras and a different architecture: a single PC with acquisition cards for 12 camera/microphone pairs.

5.2 Document centric meeting browsing

Thanks to the meeting recordings acquired in the smart meeting rooms, research on meeting analysis and browsing advanced. Concerning the meetings in which documents are discussed or projected, novel algorithms were necessary to link printable documents, that have no inherent temporality, with other media recorded during meetings such as audio and video. For this reason, research focused on multimodal document alignments and further on document-centric meeting browsers (Lalanne et al., 2003a,b, 2005).

Document-centric meeting browsers are based on the assumption that in many multimedia applications (e.g. lectures, meetings, news, etc.), printable documents play an important role in the thematic structure of discussions. The FriDoc browser, a document-enabled multimedia browsing system, considered printable documents as a portal towards multimedia content. The main purpose of the FriDoc browser was to measure the benefit of document alignments to support browsing, and more generally to assess the benefit of cross-media linking for multimedia browsing. Our hypothesis was that creating meeting browsers using links between printable documents and multimodal annotations of the audio-video streams would improve retrieval tasks.

In the FriDoc browser (see Figure 5.2), users can first search using a set of keywords within a collection of meetings. As soon as users select a document, the meeting in which it was discussed or projected is opened. All the related multimedia data (other meeting documents, audio/video clips, speech transcription, annotations) attached to this document can then be played synchronously, thanks to document alignments. Clicking on one multimedia component opens and plays the content of all the other multimedia components at the same time. For instance, clicking on a specific section of a document positions the meeting slider at the time when this section was first discussed, moves the pointer in the speech transcription at the same time,

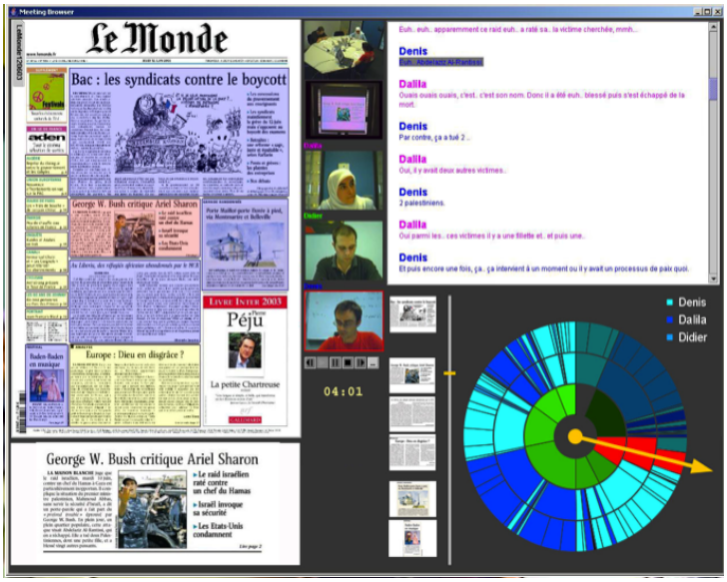

Figure 5.2: FriDoc is a meeting browser that uses printable documents as indexes to access and replay multimedia parts of meetings.

displays the document that was projected, and positions the audio/video clips at this time so that users can watch what was being said during the meeting about this section.

A user evaluation of FriDoc was performed with eight users to measure the effectiveness of using document alignments for meeting browsing. User performance in answering questions such as "Which articles from the NewYork Times have been discussed by Didier?" was measured on both a qualitative (satisfaction) and quantitative basis (e.g. success rate, task duration, number of clicks, etc.). Users had to answer several questions, with or without document alignment enabled, in the same meeting browser. This within-group experiment was properly balanced using three meetings, one for the training and two others balanced with the independent variable *with/without document alignment*. 76% of the questions were answered when document alignments were enabled in the browser, versus 66% without the alignments. For multi-modal questions, i.e. requiring information from both the speech transcript and the document discussed or projected, around 70% of the questions were solved with alignments and only half of the questions without which empirically proved the usefulness of multimodal document alignments for browsing meetings in which documents are discussed or projected.

JFriDoc (see Figure 5.3) was the improved version of the FriDoc document-centric meeting browser. It proposed novel search mechanisms and advanced

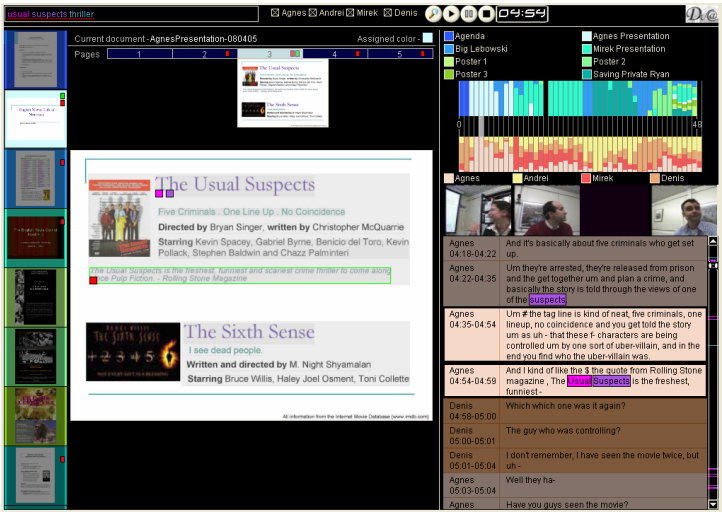

Figure 5.3: JFriDoc. The advanced document-centric meeting browser proposes a search mechanism to improve meeting browsing.

visualizations to deal with realistic data. For instance, a section of document might be discussed at several moments of a meeting, or there might be speech overlaps due to several participants speaking at the same time. For this reason, novel visualizations were developed for representing multiple alignments between static document and speech, overlapping of media, and multiple results to queries submitted by users.

Finally, in addition to meeting browsers, we have developed document-centric browsing interfaces for conference archives (SMAC) and more generally for digital libraries. This work relied in particular on three technologies: document identification, slide changes detection (Behera et al., 2004, 2008), and document-speech thematic alignments (Lalanne et al., 2003a).

5.3 Cross-meeting and ego-centric browsing

Based on our experience in multimodal document alignments, as a solution to link printable documents with other modalities such as speech, our work shifted towards cross-meeting browsing to support navigation over archives of meetings, and to support tasks such as users wanting to review the evolution of a particular topic within a series of meetings or a new employee joining a company who would like to catch up with last couple of months meetings.

FaericWorld (see Figure 5.4) is a cross-meeting navigator. It takes full advantage of the links computed between the different multimodal documents manipulated and recorded during meetings: speech transcription

of dialogs, audio-video recordings, projected slides, discussed documents, notes, emails exchanged, agenda, etc. Two corpora were integrated into the system: the IM2.DI corpus (22 meetings recorded in French) and the AMI corpus (171 meetings recorded in English). FaericWorld (Rigamonti et al., 2007) uses utterances in speech transcripts, structured content of documents, tags on videos (id codes for participants) and meeting descriptors to create links between all types of documents and annotations in the corpora.

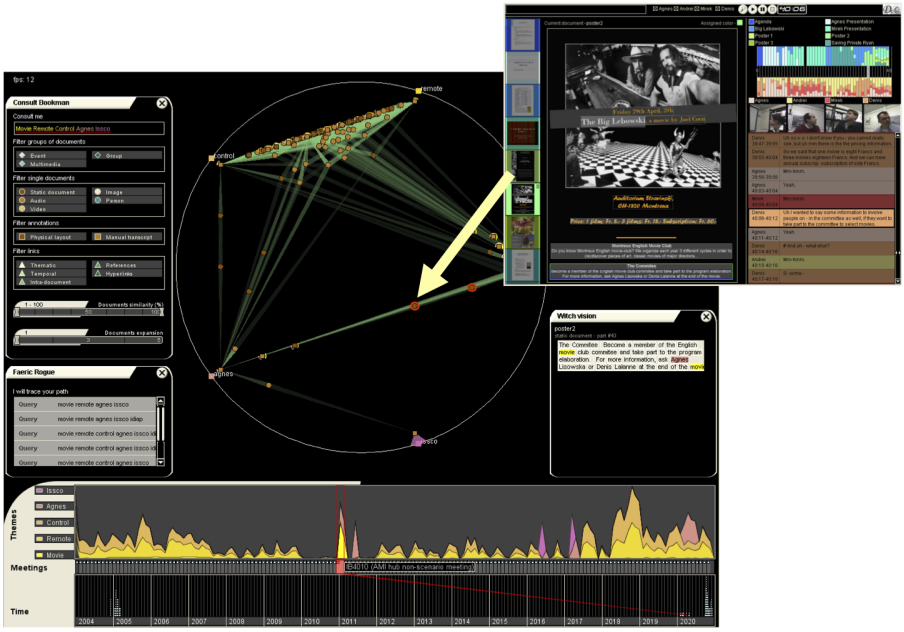

Figure 5.4: Cross-meeting browsing. The user can browse over thousands of documents from about 200 meetings. The visualization organizes the results according to a set of keywords used for browsing the archive and according to their temporal occurrences. Clicking on an item in FaericWorld visualizations opens a browser with the meeting in which the item is in the verbal or visual focus. Links to documents in the whole meeting archive change dynamically over time as the meeting is played.

Around 200 millions links (thematic or temporal) were created between parts of meetings, documents parts, speech transcripts, etc. The whole archive is displayed through a radial visualization (RadViz). Search on the archive can be performed by querying the system with several keywords and the resulting documents are displayed in the RadViz (see Figure 5.4). A default query, composed of the most recurrent words that belong to disjoint documents, provides a preliminary visualization of the archive and of its thematic structure. A document's position in the RadViz is defined using its tf.idf value for each term of the query. If the tf.idf of a term is very high for a

document (frequency of the term in the document compared to the frequency in the overall archive), the document will be greatly attracted by the related anchor. The parts of documents that match the query are also displayed in the RadViz (e.g. parts of speech transcripts, blocks of PDF documents) and linked to their parent document (e.g. the overall speech transcript). The second main visualization at the bottom of the interface is a ThemeRiver that shows the evolution of topics in the whole archive throughout time. Clicking on one item of these visualizations opens a meeting browser in which the selected item (e.g. a document part) is currently in the verbal or visual focus of the meeting (see Figure 5.4).

Another HCI activity concerned personal information management in relation to meeting recording and browsing: the TotalRecall project. The aim of this project was to support human memory in professional life, and more specifically to support humans in remembering documents exchanged during meetings, information, tasks to do, or preparing their presentations for a following meeting. In the TotalRecall project we wanted to use the implicit structure of our mailbox as a starting point to have a personalized access to meeting recordings, based on each individual user's personal interests. In the first phase of the project we collected information about how people manage meetings using digital artifacts.

We designed and performed a survey on the web inviting people to answer questions about their professional lives, meeting attendance and organization. Over about 120 participants to the survey, it turned out than more than 50% admitted that they use emails as a means to recall meeting dates, places or information exchanged between participants (Bertini and Lalanne, 2007). This was our motivation for using personal cues, derived from emails and personal documents, to access meeting records in a personalized way (Lalanne et al., 2008). For this reason, algorithms were developed to automatically extract the social network of a person based on the frequency of emails she exchanged with people. Further, an agglomerative thematic hierarchical clustering was implemented, exploiting email content similarity. Simple alignment techniques were then used to access meeting records based on person, time, or keywords. The AMI meeting corpus, which includes emails exchanged by participants between meetings, was used in order to lay the foundations of an ego-centric meeting browser, profiting from the personal information structure of each user to guide them towards the particular information they need in meetings (Evéquoz et al., 2010).

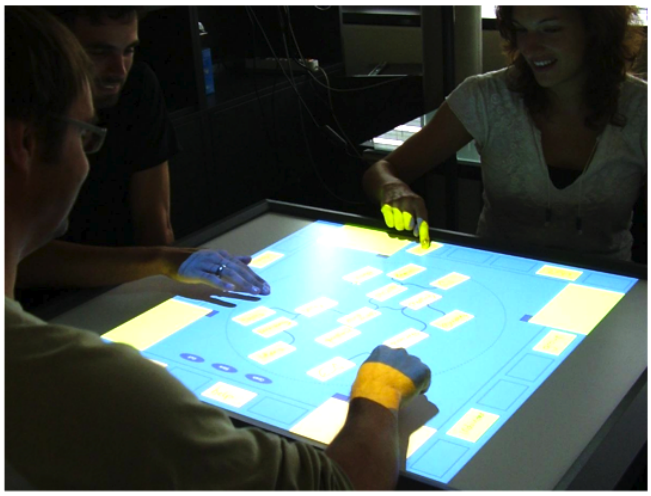

Figure 5.5: The HephaisTK toolkit helps to develop multimodal meeting assistants: here, a multi-user tabletop application supporting brainstorming.

5.4 Multimodal user interfaces prototyping for online meeting assistants

In the second phase of the IM2 NCCR, research on meeting browsers and evaluations shifted towards the development of meeting assistants, to support collaboration during meetings, and not only after to browse within meeting records. In order to support developers in the creation of multimodal interfaces, a toolkit named HephaisTK was developed (Dumas et al., 2008, 2009). This toolkit was designed to plug itself into a client application that wishes to receive notifications of multimodal events received from a set of modality recognizers, such as speech, gesture, or emotion recognizers. It was based on a multi-agents architecture, in which each recognizer is an agent that communicates with others through a central blackboard. A special agent manages fusion of input modalities, helped by a dialog agent specific to the application. A configuration file, using the SMUIML language (Dumas et al., 2010), needs to be specified to operate the toolkit. SMUIML enables the description of the multimodal dialogs: different input events, the multimodal triggers (a combination of multimodal inputs), and the actions to be performed in the specific client application. At the end the multimodal interaction is described as a finite-state machine which the fusion engine agent uses to take decisions.

Several multimodal user interfaces were developed using HephaisTK. For instance, an interactive table was created to favor brainstorming: a large interactive table on which four participants can interact at the same time using their fingers or voice as input (see Figure 5.5). Each participant has

a virtual post-it pad in front of her/him for writing down notes, sketching or drawing directly on the table with their finger. Once meeting members finish writing a note they can drag it to the center of the table using their finger to share their ideas with the other participants (i.e. mind mapping). The outcome is a file with ideas which can be processed at a later date. A user evaluation of this application showed that the multi-user capability of the application, and the fact that users can interact at the same time, encourage idea production from all the participants compared to the very same application in which only one user can interact at a time.

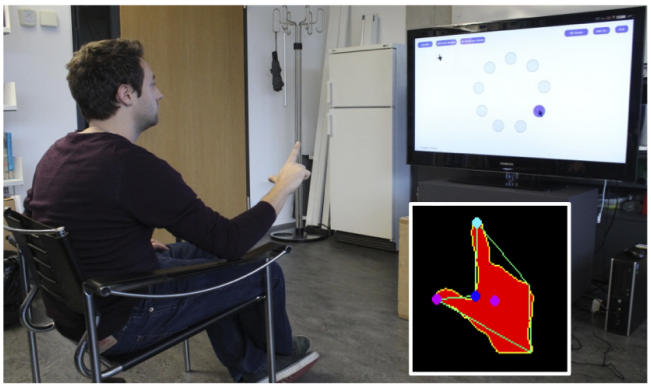

Figure 5.6: Mid-air pointing gestures to facilitate natural interactions with the Communication Board.

5.5 The Communication Board application

The Communication Board (CBoard) was a new human centered application, particularly relevant as it applies several existing technologies and assesses them through user evaluations. The CBoard is an interactive wall on which people can interact and discuss in co-presence or remotely. The CBoard enables remote collaboration since it integrates an audio-video conferencing system, and at the same time users can interact on a shared application in transparency (see picture on the right of Figure 5.7). It was inspired by the famous ClearBoard idea (Ishii and Kobayashi, 1992). Furthermore, the application served as a testbed to run user studies to evaluate multimodal technologies, and study research questions such as the impact of individual's characteristics on usability or the role of emotions in teamwork (see Chapter 2, Section 2.3.4).

Users interact on the CBoard either using 3D devices such as wiimotes or using mid-air gestures. Mid-air gesture recognizers were developed for this purpose within IM2, so that users can interact with the CBoard without the need for calibration and without having to hold a device or markers. In

this research (Schwaller and Lalanne, 2013, Schwaller et al., 2013), we have been interested in developing novel pointing and selection strategies and measuring the effect of these strategies, and of different visual feedbacks, on pointing performance and effort (see Figure 5.6).

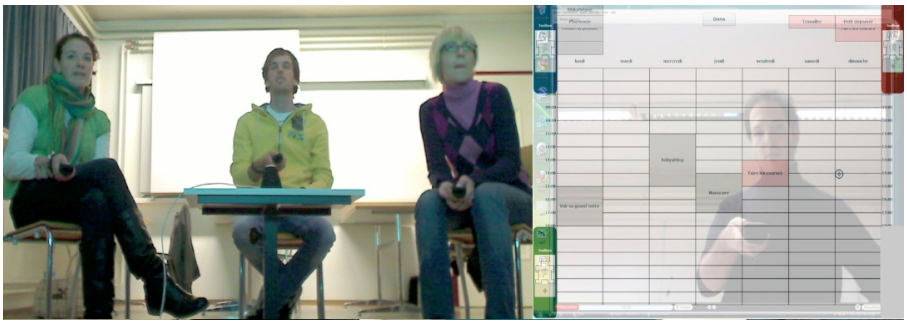

Figure 5.7: The Communication Board (CBoard) is an interactive wall on which pepole can interact and discuss in co-presence or remotely. The EmotiBoard uses the CBoard framework and displays emotional feedback of each participant on the screen, in addition to the video conference and the shared application.

Much research effort has been spent on the EmotiBoard (Sonderegger et al., 2013): an application of the CBoard in which emotional feedback of other participant is displayed on the screen in addition of the video conference and the shared application (see Figure 5.7). The EmotiBoard served as a research application to study user's affective and social behavior in the context of remote collaboration and to elicit factors influencing collaboration. IM2 technologies were used in this context both to facilitate user evaluations (for instance with automatic analysis of eye-tracking) and to setup real-time technologies to recognize and visualize emotional states of participants as well as to interact more naturally with the board.

Two large displays serve as interactive supports for the EmotiBoard application, where a continuous feedback of team-members' emotional states is included in a video conference setup. The size of the emoticon representing each participant depends on arousal (from very passive to very active) and the direction of the smile indicates the valence: from very negative to very positive.

At the time of writing, several large experiments are being performed using the CBoard technology as the main framework. The results of the first study (EmotiBoard 1) have shown the usefulness of the mood feedback tool in remote settings, because it helped meeting participants to better understand other team members moods and improved other outcome measures of team work. A second experiment (EmotiBoard 2) aimed at estimating how emotion perception and gaze of a person might be influenced by emo-

tional feedback. An eye-tracker developed during IM2 was used for this purpose. Finally the latest study (EmotiBoard 3) aimed at recording multimodal data from human affective and social interactions in a context of computer-mediated collaboration work (Ringeval et al., 2013). The corpus of data collected through EmotiBoard 3, called 'RECOLA' (Ringeval et al., 2013), is currently used to develop a real-time emotion recognizer which will be used to automatically assess team members' mood, based on speech prosody and physiological data (skin conductance, heart rate variability).

5.6 Conclusion

HCI activities in Fribourg have grown throughout the IM2 project. In the first phase, multimodal meeting data was acquired, offline multimodal analyzers were then developed and used to create offline meeting browsers that support navigation and search in multimedia recordings of meetings. With the maturity of multimodal technologies, HCI activities shifted toward online meeting assistants, using real-time multimodal analyzers to support teamwork in collocation or in remote settings.

Bibliography

Behera, A., Lalanne, D., and Ingold, R. (2004). Looking at projected documents: Event detection & document identification. In *Multimedia and Expo, 2004. ICME'04. 2004 IEEE International Conference on*, volume 3, pages 2127–2130. IEEE.

Behera, A., Lalanne, D., and Ingold, R. (2008). Docmir: An automatic document-based indexing system for meeting retrieval. *Multimedia Tools and Applications*, 37(2):135–167.

Bertini, E. and Lalanne, D. (2007). Total recall survey report. Technical report, University of Fribourg.

Dumas, B., Lalanne, D., Guinard, D., Koenig, R., and Ingold, R. (2008). Strengths and weaknesses of software architectures for the rapid creation of tangible and multimodal interfaces. In *Proceedings of the 2nd international conference on Tangible and embedded interaction*, TEI '08, pages 47–54, New York, NY, USA. ACM.

Dumas, B., Lalanne, D., and Ingold, R. (2009). Hephaistk: a toolkit for rapid prototyping of multimodal interfaces. In *Proceedings of the 2009 international conference on Multimodal interfaces*, ICMI-MLMI '09, pages 231–232, New York, NY, USA. ACM.

Dumas, B., Lalanne, D., and Ingold, R. (2010). Description languages for multimodal interaction: a set of guidelines and its illustration with smuiml. *Journal on Multimodal User Interfaces*, 3(3):237–247.

Evéquoz, F., Thomet, J., and Lalanne, D. (2010). Gérer son information personnelle au moyen de la navigation par facettes. In *IHM 2010, Conference Internationale Francophone sur I'Interaction Homme-Machine*, pages 41–48. ACM.

Ishii, H. and Kobayashi, M. (1992). Clearboard: a seamless medium for shared drawing and conversation with eye contact. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '92, pages 525–532, New York, NY, USA. ACM.

Lalanne, D., Evequoz, F., Rigamonti, M., Dumas, B., and Ingold, R. (2008). An ego-centric and tangible approach to meeting indexing and browsing. *Machine Learning for Multimodal Interaction*, pages 84–95.

Lalanne, D., Ingold, R., Von Rotz, D., Behera, A., Mekhaldi, D., and Popescu-Belis, A. (2005). Using static documents as structured and thematic interfaces to multimedia meeting archives. *Machine Learning for Multimodal Interaction*, pages 87–100.

Lalanne, D., Mekhaldi, D., and Ingold, R. (2003a). Talking about documents: revealing a missing link to multimedia meeting archives. In *Electronic Imaging 2004*, pages 82–91. International Society for Optics and Photonics.

Lalanne, D., Sire, S., Ingold, R., Behera, A., Mekhaldi, D., and Rotz, D. (2003b). A research agenda for assessing the utility of document annotations in multimedia databases of meeting recordings. *Proceedings of 3rd international workshop on multimedia data and document engineering, Berlin, Germany.*

Rigamonti, M., Lalanne, D., and Ingold, R. (2007). Faericworld: browsing multimedia events through static documents and links. *Human-Computer Interaction–INTERACT 2007*, pages 102–115.

Ringeval, F., Sonderegger, A., Sauer, J., and Lalanne, D. (2013). Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *EmoSPACE 2013, 10th IEEE Conference on Automatic Face and Gesture Recognition (FG 2013)*. IEEE.

Schwaller, M., Brunner, S., and Lalanne, D. (2013). Two handed mid-air gestural HCI: Point + command. In *Proceedings of HCI 2013 (15th International Conference on Human-Computer Interaction).*

Schwaller, M. and Lalanne, D. (2013). Pointing in the air: Measuring the effect of hand selection strategies on performance and effort. In *SouthCHI 2013, International Conference on Human Factors in Computing and Informatics*. ACM.

Sonderegger, A., Lalanne, D., Ringeval, F., and Sauer, J. (2013). Computer-supported work in partially distributed and co-located teams: the influence of mood feedback. In *Proceedings of INTERACT 201 (14th IFIP TC13 Conference on Human-Computer Interaction)*, Cape Town, South Africa. IFIP.