Towards Multi-Modal Context Recognition for Hearing Instruments

Bernd Tessendorf, Andreas Bulling, Daniel Roggen, Thomas Stiefmeier, Gerhard Tröster
Wearable Computing Lab., ETH Zurich
Gloriastr. 35, 8092 Zurich, Switzerland
Email: {lastname}@ife.ee.ethz.ch

Manuela Feilner, Peter Derleth
Phonak AG
Laubisrütistrasse 28, 8712 Stäfa, Switzerland
Email: {firstname.lastname}@phonak.com

Abstract

Current hearing instruments (HI) only rely on auditory scene analysis to adapt to the situation of the user. It is for this reason that these systems are limited in the number and type of situations they can detect. We investigate how context information derived from eye and head movements can be used to resolve such situations. We focus on two example problems that are challenging for current HIs: To distinguish concentrated from interaction, and to detect whether a person is walking alone or walking while having a conversation. We collect an eleven participant (6 male, 5 female, age 24–59) dataset that covers different typical office activities. Using person-independent training and isolated recognition we achieve an average precision of 71.7% (recall: 70.1%) for recognising concentrated work and 57.2% precision (recall: 81.3%) for detecting walking while conversing.

1 Introduction

Over the last decade significant progress has been achieved in hearing instrument (HI) technology. Today's HIs are wearable computers that comprise a variety of components such as microphones, signal processing, or wireless links. Despite these advances, HI users still experience difficulties in certain situations [3]. Auditory scene analysis (ASA) allows a HI to automatically adapt to the user's current hearing needs, e.g. by activating noise cancelation or beamforming. ASA is based on the assumption that the acoustic environment correlates with the user's hearing need in an unambiguous manner. This assumption may fail in many daily life situations in which sound alone does not provide sufficient information to derive the user's current hearing need. A multi-modal approach can address this shortcomings by collecting complementary information through different sources of context information.

We propose multi-modal context recognition as a novel means for automatic HI adaptation. The primary aim of current research is to assess the feasibility of using eye and head movements to disambiguate two example situations that are challenging for state-of-the-art HIs.

2 Experiment

To this end, we conducted an experiment that was designed to (1) record eye and head movements in a real-world office setting, and (2) to evaluate how information derived from these modalities can be used to detect concentrated work and walking while having a conversation.

Experimental Procedure The first part of the experiment involved participants performing a continuous sequence of four typical office activities: reading a book, writing on a sheet of paper, typing a text on the computer, and having a conversation. To ease the video-based offline labeling we partitioned this part of the experiment into one minute time slots. In the first minute, the participant worked concentratedly on his task. In the second minute, the participant tried to stay concentrated while the office colleague was talking to a disturber. In the third minute, the participant was interrupted and engaged in a discussion with the disturber. In the fourth minute, the disturber talked to the colleague again. In the fifth minute, the participant and the colleague had a conversation. This procedure was repeated eight times, once for each of the four office activities and with the participants being seated and standing. The second part of the experiment was designed to contain different walking situations. We included sequences of the participant walking alone, walking while talking to a conversation partner, and walking with two other people following and talking.

Sound Two state-of-the-art HIs (Phonak Exelia Art BTE) were connected to an audio recorder capturing the raw audio data on four channels with 24 bit at 48 kHz. In addition, participants wore a throat microphone.

Physical Activity For recording physical activity we used nine sensor nodes from xSens Technologies containing 3-axis accelerometers, magnetic field sensors, and gyroscopes. The sensors were placed at the wrists, lower arms, upper arms, back, head, and left upper leg. The sampling rate was set to 32 Hz.

Eye Movements For recording eye movements we used the Mobi system from Twente Medical Systems International to capture a four-channel EOG with a sampling rate of 128 Hz. EOG signals were picked up using an array of five electrodes positioned around the right eye. The horizontal

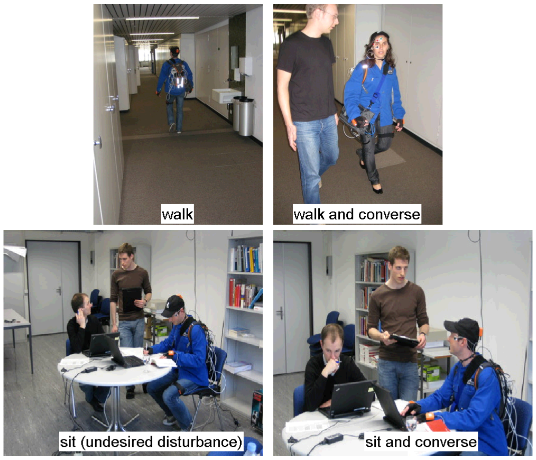

Figure 1: Activity classes in the experiment: Walking alone and while conversing as well as focused work activity classes and conversation.

signal was collected using two electrodes on the edges of both eye sockets. The vertical signal was collected using one electrode above the eyebrow and another on the lower edge of the eye socket. The signal reference electrode was placed in the middle of the forehead [1].

3 Methods

Recognition of Locomotion Since all further classification requires recognition of sitting and walking activity we first investigated the problem of recognising modes of locomotion, i.e. sitting, standing and walking. We used the mean and variance of the signals from the accelerometer, magnetometer and gyroscope mounted on the user's head. For classification, we used a linear support vector machine (cost $C = 1$, tolerance of termination criterion $\epsilon = 0.1$).

Recognition of Concentrated Work Three different eye movement types were detected from the EOG data as described in [1]: saccades, fixations, and blinks. The eye movement events returned by the detection algorithms were the basis for extracting different eye movement features using a sliding window. For classification, we selected the most relevant features using the method of max-relevance and min-redundancy (mRMR), and then used a linear support vector machine ($C = 1$, $\epsilon = 0.1$). The resulting train and test sets were standardised to have zero mean and a standard deviation of one. Feature selection was always performed solely on the training set.

Recognition of Walking while Conversing For recognising walking while having a conversation we used features describing the head turns as head movements and conversation are related [2]. For this, we calculated mean, variance and number of peaks in the yaw plane of the gyroscope, acceleration and magnet sensor signals and classified with a linear support vector machine.

4 Results

All presented results are based on a leave-one-person-out cross-validation and isolated recognition. Parameters of the algorithms were fixed to values common to all participants. Classification was scored using a frame-by-frame comparison with the annotated ground truth.

Recognition of Locomotion With a window size of one second and a step size of half a second we achieved for recognising sitting, walking and standing a precision of 95.4% and a recall of 90.6% averaged over all participants.

Recognition of Concentrated Work With a window size of 5 seconds and a stepsize of 0.25 seconds we achieved using person-independent training an average precision of 71.7% and recall of 70.1% for recognising concentrated office work and interaction.

Recognition of Walking while Conversing With a window size of 20 seconds for distinguishing walking alone from walking while having a conversation we achieved an average precision of 57.2% and recall of 81.3%.

5 Conclusion and Outlook

We proposed to use eye and head movements as additional sensing modalities for future context-aware hearing instruments. We collected a multi-person dataset that covers office situations that are challenging for HIs and showed that two example audio situations that are challenging for current HIs can be disambiguated using these modalities. These findings are promising in that both modalities may eventually be fully integrated even into in-ear hearing instruments. We strongly believe that developing context-aware HIs will eventually require to use a multi-modal sensing and inference approach to exploit different sources of context information.

We eventually will make this data set public. The use of this data set will require to cite this article. Contact the main author to discuss an early access.

Acknowledgment This work was part funded by CTI project 10698.1 PFLS-LS.

References

[1] A. Bulling, J. A. Ward, H. Gellersen, and G. Tröster. Eye movement analysis for activity recognition using electrooculography. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010.

[2] D. Heylen. Challenges Ahead. Head Movements and other social acts in conversation. In *AISB 2005, Social Presence Cues Symposium*, 2005.

[3] B. Shinn-Cunningham and V. Best. Selective attention in normal and impaired hearing. *Trends in Amplification*, 12(4):283, 2008.