Wearable Sensing to Annotate Meeting Recordings

Nicky Kern, Bernt Schiele
Perceptual Computing and Computer Vision
ETH Zurich
{kern,schiele}@inf.ethz.ch

Holger Junker, Paul Lukowicz, Gerhard Tröster
Wearable Computing Lab
ETH Zurich
{junker,lukowicz,troester}@ife.ee.ethz.ch

Abstract

We propose to use wearable computers and sensor systems to generate personal contextual annotations in audio visual recordings of meetings. In this paper we argue that such annotations are essential and effective to allow retrieval of relevant information from large audio-visual databases. The paper proposes several useful annotations that can be derived from cheap and unobtrusive sensors. It also describes a hardware platform designed to implement this concept and presents first experimental results.

1. Introduction

Interestingly, about 500 Tera Bytes of storage are sufficient to record all audio-visual information a person perceives during an entire lifespan[1]. This amount of storage will be available even for an average person in the not so distant future. A wearable recording and computing device therefore might be used to 'remember' any talk, any discussion, or any environment you saw.

Today however, the usefulness of such data is rather limited by the lack of adequate methods for accessing and indexing large audio-visual databases. Interestingly, humans not only remember events and retrieve memories based on information such as time, date, location, or content of a discussion. But humans use additional and personal experience and contextual information to remember and retrieve memories. In this paper we propose to use wearable sensors in order to enhance the recorded data to allow associative access.

In that context, wearable computers are particularly interesting since they allow a truly personal audio-visual record of the environment of a person. Using a hat- or glass-mounted camera and microphones attached to the chest or shoulders of the person enable a recording from a first-person perspective. Additionally, wearable sensors such as accelerometers and biometric sensors can enhance the recording with additional, very personal information. That sensor information can be used to annotate and structure the data stream for later access.

Obviously, automatically annotating and structuring the entire life-record of a person is an extremely ambitious and probably too general problem. Therefore, this paper deals with a more specific problem, namely the annotation of meetings, which, in itself, presents a very diverse setting. Most of us have many, maybe too many meetings every week. Using a wearable to record such meetings won't make the meetings themselves more efficient. However, it may allow the user to recall who he encountered, who discussed, who agreed or disagreed, and which arguments each participant made. It may also make it easier to reconstruct which, why, and how a decision was taken.

Meetings may take place in a room instrumented with dedicated hardware. More generally however, meetings also take place outdoors or in a mobile setting. Further, important discussions may take place during the break or on the corridor. Wearable computers, which stay with the person all the time, are particularly well suited for this more general meeting scenario.

The contributions of this paper are firstly the discussion of possible annotations of a meeting recording so as to facilitate associative retrieval (section 3). Secondly we investigate the use of audio data (section 4) and accelerometer data (section 5) to automatically generate interesting annotations. The feasibility of such annotations is shown experimentally (section 6). Thirdly, we have designed and implemented a distributed accelerometer network so as to extract information about the user's movements and postures (section 5). Finally section 7 discusses the approach and gives a brief outlook.

2. Related Work

The idea of computer-based support for human memory and retrieval is not new. Lamming and Flynn for example point out the importance of context as a retrieval key [11] but only used cues like location, phone calls, and interac-

[1] assuming a lifespan of 100 years, 24h recording per day, and 10 MB per min recording results in approximately 500 TB

Proceedings of the 6th International Symposium on Wearable Computers (ISWC'02)
0-7695-1816-8/02 $17.00 © 2002 IEEE

tion between different PDAs. The conference assistant [5] supports the organization of a conference visit, annotation of talks and discussions, and retrieval of information after the visit. Again, the cooperation and communication between different wearables and the environment is an essential part of the system. Rhodes proposed the text-based remembrance agent [17] to help people to retrieve notes they previously made on their computer.

For speech recognition the automatic speech transcription of meetings is an extremely challenging task due to overlapping and spontaneous speech, large vocabularies, and difficult background noise [1, 2]. Often, multiple microphones are used such as close-talking, table microphones, and microphone arrays. The SpeechCorder project [8] for example aims to retrieve information from roughly transcribed speech recorded during a meeting. Summarization is another topic, which is currently under investigation in speech recognition [22] as well as video processing. We strongly believe, however, that summarization is not enough to allow effective and in particular associative access to the recorded data (see section 3). It should be noted that those methods are complementary to the proposed approach and should be integrated eventually.

Richter and Le [10] propose a device which will use predefined commands to record conversations and take low-resolution photos. At the university of Tokyo [20] researchers investigate the possibilities to record subjective experience by recording audio, video, as well as heartbeat or skin conductance so as to recall one's experience from various aspects. StartleCam [6] is a wearable device which tries to mimic the wearer's selective memory. The WearCam idea of Mann [13] is also related to the idea of constantly recording one's visual environment.

3. Annotating a Meeting Recording

The ultimate goal of our system is to facilitate efficient indexing and retrieval of the audio-visual data recorded during meetings. The general idea is to support or 'extend' the human memory by means of a wearable computer. In this context it is interesting to note that the human brain heavily uses associative memory access. Humans not only retrieve memories about an event based on time, date, name of a person, or other precise attributes. Humans also remember and retrieve things by context information such as the weather, what happened before the meeting, who else was present, or if people were agitated during a discussion. Therefore, this paper proposes to enhance and annotate meeting recordings by context information in order enable associative retrieval of information.

Interesting Annotations for Meetings. It is standard to generate summaries of meetings either in written or digital form. Those summaries however are not close enough to how humans retrieve information from their memories. Looking particularly at the envisioned meeting scenario we have identified four classes of relevant annotations. Those are different meeting phases, flow of discussion, user activity and reactions, and interactions between the participants. The meeting phase includes the time of presentations, breaks, and when somebody is coming or leaving during the meeting. The flow of discussion annotations attach speaker identity and changes to the audio stream, and indicate the level of intensity of discussion. It can also help to differentiate single person presentations, interactive questions and answers, and heated debate. User activity and reactions indicate user's level of interest, focus of attention, and agreement or disagreement with particular issues and comments. By tracking the interaction of the user with other participants personal discussions can be differentiated from general discussions.

Using Wearable Sensors to Annotate Meetings. Using wearable sensors opens the opportunity to add relevant annotations from all four classes. In this paper we concentrate on using audio to identify speakers and speaker changes. We also propose a distributed accelerometer network to identify the user's reactions and activities such as walking, standing, and sitting. Also hand movements, such as shaking hands, have a clear social meaning which can be used to detect interactions between participants. Additional information can be derived by correlating those two channels. For example a speaker change together with head turning indicates a shift in the focus of attention. Similarly, a coincidence of a speaker change with head nodding or shaking indicate agreement or disagreement with what is just said. The advantage of our approach lies in the fact that complex information can be derived with small and unobtrusive low-power devices.

Collaborating Wearables. An interesting aspect of a meeting is also how much the wearable devices of the individual participants collaborate. In one extreme, a user might only have access to the data recorded and annotated by his own wearable. In that case the computing, sensing, and consequently power requirements are quite high. In the other extreme all participants possess a wearable and share at least some information. In that case the decision who is present or who is speaking is a much simpler task since it is relatively simple to detect when the user of a wearable device is talking (see section 4).

When wearables share information there is an obvious question of trust. Sharing information among trusted wearables however, is a rich source for additional information such as the slides of a presenter, the transcribed speech from the presenter, etc. In the near future it is most likely that there will be a mixture of the two extremes so that the wearable should be designed to adapt to both scenarios in an appropriate fashion.

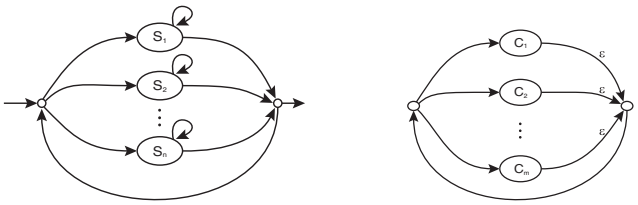

Figure 1. Identification Model for a Single Speaker (left) and Segmentation Network (right)

4. Audio Context: Speaker Segmentation

As pointed out earlier, a text-only transcript of a meeting may not be sufficient for associative retrieval. The flow of discussion indicates its intensity and allows to tell presentation from discussion. The transcription of speech in our setting (multiple speakers, spontaneous or even simultaneous speech, large vocabularies, one microphone worn by only one user) is still a very hard problem, despite recent advances in speech recognition technology. In this paper we therefore concentrate on finding speaker changes and identities.

In our setting we have the microphone attached to the user of the wearable computer. We can exploit this fact and detect, whether the user is speaking or someone else by mainly looking at the energy of the recorded audio signal. In section 6 we show that this method effectively improves speaker recognition.

Related Work The problems of speaker recognition and segmentation have been addressed by two fundamentally different approaches, either based on clustering or using Hidden Markov Models (HMMs). [9] uses agglomerative clustering and [4] distance–based segmentation to detect the most likely speaker changes and the Bayesian Information Criterion to discard invalid changes. These methods combine reliability with reasonable cost, but at the price of non real–time performance.

The identification of speakers can also be done using HMMs, similar to speech recognition. The topology of the HMM is often distorted, since the identification of a speaker is not based on any actual utterance and is therefore not really time-dependent. [3] uses a single state with 30 Gaussians and [23] uses 32 parallel states for a single speaker HMM. In the case of multiple speakers, each has his own HMM and the Viterbi–algorithm allows speaker identification in real-time. Since the speaker models need training these approaches require audio data from each speaker in advance.

Speaker Segmentation using Speaker Identification The identification of speakers can also be used for speaker segmentation [23]. In this paper we assume to know the participating speakers. This allows to take the approach of Kimber and Wilcox [9, 23], which performs in real–time and produces a segmentation, as well as an identification of the speakers. The HMMs are trained on 12 mel–cepstral coefficients over 20ms, non–overlapping windows. The topology for one speaker HMM is depicted on the left of figure 1. In our system we use $n = 32$ states.

The speaker-HMMs are combined into the segmentation network (see figure 1 on the right). The speaker models are trained separately and the segmentation network is used for recognition. The Viterbi algorithm finds the optimal path through the HMM from one speaker to another and therefore detects speaker changes. ϵ is selected empirically as to discourage short speaker changes due to isolated speech vectors. For training and recognition the HMM Toolkit (HTK) [25] is used.

5 Accelerometric Analysis

Our approach to monitoring user's activity and reactions is based on the importance of human posture and gesture. Most situations and activities can be characterized by a specific body position and/or limbs motion pattern. A person presenting a talk is likely to be standing up, possibly slowly walking back and forth, moving his arms gesticulating. By contrast somebody eating lunch would be sitting, predominantly looking down and periodically lifting a sandwich from the plate to his mouth.

To detect postures and body parts motions we rely on a network of 3 axis accelerometers distributed over the user's body. Each accelerometer provides us with information about the orientation and movement of the corresponding body part. The advantage of this approach lies in the small size and energy efficiency of acceleration sensors. In addition with a modest amount of preprocessing only minimal communication bandwidth is needed to read the relevant information. Thus such a network can be unobtrusively integrated in an arbitrary outfit. While much work has been devoted to accelerometric context detection [24, 19, 18, 7] with the exception of [15] which relied on several sensors distributed over the hand the use of a distributed network has not been studied.

A detailed description of our approach to movement and posture recognition is beyond the scope of this paper. Instead the following section provides an overview of our hardware and the principles used for the classification of postures and motions from the sensor data.

5.1 Hardware

Each sensor node consists of two dual-axis accelerometers from Analog Devices ADXL202E (combined allow

measurement of linear acceleration in the 3D-space) and the MSP430F149 low power 16-Bit mixed signal microprocessor (MPU) from Texas Instruments running at 6 MHz maximum clock speed. The MPU reads out the sensor signals and handles the communication between modules through dedicated I/O pins. Since our setup relies on the analog outputs of the accelerometers three second order Sallen-Key low pass filters are also used. Optionally, a single-axis gyroscope can be mounted on the board.

Although the modules are miniaturized (28x34 mm) even smaller devices are desirable at some locations such as the head or fingers. Therefore, the modules are partitioned and consist of two parts each: the main part with the microcontroller, the filters and amplifiers and a sub-part with the sensors only which can either be mounted directly on the main unit or connected by wires. Figure 2 illustrates the assembly of a node and its block diagram. All modules are

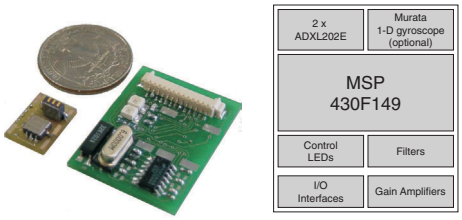

Figure 2. Left: Main-board, sub-board of sensor node and 1/4 dollar coin, right: corresponding block diagram of sensor node

powered from a single central power supply consisting of a step down regulator and a small mobile phone or camcorder battery. The power supply unit is part of a central control module. This module is based on a GPS receiver (u-blox GPS-MS1E) with an Hitachi SH-1 processor. Apart from serving as a central control unit of the network and a serial I/O-interface to a computer this module can also provide absolute location information.

Network The communication within our sensor network is based on a 3-wire bus. Two wires implement the communication between the nodes using the I2C-bus and the third is used to synchronize all sensors. The sensor platform is partitioned into subnetworks reflecting anatomical relations between the body parts. For example all sensors on the upper and lower leg and possibly foot constitute a single subnetwork. Within such a subnetwork a particular sensor module acts as a master which handles communication with the other sensor nodes (slaves) within the channel. All the masters of the subnetworks are slaves to an upper network layer in which the central module with the Hitachi SH-1 processor serves as master. This two-layered hierarchical network architecture allows to optimize communication in terms of overall network load, since a considerable amount of pre-processing can be done locally within a subnetwork. Thus most of the communication between layers consists of high level features represented by a few numbers rather the large amounts of raw data. As a second advantage this distributed data processing approach allows to reduce computational load of the central master node in the first layer. Figure 3 shows the hierarchical network structure with possible sensor locations. Sensors labeled 1 to 7 are used for our experiments.

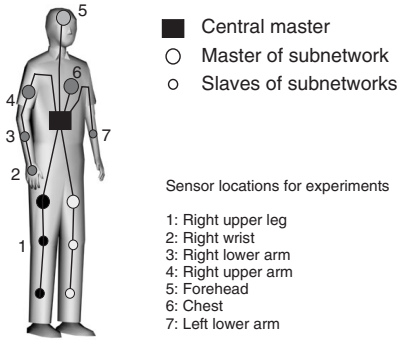

Figure 3. Hierarchical network with observation channels.

5.2 Recognition Methodology

Most approaches to activity recognition using wearable accelerometers rely on parameters that are more or less directly derived from the raw accelerometer data [14, 16]. Often automatic clustering algorithms are used to derive features used for classification [21]. By contrast our approach emphasizes physical models of human motion and the decomposition of complex motions into elementary postures and gestures of each body part. This is possible since the distributed, multisensor accelerometer network provides detailed information on every relevant body part. The following summarizes the underlying physical model and sketches our feature extraction procedure. The features provide an excellent separation between the relevant movements and postures so that in many cases simple, easy to derive decision trees are sufficient for reliable recognition. For complex, more ambiguous motions statistical pattern classification algorithms like HMM and neural networks can be applied.

Physical Model The readings of the accelerometers consist of three components: gravity, change of speed and centripetal forces. The gravity component can be used to determine the orientation of a sensor (and with it the corresponding body part) in the vertical plane. The change of speed part of the reading is the basis for motion analysis. The centripetal force results from rotational motions of the limbs with respect to the corresponding joints. In most cases this component of the reading is not used.

In general for any single sensor the above components are non separable. However in most cases approximate separation can be achieved either by frequency separation of the acceleration signal or by appropriate sensor placement. The gravity contribution showing the orientation of the body parts is predominantly contained in the low frequency part of the sensor signal often remaining unchanged for seconds. By contrast any strong acceleration of the body parts is likely to last no longer then a few tenths of a second. In terms of sensor placement we utilize the fact that for sensors placed close to a joint of an arm or leg the readings will be dominated by the gravity contribution.

Feature Extraction The features used for the recognition of user activity are the approximate orientation and motion patterns of the relevant body parts. They are derived from the sensor reading in several steps.

First for each sensor the output is filtered and separated into low ($< 2Hz$) and 'high' ($> 2Hz$) frequency components. In the next steps the readings from the sensors on the torso are propagated down in the network hierarchy allowing the bottom sensors (located at the limbs and the head) to compute their motion relative to the torso. To this end the low frequency component is used to compute the orientation of the corresponding body parts in the vertical plane. The high frequency component is analyzed for motion artifacts. After this initial evaluation phase the results are propagated up through the network to the control node. Using the knowledge about the anatomical constraints of the human body together with the previous position and motion state the controller combines the data from the individual sensors to an overall approximate description of the posture and motion pattern. This is forwarded to the main computer unit that performs the actual classification of the situation and user activity.

6 Experiments

We acquired audio and acceleration data to validate our idea and evaluate the methods described in sections 4 and 5.

6.1. Speaker Recognition

To validate the audio retrieval algorithms we conducted several experiments with increasing difficulty. We used either desktop microphones or a wearable clip–on microphone (Sony ECM TS-125).

User vs. the World Since we are considering *personal* annotations, the information whether the user or somebody else is speaking is highly interesting. The wearable microphone allows to use a simple energy-thresholding algorithm for distinguishing between 'Me' and 'Not-Me', which proves to be very successful. The recording of a 41 minute meeting has been labelled and tested. Using energy thresholding on 0.1 second intervals we achieve an error rate of 1.2 %, which will be sufficient for many retrieval applications.

Sequence	Recognition Error
Reading One	1.6 %
Reading Two	0.0 %
'Clean' Dialog	11.5 %
'Normal' Dialog	11.4 %
Clip–on Mic 1	5.9 %
Clip–on Mic 2	9.2 %

Table 1. Speaker Recognition Experiments

Controlled Setting A set of six recordings of increasing difficulty is used to evaluate the performance of the speaker recognition algorithm. See table 1 for a summary of the retrieval results. One male and one female speaker recorded them. Two sequences of 4 minutes each recorded with desktop microphones are used as training data. The first two lines of table 1 show recognition results on these training sets, which are obviously good.

The same two speakers were recorded, using the same desktop microphones, involved in a dialog. The first lasts about 10 min and is 'clean', i.e. contains distinct pauses between the speakers, no simultaneous speech and little laughter. The second (4 min) is a 'normal' unconstrained dialog. Considering, that the problem is harder, the lower performance is quite natural (see the 3rd and 4th row of table 1).

Finally the same speakers recorded two dialogs where one of them wore the wearable microphone. Since the microphones changed, both sequences had to serve both as test and as training sequences. The last two rows of table 1 show the corresponding results.

Wearable Meeting In order to validate our approach we recorded an entire meeting of 31 minutes. One of the four participants (the 'user' in the following) was equipped with a clip-on microphone. Using HMMs as described in section 4 results in a recognition error of 18 %. As described above, the distinction between the wearer of the microphone and 'the others' can reliably be made using an energy threshold. Using an energy threshold and only applying the HMM-algorithm for the remaining part of the audio sequence results in an overall decrease of the error rate down to 9%. Given, that the user was relatively far apart from the others and constantly moving, these results are very promising. Again we should point out that the raw error rate is not a very good measure to evaluate the usefulness for retrieval. We do believe that the obtained recognition rates are sufficient for retrieval.

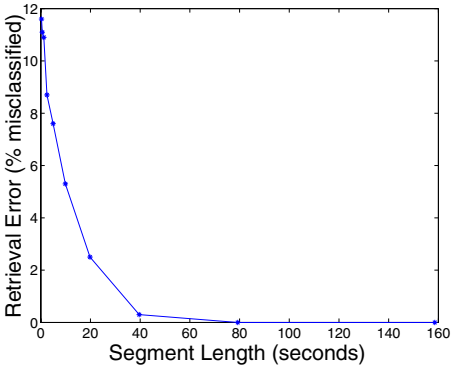

Figure 4. Retrieval Error on different segment lengths

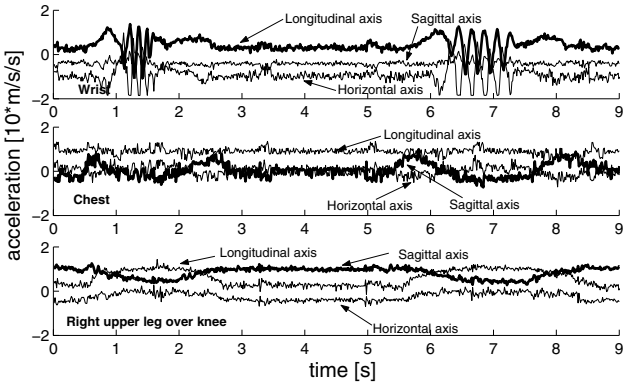

Figure 5. Raw accelerometer data for the 'stand up, handshake, sitdown' sequence.

6.2. Retrieval

Analyzing the results from the previous section reveals that most speakers are identified correctly but that sometimes the time of the speaker change is not very accurate. A substantial part of the error rates of table 1 are due to this. In many retrieval scenarios however we are more interested who is speaking i.e. which speakers participate in a discussion rather than to know exactly at what time somebody speaks.

Motivated by this fact we propose a scheme that allows trading error rate of speaker identification against its time accuracy. More specifically we start looking for a specific constellation of speakers using long segments and shorten them incrementally. In order to avoid looking into wrong segments and missing correct segments, we need a low error rate for long segments. At the same time however we can allow for a rather coarse time accuracy. Once we are looking into shorter segments, we are interested in listening to actual utterances of specific people therefore relying on a higher accuracy at what time somebody speaks or not.

Figure 4 shows a plot of the error rate of the speaker identification versus the length of the time segments used for recognition. The figure corresponds to the result for the 'clean' dialog, which has the highest error rate in our experiments (see table 1). It can easily be seen, that the error rate drops quickly under 1% for segments of only 40 seconds. We can hence decrease the error significantly by enlarging the segment in question. The price is obviously that the time accuracy of the speaker is decreased. This result supports the validity of the above mentioned retrieval scheme.

6.3 Accelerometric Activity Detection

To verify that relevant indexing cues can be detected using our distributed accelerometer network we have looked at different event sequences typical for a meeting scenario. In this section we present the results from three selected sequences.

Measurement Setup All measurements have been made with the accelerometer axes aligned with the three principal body axes: Longitudinal axis (vertical axis through the body in the upright position), Horizontal axis (perpendicular to longitudinal axis and runs from left to right), and Sagittal axis (axis that runs from front to back). The assignment of the body axes to the axes of the accelerometers have been made for the anatomical position that is when a person is standing upright with the head, eyes and toes pointing forward, feet together with arms by the side. The palms of the hands are also pointing forward.

Greeting a New Participant In the first experiment the subject is sitting on a chair, hands on the table. From this position he stands up, shakes hands with a newly arrived meeting participant and sits down again. In this context the acceleration of the vertical and sagittal axes of the upper legs and the chest as well as the vertical and longitudinal axes of the the right wrist are of particular interest. The raw sensor data from the above channels is shown in figure 5. The left leg is omitted since it is virtually identical to the right leg signal. The sequence contains two 'standup and shake hands' events. For the first one the features extracted using our signal processing algorithm are shown in the six diagrams in figure 6. For each channel the left diagram shows the filtered low frequency component which is proportional to the vertical orientation of the corresponding body part. The leg channel shows the transition of the upper leg from horizontal (sitting) to vertical (standing) position and back. The chest channel shows a forward leaning motion of the torso characteristic for sitting down and

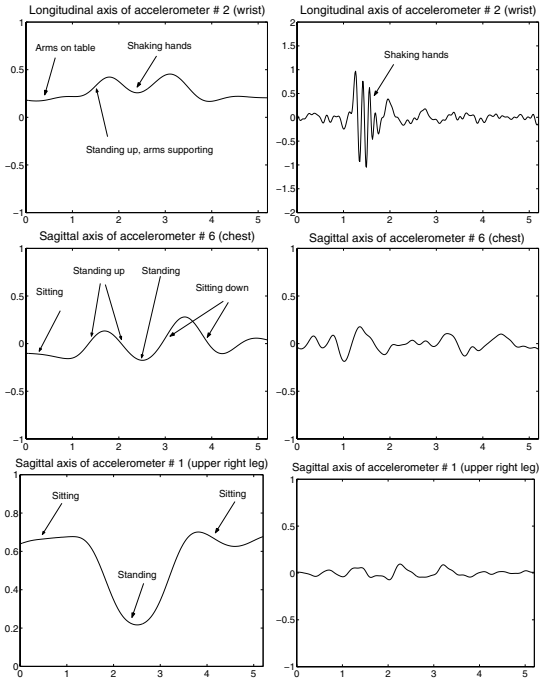

Figure 6. Low (left) and high (right) frequency components from the relevant channels in the 'stand up, handshake, sitdown' sequence.

standing up. Taken together the two channels provide a reliable indication of the user standing up and sitting down. The low frequency component of the hand channels shows the user's arm falling into a vertical position as he stands up followed by a more horizontal orientation during the actual handshake and another vertical horizontal transition as the user sits down. Of the high frequency components shown on the far right only the hand channel shows a significant change of amplitude which can easily be identified as the vertical handshake motion.

Head Movements during Discussion The second scenario concentrates on head movements. As described in section 3 head motion is an important indicator of the user's reaction to events and his focus of interest. In particular, spontaneous nodding and head shaking is a good sign of agreement or disagreement with a particular issue or comment. Using the head channel from a sensor mounted on the forehead we have looked at nodding, head shaking, and head turning events. Figure 7 shows the features extracted from the longitunal channel of the sensor for a typical nodding event. The low frequency channel is essentially constant, since the amplitude of the nodding motion is too small and the frequency too high. The high frequency channel contains a sequence of 'bumps' corresponding to the individual nods, which can easily be identified with simple signal analysis techniques. Looking at the other sensors in the network we can be sure that the bumps result from head motions.

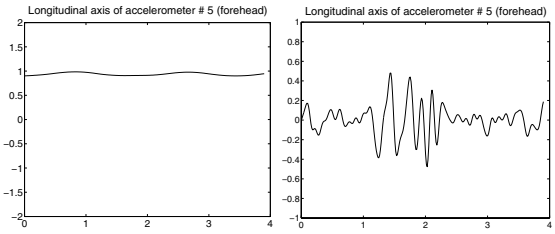

Figure 7. Low (left) and high (right) frequency components from the head channel for a typical head nodding motion.

Figure 8. Examples of different postures that can be recognized by our system.

Thus head nodding can be reliably detected using the features extracted with our approach. A corresponding pattern can be found for head shaking and head turns in the signal of the horizontal or sagittal axis.

Complex Gestures and Posture Simple gestures like nodding, shaking ones head or raising shoulders constitute just a small and simple subset of human body language. As shown in [12] the human body language is rich and complex with a variety of postures and gestures that can be used to deduce peoples attitudes and emotions. Many elements of body language involve facial expressions and subtle gesture nuances that are beyond the scope of our network of accelerometers recognition approach. However, there are a number of other potentially interesting expressions that we can reliably recognize. The third scenario exemplifies this by showing how sensors placed on the arm, the chest the head and the upper leg can be used to distinguish two different postures. To this end we have considered the postures shown in figure 8, which could be interpreted as 'concentrated' (left) and 'laid back' (right). Figure 9 shows the appropriately processed low frequency components of the sensors axes most affected by the change of orientation of the corresponding body parts.

7 Discussion and Outlook

Using wearable computers to record and annotate meetings is an interesting application of wearable computing with great potential. In this paper we propose to automat-

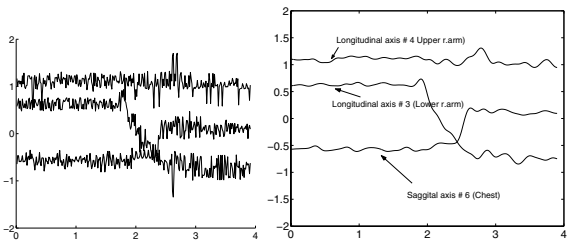

Figure 9. Raw data (left) and the preprocessed low frequency component of the relevant channels for a change between the postures shown in figure 8

ically generate *personal* annotations and *personalized* contextual information from wearable sensors. In particular, we show the feasibility to segment an audio stream into sequences assigned to different speakers. This allows us to reconstruct the speaker flow. We also introduce a distributed accelerometer network to analyze the body movements and postures of the user. Those cues are a first step to enable and facilitate efficient indexing and retrieval of audio-visual meeting recordings.

Obviously, many issues remain to be addressed. As far as complex postures and gestures are concerned, recognition is just part of the problem. In many cases the correct interpretation of complex body language poses a great challenge. Obviously, there are a number of other sensor types that could potentially be useful to annotate meeting recordings. In particular, physiological data about the user may prove valuable to asses the user's reaction to events and the importance of issues. Wearable cameras are also a rich source to recognize faces, objects, and situations. Since the proposed system should be used by a human an appropriate interface to the indexing methodology has to be developed and evaluated. Also the usefulness of the proposed cues as well as additional cues should be evaluated on a large number of real-life meetings. We would also like to point out the important issue of privacy which is a legal and even more so an ethical issue linked to audio-visual recordings [8, 10].

References

[1] ICSI Berkeley, The Meeting Recorder Project at ICSI. http://www.icsi.berkeley.edu/Speech/mr/.

[2] NIST Automatic Meeting Transcription Project. http://www.itl.nist.gov/iad/894.01/.

[3] T. Choudhury, B. Clarkson, T. Jebara, and A. Pentland. Multimodal person recognition using unconstrained audio and video. In *Audio- and Video-based Biometric Person Authentication*, 1999.

[4] P. Delacourt and C. Wellekens. Detection of speaker changes in an audio document. In *Eurospeech*, 1999.

[5] A. Dey, D. Salber, G. Abowd, and M. Futakawa. The conference assistant: Combining context-awareness with wearable computing. In *ISWC*, pages 21–28, 1999.

[6] J. Healey and R. Picard. Startlecam: A cybernetic wearable camera. In *ISWC*, pages 42–49, 1998.

[7] K. Hinckley, J. Pierce, M. Sinclair, and E. Horvitz. Sensing techniques for mobile interaction. In *User Interface Software and Technology*, pages 91–100, 2000.

[8] A. Janin and N. Morgan. Speechcorder, the portable meeting recorder. In *Workshop on Hands-Free Speech Communication*, 2001.

[9] D. Kimber and L. Wilcox. Acoustic segmentation for audio browsers. In *Proc. Interface Conference*, 1996.

[10] T. Kontzer. Recording your life. http://www.informationweek.com, Dec, 18 2001.

[11] M. Lamming and M. Flynn. Forget-me-not: intimate computing in support of human memory. In *FRIENDS21*, pages 125–128, 1994.

[12] D. Leathers. *Successful Nonverbal Communication: Principles and Applications*. Allyn & Bacon, 1997.

[13] S. Mann. Smart clothing: The wearable computer and wearcam. *Personal Technologies*, 1(1), 1997.

[14] J. Mantyjarvi, J. Himberg, and T. Seppanen. Recognizing human motion with multiple acceleration sensors. In *Systems, Man and Cybernetics*, pages 747–752, 2001.

[15] J. Perng, B. Fisher, S. Hollar, and K. Pister. Acceleration sensing glove (asg). In *ISWC*, pages 178–180, 1999.

[16] C. Randell and H. Muller. Context awareness by analysing accelerometer data. In *ISWC*, pages 175–176, 2000.

[17] B. Rhodes. The wearable remembrance agent: A system for augmented memory. In *ISWC*, pages 123–128, 1997.

[18] A. Schmidt, K. Aidoo, A. Takaluoma, U. Tuomela, K. V. Laerhoven, and W. V. de Velde. Advanced interaction in context. In *HUC*, pages 89–101, 1999.

[19] M. Sekine, T. Tamura, T. Fujimoto, and Y. Fukui. Classification of walking pattern using acceleration waveform in elderly people. In *Engineering in Medicine and Biology Society*, volume 2, pages 1356–1359, 2000.

[20] R. Ueoka, M. Hirose, K. Hirota, A. Hiyama, and A. Yamamura. Study of experience recording and recalling for wearable computer. *Correspondences on Human Interface*, 3(1):13–16, 2001.02.

[21] K. van Laerhoven, K.Aido, and S. Lowette. Real-time analysis of data from many sensors with neural networks. In *ISWC*, pages 115–123, 2001.

[22] A. Waibel, M. Bett, and M. Finke. Meeting browser: Tracking and summarizing meetings. In *Proceedings of the DARPA Broadcast News Workshop*, 1998.

[23] L. Wilcox, D. Kimber, and F. Chen. Audio indexing using speaker identification. In *Eurospeech*, pages 25–28, 1994.

[24] Y. Yoshida, Y. Yonezawa, K. Sata, I. Ninomiya, and W. Caldwell. A wearable posture, behavior and activity recording system. In *Engineering in Medicine and Biology Soc.*, volume 2, page 1278, 2000.

[25] S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. *The HTK Book*. Cambridge, 1995.