Multimodal Recognition of Reading Activity in Transit Using Body-Worn Sensors

ANDREAS BULLING, University of Cambridge and Lancaster University
JAMIE A. WARD and HANS GELLERSEN, Lancaster University

Reading is one of the most well-studied visual activities. Vision research traditionally focuses on understanding the perceptual and cognitive processes involved in reading. In this work we recognize reading activity by jointly analyzing eye and head movements of people in an everyday environment. Eye movements are recorded using an electrooculography (EOG) system; body movements using body-worn inertial measurement units. We compare two approaches for continuous recognition of reading: String matching (STR) that explicitly models the characteristic horizontal saccades during reading, and a support vector machine (SVM) that relies on 90 eye movement features extracted from the eye movement data. We evaluate both methods in a study performed with eight participants reading while sitting at a desk, standing, walking indoors and outdoors, and riding a tram. We introduce a method to segment reading activity by exploiting the sensorimotor coordination of eye and head movements during reading. Using person-independent training, we obtain an average precision for recognizing reading of 88.9% (recall 72.3%) using STR and of 87.7% (recall 87.9%) using SVM over all participants. We show that the proposed segmentation scheme improves the performance of recognizing reading events by more than 24%. Our work demonstrates that the joint analysis of eye and body movements is beneficial for reading recognition and opens up discussion on the wider applicability of a multimodal recognition approach to other visual and physical activities.

Categories and Subject Descriptors: H.1.2 [**Models and Principles**]: User/Machine Systems—*Human information processing*; I.5.2 [**Pattern Recognition**]: Design Methodology—*Pattern analysis*; I.5.4 [**Pattern Recognition**]: Applications—*Signal processing*

General Terms: Algorithms, Experimentation, Measurement

Additional Key Words and Phrases: Recognition of reading, eye movement analysis, multimodal sensing, sensorimotor coordination, head movements, electrooculography (EOG)

ACM Reference Format:

Bulling, A., Ward, J. A., and Gellersen, H. 2012. Multimodal recognition of reading activity in transit using body-worn sensors. ACM Trans. Appl. Percept. 9, 1, Article 2 (March 2012), 21 pages.
DOI = 10.1145/2134203.2134205 http://doi.acm.org/10.1145/2134203.2134205

1. INTRODUCTION

Machine recognition of human physical activities has a long history in computer vision research (see [Mitra and Acharya 2007; Turaga et al. 2008] for surveys). The growing use of body-worn sensors, in

Authors' addresses: Contact author: A. Bulling, University of Cambridge, Computer Laboratory, 15 JJ Thomson Avenue, William Gates Building, Cambridge CB3 0FD; email: andreas.bulling@acm.org.; J. A. Ward and H. Gellersen, Lancaster University, School of Computing and Communications, Lancaster LA1 4WA, UK; email: {j.ward, hwg}@comp.lancs.ac.uk.

DOI 10.1145/2134203.2134205 http://doi.acm.org/10.1145/2134203.2134205

particular in ubiquitous computing and human-computer interaction, has paved the way for a new class of activity recognition systems [Davies et al. 2008]. Considerable advances in activity recognition were achieved by using modalities such as body movement and posture [Najafi et al. 2003], sound [Ward et al. 2006], or interactions between people [Kern et al. 2007].

In earlier work we introduced eye movements as a new modality for activity recognition [Bulling et al. 2011]. The movement patterns our eyes perform as we carry out specific activities reveal much about the activities themselves - independently of what we are looking at. This includes information on physical activities, such as driving a car [Ji and Yang 2002]; cognitive processes of visual perception, such as attention [Liversedge and Findlay 2000]; or saliency determination [Henderson 2003], and information on visual tasks, such as reading.

In human vision research, eye movements during reading have been studied for over 30 years (see Rayner [1998] for a review). A large number of studies have analyzed people's visual behavior while reading written text. Reading is a pervasive visual activity e.g., on computer screens at work, advertisements and signs in public, or books read at home or while traveling. Therefore, information on a person's reading activities is a useful indicator of his daily situation [Logan et al. 2007]. A computer interface capable of detecting reading activity could comprise the users' current level of interruptibility or task engagement, provide assistance to users with reading disabilities by automatically magnifying or explaining words or context in the text [Sibert et al. 2000; Maglio et al. 2000; Biedert et al. 2010], or infer and adapt to the users' intention [Young 2010].

1.1 Paper Scope and Contributions

We previously investigated reading recognition from horizontal eye movements using string matching and hidden markov models [Bulling et al. 2008]. The current work expands on this by demonstrating the feasibility of recognizing reading activity in different daily situations using three different types of eye and head movements. To leverage the information provided by these modalities, we introduce a flexible feature-based method for reading recognition, as well as a head-based segmentation to improve recognition performance. Using this methodology, we provide an in-depth analysis and performance evaluation of multimodal reading recognition. The specific contributions are (1) a wearable sensing approach to capture eye and head movements as a basis for reading recognition; (2) the implementation and detailed comparison of two methods for continuous recognition of reading based on string matching (STR) and a support vector machine (SVM); and (3) a new method for segmenting reading activity using information derived from head movements.

1.2 Article Organization

We first survey related work and introduce EOG as well as the main eye movement types that we identify as useful for reading recognition. We describe the algorithms developed for removing noise and baseline drift from EOG signals and for detecting three different types of eye movements: saccades, fixations, and blinks. We then introduce the classification algorithms and the features extracted from these eye movements. Finally, we present and discuss the results of a user study on reading recognition involving participants to read text while sitting at a desk, standing, walking indoors and outdoors, and riding a tram.

2. RELATED WORK

2.1 Mobile Eye Tracking

Developing sensors to track eye movements in daily life is still an active topic of research. Mobile settings call for highly miniaturized, low-power eye trackers with real-time processing capabilities. These requirements are increasingly addressed by portable video-based eye trackers, such as the commercial

Mobile Eye system by Applied Science Laboratories (ASL) or the iView X HED by SensoMotoric Instruments (SMI). However, these systems require bulky headgear and additional equipment such as digital video recorders or laptops to store and process the video streams. Due to the demanding video processing, state-of-the-art video-based eye trackers are also limited to a couple of hours of recording time.

Electrooculography (EOG)—the measurement technique used in this work—is an inexpensive method for mobile eye movement recordings; it is computationally lightweight and can be implemented using on-body sensors [Bulling et al. 2009]. These characteristics are crucial with a view to long-term eye movement recordings in daily life settings. For EOG to be truly unobtrusive, particularly in daily life settings, the design of robust electrode configurations is critical. Manabe et al. proposed the idea of an EOG gaze detector using an electrode array mounted on ordinary headphones [Manabe and Fukumoto 2006]. While this placement might reduce the problem of obtrusiveness, it raises two other issues, namely, low signal-noise ratio (SNR) and poor separation of horizontal and vertical EOG signal components. In another work, Vehkaoja et al. made electrodes from conducting fibers and sewed them into a head cap [Vehkaoja et al. 2005]. As yet, however, the device is still to be evaluated in operation. In earlier work, we introduced the wearable EOG goggles, a lightweight eye tracker based on EOG and integrated into ordinary security goggles [Bulling et al. 2009]. The device uses dry electrodes, offers real-time EOG signal processing and adaptive signal artifact removal, and allows for more than seven hours of mobile eye movement recordings.

2.2 Electrooculography Applications

Eye movement characteristics such as saccades, fixations, and blinks, as well as deliberate movement patterns detected in EOG signals, have already been used for hands-free operation of a static human-computer [Ding et al. 2005] and human-robot [Chen and Newman 2004] interfaces. EOG-based interfaces have also been developed for assistive robots [Wijesoma et al. 2005] or as a control for an electric wheelchair [Barea et al. 2002]. Such systems are intended to be used by physically disabled people who have extremely limited peripheral mobility but still retain eye-motor coordination. These studies showed that EOG is a measurement technique that is inexpensive, easy to use, reliable, and relatively unobtrusive when compared to head-worn cameras used in video-based eye trackers. While these applications all used EOG as a direct control interface, our approach is to use EOG as a source of information on a person's activity.

2.3 Eye Movement Analysis

A growing number of researchers use video-based eye tracking to study visual behavior in natural environments. This has led to important advances on our understanding of how the brain processes tasks, and of the role that the visual system plays in this [Hayhoe and Ballard 2005]. Eye movement analysis has a long history as a tool to investigate visual behavior. In an early study, Hacisalihzade et al. used Markov processes to model visual fixations of observers recognizing an object [Hacisalihzade et al. 1992]. They transformed fixation sequences into character strings and used the string edit distance to quantify the similarity of eye movements. Elhelw et al. used discrete time Markov chains on sequences of temporal fixations to identify salient image features that affect the perception of visual realism [Elhelw et al. 2008]. They found that fixations were able to uncover the features that most attract an observer's attention. Dempere-Marco et al. presented a method for training novices in assessing tomography images [Dempere-Marco et al. 2002]. They modeled the assessment behavior of domain experts based on the dynamics of their saccadic eye movements. Salvucci et al. evaluated means for automated analysis of eye movements [Salvucci and Anderson 2001]. They described three methods based on sequence-matching and hidden Markov models that interpreted eye movements as accurately as human experts but in significantly less time. Canosa analyzed different tasks such as reading, counting, talking, sorting, and walking to determine the extent to which everyday tasks performed in real-world environments affect visual perception [Canosa 2009]. She showed that these tasks could be compared and distinguished from one another by using eye movement features such as mean fixation duration or mean saccade amplitude.

All of these studies aimed to model visual behavior during specific tasks using well-known eye movement characteristics. They explored the link between the task and eye movements, but did not recognise the task or activity using this information.

2.4 Activity Recognition

In a recent work, Logan et al. aimed to recognize common activities in a real-world setting using a large variety and number of environmental sensors such as wired reed switches, RFID tags, and infra-red motion detectors [Logan et al. 2007]. They discovered that among the activities they investigated reading was one of the most difficult activities to detect, and concluded that for covering all types of physical activity in daily life, improved algorithms need to be developed.

Most previous attempts to recognize reading have been based on video-based eye trackers. With the goal of building a more natural computer interface, Campbell et al. investigated on-screen reading recognition using infra-red cameras to track eye movements [Campbell and Maglio 2001]. The approach was participant-independent, robust against noise and had a reported accuracy of 100%. However, the system required that each participant's head was kept still by using a chin rest.

In a later work, Keat et al. proposed an improved algorithm to determine whether a user is engaged in reading activity on a computer monitor [Keat et al. 2003]. Using an ordinary video camera placed between the participant and monitor, 10 participants were asked to read an interesting text from a list of preselected articles. The participants were explicitly asked to undertake other types of common computer-related activities such as playing computer games or watching video clips during the course of the experiment. Using user-dependent training, they achieved an average reading detection accuracy of 85.0% with a false alarm rate of 14.2%. However, to ensure correct detection of gaze direction, participants were required to face the screen throughout the experiments.

Motivated by the goal of improving reading skills for people with reading disabilities, Sibert et al. developed a system for remedial reading instruction [Sibert et al. 2000]. Based on visual scanning patterns, the system used visually controlled auditory prompting to help the user with recognition and pronunciation of words. Following the study, participants reported that the most obtrusive part of the system was the video camera used to track eye movements.

In earlier work, we used eye movement analysis to recognize a set of common office activities in a work environment: copying a text between two screens, reading a printed paper, taking hand-written notes, watching a video, and browsing the web [Bulling et al. 2011]. We also included periods of rest (the NULL class) during which the participants were asked not to engage in any of the other activities. Using a SVM classifier and person-independent training, we obtained an average precision of 76.1% and recall of 70.5% over all classes and participants.

All of these studies used eye movement analysis for activity recognition. The current work, however, is the first to fuse information derived from eye movements with that from other modalities, in this case head movements, to increase recognition performance.

3. SENSING AND SIGNAL PROCESSING

3.1 Electrooculography

The eye can be modeled as a dipole with its positive pole at the cornea and its negative pole at the retina. Assuming a stable corneo-retinal potential difference, the eye is the origin of a steady electric

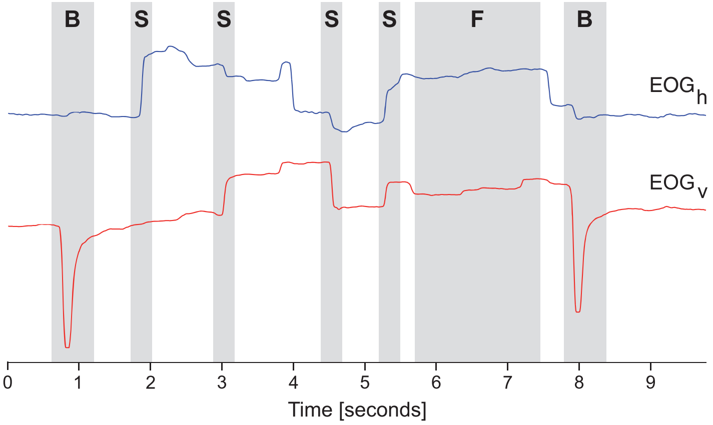


Fig. 1. Denoised and baseline drift removed horizontal (EOG_h) and vertical (EOG_v) signal components. Examples of the three main eye movement types detected in both signal components are marked in grey: saccades (S), fixations (F), and blinks (B).

potential field. The electrical signal that can be measured from this field is called the electrooculogram. If the eye moves away from the center position, the retina approaches one electrode while the cornea approaches the opposing one. This change in dipole orientation causes a change in the electric potential field, and thus the measured EOG signal amplitude. By analyzing these changes, eye movements can be tracked. Using two pairs of skin electrodes placed at opposite sides of the eye and an additional reference electrode on the forehead, two signal components (EOG_h and EOG_v), corresponding to two movement components—a horizontal and a vertical—can be identified. EOG signal amplitudes typically range from 5 μV/degree to 20 μV/degree with an essential frequency content between 0 Hz and 30 Hz [Brown et al. 2006].

3.2 Eye Movement Types

To be able to use eye movement analysis for reading recognition, it is important to understand the different types of eye movements. We identified three basic types that can be detected using EOG: saccades, fixations, and blinks (see Figure 1).

3.2.1 *Saccades.* The eyes do not remain still when viewing a visual scene but move constantly to build up a mental "map" from interesting parts of that scene. The main reason for this is that only a small central region of the retina, the fovea, is able to perceive with high acuity. The fast movement of the eyes is called a saccade. The duration of a saccade depends on the angular distance the eyes travel during this movement: the so-called saccade amplitude. Typical characteristics of saccadic eye movements are 20 degrees of visual angle for the amplitude, and 10 ms to 100 ms for the duration [Duchowski 2007]. Eye movements during reading are characterised by repetitive saccades of different amplitude in horizontal direction [Rayner 1998].

3.2.2 *Fixations.* Fixations are the stationary states of the eyes during which gaze is held upon a specific location in the visual scene. Fixations are typically defined as the time between each two saccades. The average fixation duration lies between 100 ms and 200 ms [Manor and Gordon 2003].

3.2.3 *Blinks.* The frontal part of the cornea is coated with the so-called precornial tear film. To spread this fluid across the corneal surface, regular opening and closing of the eyelids, or blinking, is required. The average blink rate varies between 12 and 19 blinks per minute while at rest [Karson et al. 1981]; it is influenced by environmental factors such as relative humidity, temperature or brightness, but also by physical activity, cognitive workload, or fatigue [Schleicher et al. 2008]. The average blink duration lies between 100 ms and 400 ms [Schiffman 2001].

3.3 EOG Signal Processing

3.3.1 *Noise and Baseline Drift Removal.* $\mathrm{EOG_h}$ and $\mathrm{EOG_v}$ are first stripped of high frequency noise using a median filter. For baseline drift removal, we then perform an approximated multilevel 1-D wavelet decomposition at level nine using Daubechies wavelets on $\mathrm{EOG_h}$ and $\mathrm{EOG_v}$. The reconstructed decomposition coefficients give a baseline drift estimation. Subtracting this estimation from the original signals yields the corrected signals with reduced drift offset (see Tinati and Mozaffary [2006] for further details).

3.3.2 *Saccade Detection.* For saccade detection we developed the *continuous wavelet transform - saccade detection* (CWT-SD) algorithm. CWT-SD detects saccades by thresholding on the continuous 1-D wavelet coefficient vector computed from the de-noised and baseline drift removed $\mathrm{EOG_h}$ and $\mathrm{EOG_v}$. Small and large saccades are distinguished using different thresholds; saccade direction is obtained from the sign of the first derivative of the signal. To improve the algorithm's robustness to differences in EOG signal quality, an additional step removes all saccade candidates that do not comply to typical physiological saccade characteristics described in the literature [Duchowski 2007].

3.3.3 *Fixation Detection.* Our algorithm for fixation detection exploits the fact that fixation points tend to cluster together closely in time. Thus, by thresholding on the dispersion of these points, fixations can be detected. $\mathrm{EOG_h}$ and $\mathrm{EOG_v}$ are first divided into saccadic and nonsaccadic segments using the output from CWT-SD. For each nonsaccadic segment, the algorithm then calculates the corresponding dispersion and duration values. If the dispersion is below a maximum threshold, and the duration above a minimum threshold, a fixation is detected (see Widdel [1984] for typical values).

3.3.4 *Blink Detection.* Similar to the saccade detection, the so-called *continuous wavelet transform - blink detection* (CWT-BD) algorithm uses thresholding of wavelet coefficients to detect blinks in $\mathrm{EOG_v}$. In contrast to saccades, a blink is characterized by a short sequence of two large peaks in the coefficient vector: one positive, the other negative. The time between these peaks is much smaller than for saccades. Thus, blinks are distinguished from saccades by applying a maximum threshold on this time difference.

4. METHODOLOGY

We first provide an overview of the classification algorithms used in this work. We then detail the method developed for analyzing repetitive eye movement sequences, and the features used by the SVM classifier. Finally, we detail the head-based segmentation and the methods used for evaluating the classifiers' performance.

4.1 Classification Algorithms

Figure 2 shows the two processing chains for STR and SVM. Input to both chains are the processed EOG signals capturing the horizontal and vertical eye movement components. In the first stage, these signals are processed to remove any artifacts that might hamper eye movement analysis. Only this initial processing depends on the particular eye tracking technique used; all further stages are completely independent of the underlying type of eye movement data. In the second stage, for STR, horizontal saccades are detected from the processed eye movement data, encoded into a saccade sequence, and fed into the string matching classifier. For SVM, three eye movement types are detected from the eye movement data: saccades, fixations, and blinks. The corresponding eye movement events returned

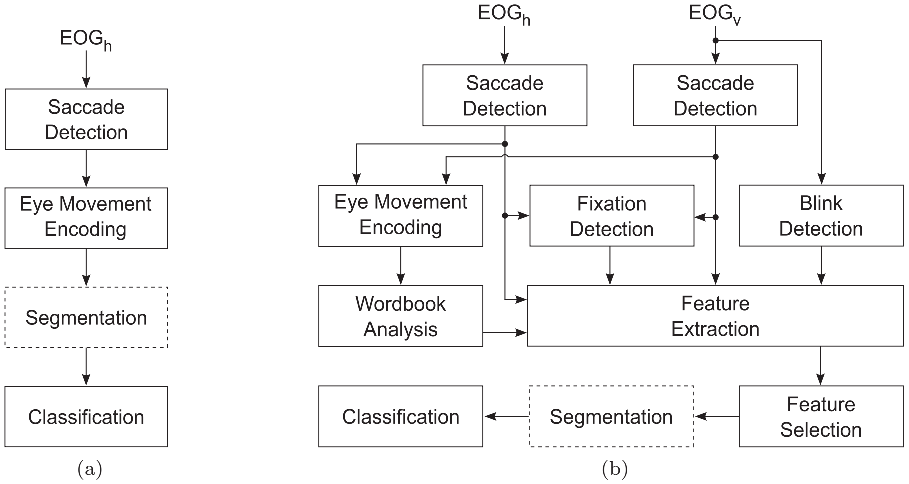


Fig. 2. Eye-based reading recognition using string matching (a); a support vector machine classifier (b) with optional head-based reading segmentation.

Table I. Qualitative Comparison of the Two Classifiers in this Work (with respect to memory footprint, computational complexity, for classifier training, and flexibility of application to other activities)

	String Matching	Support Vector Machine
Memory footprint	low	high
Computational complexity	low	high
Classifier Training	not required	required
Flexibility	specific	generic

by the detection algorithms are the basis for extracting different eye movement features using a sliding window. Finally, a hybrid method selects the most relevant of these features, and uses them for classification.

Particularly in mobile settings, visual tasks such as reading require spatial and temporal coordination between movements of the eyes and other parts of the human body [Sailer et al. 2005; Pelz et al. 2001; Johansson et al. 2001]. It is therefore conceivable that the joint analysis of eye and body movements is beneficial for reading recognition. Thus, in both processing chains, a segmentation stage that segments reading activity based on head movements may precede the classification stage.

As the processing chains for STR and SVM differ, so do their characteristics in terms of memory footprint, computational complexity, and applicability to other activities (see Table I for a qualitative comparison). String matching has a low memory footprint as well as computational complexity and does not require training prior to classification. In contrast, the SVM algorithm does not rely on a specific reading template but uses a feature-based classification approach. This approach, while being more flexible, has a considerably higher memory footprint as well as computational complexity and requires classifier training.

4.1.1 *String Matching.* We first encode the direction and amplitude of saccades in EOG$_h$ into a string consisting of four different characters: “L” for large saccades to the left; “R” for large saccades to the right; “l” for small saccades to the left; and “r” for small saccades to the right (c.f., Figure 3). During reading, small saccades correspond to jumps between words while large saccades are those

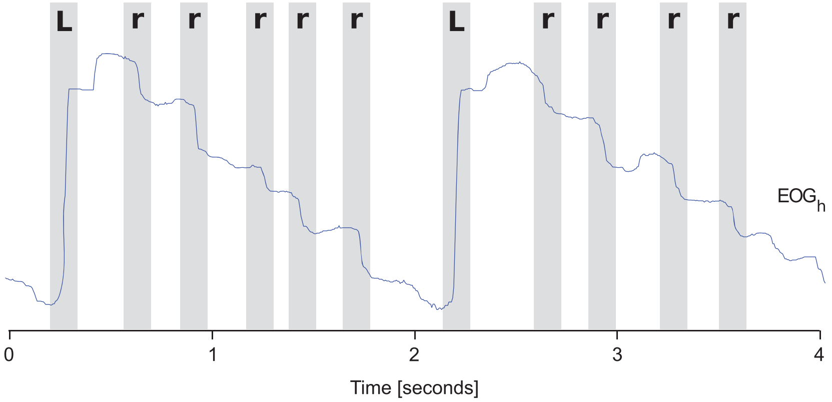


Fig. 3. De-noised and baseline drift removed horizontal EOG reading signal and corresponding string encoding.

observed during an end of line "carriage return". This encoded string is the basis for classification using string matching. The matching is performed by moving a string template, representing a typical reading instance, over the encoded string sequence, character-by-character. In each of these steps, the algorithm calculates the Levenshtein distance between the template and the current string sequence. The Levenshtein distance between two strings is given by the minimum number of operations needed to transform one string into the other, where an operation is an insertion, a deletion, or a substitution of a single character [Levenshtein 1966].

The algorithm then applies a threshold T_{ed} on the Levenshtein distance vector to separate the two classes "reading" and "not reading". This threshold defines how tolerant the classification is towards relative error in the edit distance. As the method does not yet adapt the string template to the signal while calculating the distances, it is sensitive to fluctuations in the number of small saccades. This results in a high number of false insertions. To counter this we slide a majority vote window W_{str} over the event-based classifier output to "smooth" the final result.

4.1.2 *Support Vector Machine.* We chose a linear support vector machine that yielded good performance in an earlier study on eye-based activity recognition [Bulling et al. 2011]. Our particular SVM implementation uses a fast sequential dual method for dealing with multiple classes [Crammer and Singer 2003; Lin 2008]. This reduces training time considerably while retaining recognition performance.

Analysis of Eye Movement Sequences. We first encode eye movements by mapping saccades with different direction and amplitude to a character-based representation. Our algorithm takes the CWT-SD saccades from EOG$_\mathrm{h}$ and EOG$_\mathrm{v}$ as its input. It then checks for simultaneous saccades in both components, as these represent diagonal eye movements. Simultaneous saccades are characterized by overlapping saccade segments in the time domain. If no simultaneous saccades are detected, the saccade is directly encoded. If two saccades are detected, the algorithm maps them onto one of 24 discrete characters.

An eye movement sequence is defined as a string of l successive characters. As an example, with $l = 4$, "LrBd" translates to large left (L) → small right (r) → large diagonal right (B) → small down (d). To extract information on repetitive eye movement sequences, such strings are collected in wordbooks using a sliding window of length l and a step size of one. Each newly found string is added to the corresponding wordbook Wb_l. For a string that is already included in Wb_l, its occurrence count is increased by one.

Table II. Summary of 90 Eye Movement Features in this Work

Group	Features
saccade	mean, variance and maximum EOG signal amplitudes within a saccade, rate of small and large, positive and negative saccades in horizontal and vertical directions
fixation	mean and variance of the horizontal and vertical EOG signal amplitude within a fixation, fixation duration, fixation rate
blink	mean and variance of the blink duration, blink rate
wordbook	wordbook size, maximum and difference between maximum and minimum occurrence count, mean and variance of all occurrence counts in the wordbook (all for different pattern lengths)

Feature Extraction. We extract four groups of features based on the detected saccades, fixations, blinks, and the wordbooks. Table II details the naming scheme used for all of these features. The features are calculated using a sliding window (window size $W_{fe} = 5s$ and step size $S_{fe} = 0.25s$) on $\mathrm{EOG_h}$ and $\mathrm{EOG_v}$.

Features calculated from saccadic eye movements make up the largest proportion of extracted features. In total, there are 62 such features comprising the mean, variance and maximum EOG signal amplitudes of saccades, and the normalised saccade rates. These are calculated for both $\mathrm{EOG_h}$ and $\mathrm{EOG_v}$; for small and large saccades; for saccades in positive or negative direction; and for all possible combinations of these.

We calculate five different features using fixations: the mean and variance of the EOG signal amplitude within a fixation; the mean and the variance of fixation duration; and the fixation rate over window W_{fe}.

For blinks, we extract three features: blink rate, and the mean and variance of the blink duration.

We use four wordbooks. This allowed us to account for all possible eye movement patterns up to a length of four ($l = 4$), with each wordbook containing the type and occurrence count of all patterns found. For each wordbook we extract five features: the wordbook size, the maximum occurrence count, the difference between the maximum and minimum occurrence counts, and the variance and mean of all occurrence counts.

Feature Selection. For feature selection, we choose a filter scheme over the commonly used wrapper approaches because of the lower computational costs and thus shorter runtime given the large dataset. We use minimum redundancy maximum relevance feature selection (mRMR) for discrete variables [Peng et al. 2005; Peng 2008]. The mRMR algorithm selects a feature subset of arbitrary size S that best characterizes the statistical properties of the given target class based on the ground truth labeling. In this work, we fixed S to 90. In contrast to other methods such as the F-test, mRMR also considers relationship between features during the selection. Among the possible underlying statistical measures described in the literature, mutual information was shown to yield the most promising results, and thus was selected in this work.

4.2 Head-Based Segmentation

With a view to a later implementation on an embedded device with limited processing power, we developed a segmentation approach that only requires one single-axis head-mounted accelerometer. The segmentation is based on two assumptions: (1) that reading only occurs while the participant's head is down; and (2) that up and down movements of the head can be reliably detected using the head-mounted accelerometer.

We detect these head movements by thresholding on the x-component of the de-noised and mean subtracted head acceleration signal (see white arrow in Figure 5). Time periods during which this signal is equal or below a threshold T_{segd} are interpreted as "head down" and the remaining time periods as "head up". This procedure splits the whole dataset into "head up" and "head down" segments.

While this segmentation scheme is computationally lightweight and fast, it is also sensitive to artifacts in the acceleration signal. For example, strikes to the motion sensor may cause signal artifacts that are similar to signal changes caused by a person moving his head. To counter these, we remove "head down" segments that are shorter than a minimum time period T_{segt}.

During classification, the output is initially assumed to be "not reading" throughout. We then run the head-based segmentation on the eye movement data and apply isolated classification on the "head down" segments only.

5. EXPERIMENT

The experimental setup was designed with two main objectives in mind: (1) to record eye and head movements in an unobtrusive manner in a real-world setting; and (2) to evaluate how well the reading by people in transit can be recognized using these modalities. We defined a scenario of traveling to and from work containing a semi-naturalistic set of reading activities. It involved participants reading text while engaged in a sequence of activities such as sitting at a desk, walking along a corridor, walking along a street, waiting at a tram stop, and riding a tram.

5.1 Apparatus

For EOG data acquisition we used a commercial system, the Mobi from Twente Medical Systems International (TMSI), which was worn on a belt around each participant's waist (see Figure 5). The device is capable of recording a four-channel EOG with a joint sampling rate of 128Hz.

The data was collected using an array of five electrodes positioned around the right eye (see Figure 5). The electrodes used were the 24mm Ag/AgCl wet ARBO type from Tyco Healthcare equipped with an adhesive brim to stick to the skin. The horizontal signal was collected using one electrode on the nose and another directly across from this on the edge of the eye socket. The vertical signal was collected using one electrode above the eyebrow and another on the lower edge of the eye socket. The fifth electrode, the signal reference, was placed away from the other electrodes in the middle of the forehead.

For motion tracking we used three MTx sensors from XSens Technologies. The sensors were positioned on top of each participant's head and on the back of their hands (see Figure 5). Unfortunately, the MTx system performed poorly using Bluetooth, and so we were forced to use a wired connection. This was the only physical connection between the participant and assistant. Care was taken by the assistant to ensure that the trailing wire did not interfere or distract the participant.

All recorded data was sent to a laptop in the backpack worn by the assistant. Data recording and synchronization were handled using the context recognition network (CRN) toolbox [Bannach et al. 2008]. We made two extensions to the toolbox: the first was a reader to acquire the annotation from the Wii remote controller; the second extension was a "heartbeat" component that provided audio feedback to the assistant on whether the toolbox was running and recording data, thus providing instant notification of device failures. The "heartbeat" allowed the assistant to concentrate on labeling and observing the participants rather than continually disturbing the procedure by checking the recording status.

5.2 Procedure

Before the experiment participants were only informed about the measurement equipment–neither the later use of the sensors nor how reading would be recognized were explained to them. Participants were asked to follow two different sequences: A first calibration sequence involved walking around a circular corridor for approximately two minutes while reading continuously. The second sequence involved a walk and tram ride to and from work (see Figure 4). This sequence was repeated in three runs. The first run was carried out as a baseline case without any reading task. This was both to

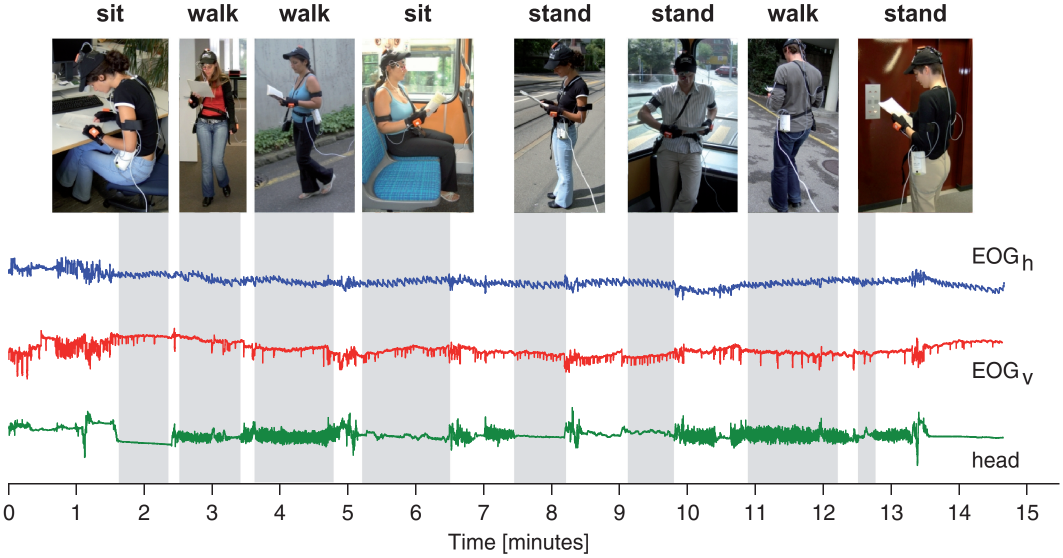


Fig. 4. Experimental procedure involving a sequence in reading activity while sitting at a desk, walking along a corridor, walking along a street, waiting at a tram stop, and riding a tram. The figure also shows the corresponding horizontal (EOG$_h$) and vertical (EOG$_v$) EOG signals as well as acceleration data from the head of one complete dataset.

accustom the participants to the route, but also to provide a reasonable amount of NULL data, which contributed to the objective of obtaining a realistic dataset. In two subsequent runs the participants were asked to read a text throughout. Between each run the participants rested for about five minutes. The total experiment time for each participant was about one hour. At the end of each experiment, the participants were asked on their experiences on the procedure in a questionnaire.

In contrast to a previous study [Campbell and Maglio 2001], we opted to allow a free choice on reading material. Only two conditions were made: (1) that the material was text-only, that is, no pictures; and (2) that participants only chose material they found interesting and long enough to provide up to an hour's worth of reading. Thus the type of text, its style (e.g., number of columns), as well as page and font size could be chosen to each participant's personal preference, and were in fact different across participants. Our objective was to induce a state where readers were engrossed in the task for the relatively long recording time, thus allowing us to gather realistic data without having to coerce participants. A further benefit was that if participants were engrossed in the task, they would be less likely to be distracted by other people.

We were able to collect data from eight participants (four female and four male), aged between 23 and 35 years. Originally there were 10 participants, but two had to be withdrawn due to recording problems resulting in incomplete data. Most of the experiments were carried out in well-lit, fair to cloudy conditions, with two exceptions: One of the male participants was recorded at night where we had to rely on street lights while walking around outdoors. Another male was recorded in the rain where an assistant had to hold an umbrella over the participant to protect the sensors and reading material. However, as neither of the datasets showed a decrease in signal quality, both were used for the analysis.

5.2.1 *Annotation of Ground Truth.* Participants were tailed by an assistant who annotated both the participants' activity (sitting, standing, walking) and whether they were reading. For this level of detail (are the participants' eyes on the page or not) the assistant had to monitor the participants

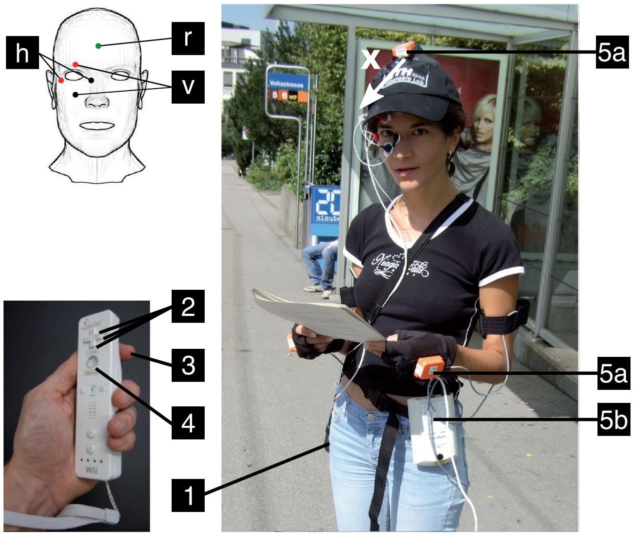


Fig. 5. Experimental setup consisting of five EOG electrodes (h: horizontal, v: vertical, r: reference) and the Mobi (1). The bottom figure shows the wireless Wii controler used for annotating the ground truth with three thumb-control buttons (2); the reading trigger button (3); and a button for labeling special events (4). Also shown are the three XSens motion sensors on the head and on the back of both hands (5a); and the XBus Master (5b). The x-direction of the head-mounted accelerometer is indicated with a white arrow.

from a close proximity, but without being so close as to cause a distraction. For this purpose we used a wireless Nintendo Wii remote controller (see Figure 5). Using the Wii's thumb-control buttons "up," "down," and "right," the assistant could annotate the basic activities of standing, sitting, and walking. In parallel, the trigger button was held down whenever the participants appeared to be reading and released when they stopped. A fifth button was used to annotate special events of interest, such as when the participants passed through a door and while entering or leaving the tram.

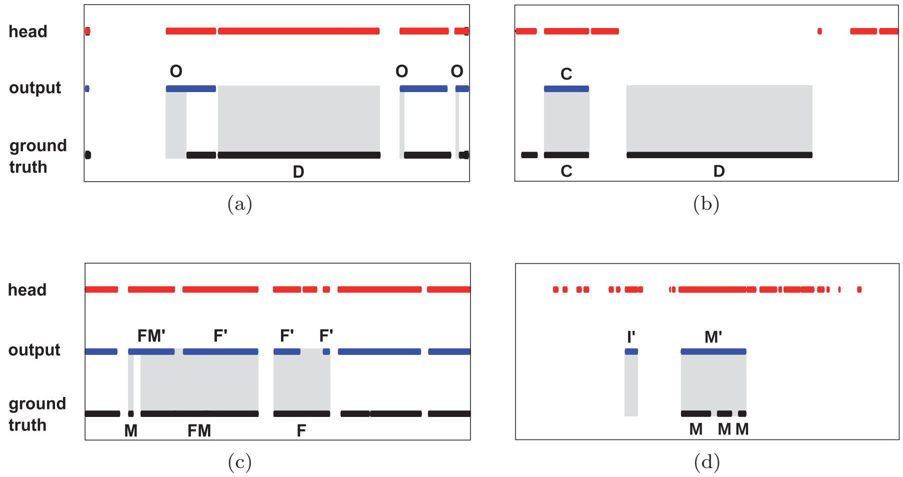
6. EVALUATION

6.1 Parameter Selection

6.1.1 *Saccade Detection.* To determine the threshold parameters of the CWT-SD algorithm, we used a sweep on a manually cut subset of the data. On average, for all participants, this subset of data contained 15 large reading saccades plus noise and artifacts caused by interrupting eye movements. For each threshold, we counted the number of large saccades that were detected and calculated the *relative error* ($\frac{Total-Detected}{Total}$). Based on this sweep, we chose the large saccade threshold at $T^{L}_{sac} = 7000$. Due to the difficulties in manually segmenting samples of small saccades, we approximated the small threshold $T^{s}_{sac} = 1500$.

6.1.2 *String Matching and Support Vector Machine.* The string matching parameters were evaluated across a sweep of the majority vote window length W_{str}, the distance threshold T_{ed}, and different template lengths. Based on this sweep, we chose $W_{str} = 30$, $T_{ed} = 3$, and template "*Lrrr*" for all participants.

The two main parameters of the SVM algorithm, the cost C and the tolerance of termination criterion ϵ, were fixed to $C = 1$ and $\epsilon = 0.1$.



(a) (b) (c) (d)

Fig. 6. Example time- and event-based errors in continuous, multimodal recognition of reading indicated by mismatches between ground truth, head-based segmentation, and recognition output. (a) Overfill errors introduced by head-based segmentation (O) and deletion due to misclassified reading segment (D); (b) correctly classified ground truth segment (C) and deletion introduced by segmentation (D); (c) fragmented only (F) and fragmented and merged (FM) ground truth segments and corresponding fragmenting (F'), as well as fragmenting and merging (FM') outputs; (d) insertion (I') and three ground truth segments (M) merged into a single reading output (M').

6.1.3 *Head-Based Segmentation.* To determine the parameters of the head-based segmentation, we swept $T_{segd} = 0.001 \ldots 0.75$ (in 750 steps) and $T_{segt} = 0.25s, 0.5s, 0.75s, 1s, 2s, 3s, 4s, 5s$. We then chose those parameters that minimized the sum of deletions (D) and insertions (I) minus the number of correctly classified ground truth segments (C, see next paragraph for a description) according to Eq. (1).

$$\min_{D,I,C \in \mathbb{N}_0} (D + I - C). \tag{1}$$

Based on this analysis, we set $T_{segd} = 0.64$ and $T_{segt} = 1s$.

6.2 Comparison of Classification Algorithms

Classification and feature selection were evaluated using a leave-one-person-out scheme: we combined the datasets of all but one participant and used this for training; testing was done using the datasets of the remaining participant. This was repeated for each participant. The resulting train and test sets were standardized to have zero mean and a standard deviation of one. Feature selection was always performed solely on the training set.

We first investigated the robustness of reading recognition with head-based segmentation across participants. We evaluated the performance of both classifiers with respect to the time and event errors that occur in continuous recognition of reading (see Figure 6 for examples). To this end, we first divided the recognition output and the ground truth into segments, that is, contiguous sequences of same class samples. For time-based performance evaluation we then calculated performance measures using a direct timewise (sample-by-sample) comparison of the ground truth with the classifier output. For event-based performance evaluation we interpreted each segment as a reading activity event and compared ground truth events with recognition output events to calculate performance measures. For a fair comparison with STR, we also investigated the case of only using the horizontal eye movement

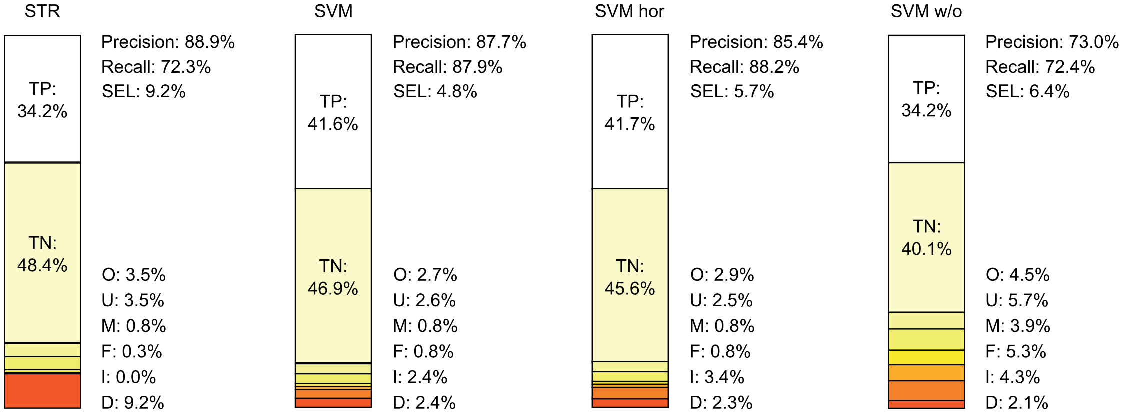


Fig. 7. Error division diagrams (EDD) for string matching (STR) as well as support vector machine with head-based segmentation using both eye movement components (SVM), only using the horizontal component (SVM hor), and both components but without head-based segmentation (SVM w/o). The EDDs show the proportion of the total dataset comprising true positives (TP), true negatives (TN), overfill (O), underfill (U), merge (M), and fragmentation (F) errors; insertion (I) and deletion (D) errors form the serious error level, SEL (see Figure 6 for a description of these error classes). Results are averaged over all participants using person-independent parameters.

component (two horizontal electrodes plus reference) for SVM. In this case, only 28 saccade features could be extracted from the horizontal EOG signal and used for classification.

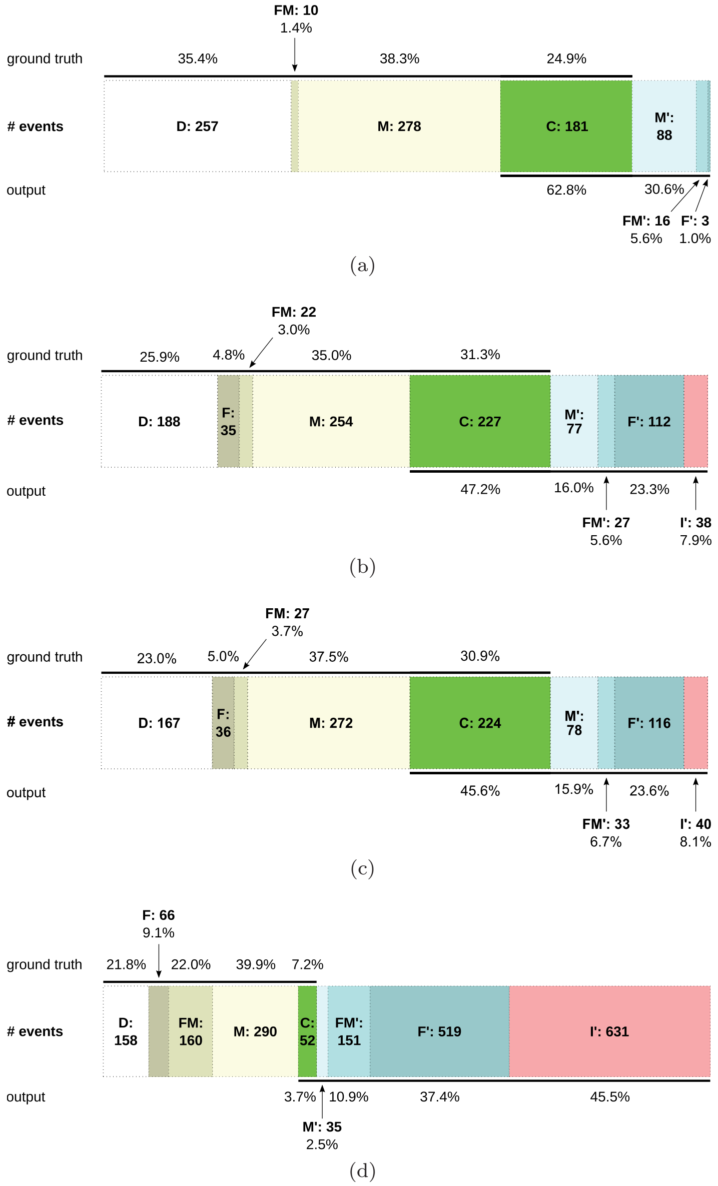
The error division diagrams (EDD) in Figure 7 provide a detailed breakdown of the time-based recognition performance for both classifiers averaged over all participants. An EDD summarizes the typical time errors that occur in continuous recognition systems. Specifically, there are three classes of error that we consider in addition to the true negative (TN) and true positive (TP) times (see Figure 6). Details on how these error classes are derived is outside the scope of the current work, but can be found in Ward et al. [2011].

Serious errors. Insertions (I) describe when a reading segment is detected, but there is none in the ground truth. Deletions (D) are failures to detect a reading segment. Taken together, both errors are called the serious error level, SEL.

Fragmentation and merge. Fragmentation errors (F) describe when a reading segment in the ground truth corresponds to several segments in the recognition system output. Merge (M) is the opposite: several ground truth reading segments are combined into one segment. Both errors are serious only if counts of reading segments are to be taken (e.g., as part of a statistical analysis)

Timing errors. Overfill errors (O) ocur when a segment in the system output extends into regions of NULL. The opposite of overfill is underfill (U), in this case the segment recognized by the system fails to “cover” parts of the ground truth segment. These errors are typically not considered serious, as they typically caused by slightly offset labeling or delays in the recognition system.

Using this breakdown, Figure 7 shows that SVM performs best with an overall time-based recognition performance of 87.7% precision (recall 87.9%) and a SEL of 4.8%. The largest sources of error are overfills and underfills (5.3%), as well as deletions and insertions, each accounting for 2.4% of the total experiment time. Only the use of the horizontal eye-movement component reduces the recognition performance to 85.4% precision (recall 88.2%) and a SEL of 5.7%. STR yields a slightly higher precision of 88.9%, but a recall of only 72.3% and a SEL of 9.2% (caused only by deletions). The reduced performance is mainly caused by a drop in TP to 34.2% and an increase in overfill and underfill errors.



(a)

(b)

(c)

(d)

Fig. 8. Event analysis diagram (EAD) for string matching (a), as well as a support vector machine with head-based segmentation using both eye movement components (b); only using the horizontal component (c); and both components but without head-based segmentation (d). Results are averaged over all participants using person-independent parameters. The EADs show counts of deleted (D), fragmented (F), fragmented and merged (FM), merged (M), and correct (C) events as a proportion of the total ground truth event count; and C, M', FM', F', and insertions (I') as a proportion of the total recognition output event count (see Figure 6 for a description of these error classes).

The event analysis diagram (EAD) in Figure 8 shows that SVM also outperforms STR in terms of event-based reading recognition performance. An EAD summarizes the typical event errors that occur in continuous recognition systems. Specifically, we distinguish between nine possible outcomes when ground truth events are compared against recognition output events (see Figure 6): correct (C), when

Table III. Selected Event Error Counts for SVM with Head-Based Segmentation for Each Participant. (Results referred to in the text are marked in bold and italic. The table also shows the participants' gender (f: female, m: male); the dataset recorded at night is marked with *, the one recorded in heavy rain with **)

	P1 (f)	P2 (m*)	P3 (f)	P4 (m**)	P5 (f)	P6 (m)	P7 (f)	P8 (m)
Deletion	26	17	18	**59**	**62**	*5*	14	14
Merge	*7*	34	51	**72**	29	*4*	25	13
Correct	*46*	20	17	30	7	*36*	29	19
Total	81	76	94	167	102	47	70	56
Merging Output	3	9	10	**24**	9	2	10	6
Fragmenting Output	3	9	10	7	2	4	4	21
Insertion	10	25	9	12	2	12	14	5
Total Output	63	65	52	78	25	54	57	52

ground and output match; deleted (D), when there is no matching output; inserted (I'), when there is no matching ground; fragmented (F), when a ground is recognized by more than one matching output; and merged (M), when a ground truth event is connected to one or more others by a single large matching output event. We further refer to a "fragmenting" output (F') as any one of the matching outputs that help create a fragmented ground event. Similarly, we refer to a "merging" output (M') as the large output event that merges two or more ground events. If a ground event is both fragmented and merged, we refer to it as FM; similarly an output event can be both fragmenting and merging (FM').

Figures 8(a) and 8(b) show that SVM correctly classifies 227 reading events while STR is correct in 181 cases. These events account for 31.3% (SVM) and 24.9% (STR) of the total number of actual reading events. The largest sources of errors are event deletions (SVM: 188, STR: 257) and event merges (SVM: 254, STR: 278). Comparing the number of merges (M) with the number of merging outputs (M'), the EAD also indicates that, for both classifiers, a large number of short reading events got merged into a much smaller number of reading output events. Figure 8(c) shows that only using the horizontal eye-movement component does not considerably reduce the event-based recognition performance (224 correctly classified reading events).

6.3 Event-Based Performance for Each Participant

Table III reveals large interperson differences in the event-based recognition performance for SVM with head-based segmentation. More than 50% of the total number of deleted events are caused by only two participants, that is, P4 (59 deletions) and P5 (62 deletions). P4 has the highest error counts for merged (72 events) and merging returns (24 events). In contrast, P1 and P6 achieved the best performance with 46 (P1) and 36 (P6) correctly classified events, 7 (P1) and 4 (P6) merged events, as well as only 5 deleted events (P6).

6.4 Contribution of Head-Based Segmentation

Based on these results we further investigated the contribution of the head-based segmentation to the overall recognition performance of the SVM classifier. The right EDD in Figure 7 shows that—without head-based segmentation—the time-based recognition performance of SVM decreases to 73.0% precision (recall 72.4%) and a SEL of 6.4%. Comparing these results with the left EDD in Figure 7 clearly shows a drop in recognition performance but one that, depending on the particular end application, may still be considered acceptable.

The event-based performance analysis shown in Figures 8(b) and 8(d) draws a different picture. Without head-based segmentation, SVM correctly classifies only 52 events while causing 158 event deletions; 160 ground truth events to be fragmented and merged; 290 merged events; 151 fragmenting and merging outputs; 519 fragmenting outputs; and 631 insertions. Expressed as a percentage of the total number of actual reading events, the percentage of correctly classified events drops to 7.2%.

7. DISCUSSION

7.1 On the Recognition Performance

Two different approaches to recognizing reading based on EOG in a mobile setting have been described and investigated in this work. String matching is computationally more lightweight, as it only uses simple arithmetic; it can be easily adapted to a future online implementation, for example on a wearable device. As we have shown in Bulling et al. [2011] a feature-based approach such as SVM is more flexible and scales better with the number and type of activities that are to be recognized. However, this also comes at a higher computational complexity.

Using person-independent parameters and a NULL class of over 50%, SVM performs best with an average precision of 87.7% (recall 87.9%) over all participants (see Figure 7). SVM also outperforms string matching with respect to the event-based performance with 31.3% correctly classified reading events (see Figure 8). As can be seen from Figures 7 and 8(c) these findings hold if only those features extracted from the horizontal eye-movement component are used for recognition; features extracted from the vertical eye-movement component contribute little to the overall recognition performance. More than half of all event errors are merges and deletions. For many end applications, merge errors may be considered less serious, as the actual reading events are still recognized correctly. Deletions are a more serious type of error, as they refer to reading events that got removed completely. Table III shows that P4 and P5 alone contribute more than 50% of all deletions. It remains to be investigated whether in these cases the errors are caused by inaccuracies of the labeling process, by errors in the head-based segmentation or by misclassification.

The proposed segmentation scheme contributes considerably to the overall recognition performance. Without head-based segmentation, the time-based recognition performance drops by 14.7% in precision and 15.5% in recall (see Figure 7). This drop in overall performance is caused by an increase in all error classes except for deletions. This was to be expected, as the segmentation does not help in classifying but in spotting the begin and end of reading activity in the continuous data stream. The percentage of correctly classified reading events even drops by 24.1%, with the majority of errors being fragmentation and insertion errors (see Figure 8). These findings suggest that time-based performance evaluation, while commonly used in activity recognition research, is not sufficient to assess the performance of an activity recognition system. It is crucial, too, to assess a system's recognition performance in terms of activity events. Multimodal sensing is a promising approach to improve event-based recognition performance, particularly if, as shown in this work, it adds only little to the complexity of the sensor setup.

7.2 On Electrooculography

Our study demonstrates that EOG is a robust technique for recording eye movements in mobile settings. The main advantage of EOG over common video-based systems is the fact that the participants only have to wear relatively unobtrusive and lightweight equipment. This contributes to the participants feeling unconstrained during the experiments, and therefore allows for natural reading behavior.

One drawback is that EOG electrodes require good skin contact. Poor placement of electrodes was the reason for many of the problems in our work. We usually solved them by removing and reataching fresh electrodes. The fact that the electrodes are stuck to the face may be regarded as inconvenient. In the post-experiment questionnaire the participants reported that they did not feel physically constrained by the electrodes, sensors, or wires. It is clear, however, that for long-term use a more comfortable and robust solution, such as the wearable EOG goggles described in Bulling et al. [2009], is desirable.

Baseline drift is perhaps an unavoidable problem for EOG recordings. It is for this reason that accurate gaze tracking, for purposes such as target detection, might be difficult to achieve using wearable EOG. By analyzing the dynamics of eye-movements, however, we can detect activities (such as reading) without the need for accurate gaze tracking. It is also important to note that the recognition methodology presented here is not limited to EOG. All eye-movement features could be extracted equally well from eye-movement data recorded using a video-based eye tracker.

7.3 On the Experiment

The Wii-remote proved to be a useful annotation tool, and was certainly preferable to video-based offline annotation. This method is subject to inaccuracies when, for example, the assistant is distracted, or when buttons are pressed and released too early or too late. However, labeling errors are an intrinsic problem, especially in mobile settings, and a satisfying solution has not been found yet.

In the questionnaire, all participants declared that they did not feel distracted by people in the street and were only partially aware of the experiment assistant. Half of the participants did report a feeling of unease while reading and walking. This unease could clearly be seen in the EOG signal by the occasional presence of small vertical saccades during reading whenever a participant looked up from the text to check the way ahead.

Ideally, the most natural scenario would have involved recordings over a period of days or weeks. This would allow us to better study the reading behavior of our participants and to open up interesting questions regarding daily reading habits. Unfortunately, the battery lifetime of our recording equipment limited recordings to a few hours. Therefore, the main improvement for future studies is to use equipment that does not impose such restrictions but allows for long-term eye-movement recordings.

7.4 Limitations

One limitation of the current work is the assumption that reading only occurs during “head down” periods. Our results show that these head movements do provide valuable information for reading recognition in the current scenario. However, the assumption needs to be validated in other scenarios, particularly those involving shorter reading sequences. Although half of our participants reported to occasionally reading the newspaper while in transit, it was more common for people to read shorter texts, such as advertisements, road signs, or timetables. As short reading sequences may not always involve similar head movements, the proposed segmentation approach will likely need to be adapted.

The study also reveals some of the complexity we might face in using the eyes to detect a person’s reading activity. The ubiquity of the eye involvement in everything a person does means that it is challenging to annotate precisely when a person is reading. It is challenging, too, to identify brief reading events and to separate relevant eye movements from momentary distractions, such as reading while walking. These problems may be solved, in part, by using video and gaze tracking for annotation. Reading activity could also be studied at larger timescales to perform behavioral analysis rather than activity recognition. Annotation will still be an issue, but one that may be alleviated using unsupervised or self-labeling methods [Huynh et al. 2008; Bao and Intille 2004].

7.5 Considerations for Future Work

Additional modalities that are potentially useful for reading recognition - such as hand gestures or upper-body postures - were not investigated in this work. Such modalities are still worth investigating in the future as they may carry information that complements that derived from head movements. It will be interesting to see whether a similar segmentation approach can be used with these modalities.

This may, for example, allow us to detect the hand movements of a person taking a mobile phone out of the pocket to spot the onset of the person's subsequent activity of reading a text message.

Using multiple modalities will require to investigate more complex fusion approaches than simple thresholding. For example, hand gestures or body movements could be recognised separately and the corresponding classifier outputs be fused on the classifier level with that of an eye movement classifier. The light-weight string matching algorithm could not only be extended to support both eye movement components but also to support these additional modalities, e.g. to detect typical eye-motor coordination activities.

Finally, eye movements are also linked to a number of cognitive processes of visual perception - such as visual memory, learning, or attention - and are therefore often called "a window to mind and brain". If it were possible to infer such processes from eye movements, this may lead to cognition-aware systems - systems that are able to sense and adapt to a person's cognitive state [Bulling et al. 2011].

8. CONCLUSION

In this work we have shown that multimodal fusion of information derived from eye and head movements is a robust approach for recognizing reading activity in daily life scenarios across different people. The proposed method of exploiting the sensorimotor coordination of eye and head movements during reading increased the recognition performance considerably. This raises the question of whether different reading behaviors and attention levels to written text can be detected automatically. The proposed segmentation approach is computationally lightweight and paves the way for developing robust end applications that include reading recognition. For example, a "reading detector" could enable novel attentive user interfaces which take into account aspects such as user interruptibility or the level of task engagement.

REFERENCES

BANNACH, D., LUKOWICZ, P., AND AMFT, O. 2008. Rapid prototyping of activity recognition applications. *IEEE Pervasive Comput. 7*, 2, 22–31.

BAO, L. AND INTILLE, S. S. 2004. Activity recognition from user-annotated acceleration data. In *Proceedings of the 2nd International Conference on Pervasive Computing*. Springer, Berlin, 1–17.

BAREA, R., BOQUETE, L., MAZO, M., AND LOPEZ, E. 2002. System for assisted mobility using eye movements based on electrooculography. *IEEE Trans. Neural Syst. Rehab. Eng. 10*, 4, 209–218.

BIEDERT, R., BUSCHER, G., SCHWARZ, S., HEES, J., AND DENGEL, A. 2010. Text 2.0. In *Extended Abstracts of the SIGCHI Conference on Human Factors in Computing Systems.* ACM, New York, 4003–4008.

BROWN, M., MARMOR, M., AND VAEGAN. 2006. ISCEV standard for clinical electro-oculography (EOG). *Documenta Ophthalmologica 113*, 3, 205–212.

BULLING, A., ROGGEN, D., AND TRÖSTER, G. 2009. Wearable EOG goggles: Seamless sensing and context-awareness in everyday environments. J. *Ambient Intell. Smart Environ.1*, 2, 157–171.

BULLING, A., ROGGEN, D., AND TRÖSTER, G. 2011. What's in the eyes for context-awareness? *IEEE Pervasive Comput. 10*, 2, 48–57.

BULLING, A., WARD, J. A., GELLERSEN, H., AND TRÖSTER, G. 2008. Robust recognition of reading activity in transit using wearable electrooculography. In *Proceedings of the 6th International Conference on Pervasive Computing*, Springer, Berlin. 19–37.

BULLING, A., WARD, J. A., GELLERSEN, H., AND TRÖSTER, G. 2011. Eye movement analysis for activity recognition using electrooculography. *IEEE Trans. Patt. Anal. Machine Intell. 33*, 4, 741–753.

CAMPBELL, C. S. AND MAGLIO, P. P. 2001. A robust algorithm for reading detection. In *Proceedings of the Workshop on Perceptive User Interfaces*. ACM, New York, 1–7.

CANOSA, R. L. 2009. Real-world vision: Selective perception and task. *ACM Trans. Appl. Percept. 6*, 2, 1–34.

CHEN, Y. AND NEWMAN, W. S. 2004. A human-robot interface based on electro-oculography. In *Proceedings of the International Conference on Robotics and Automation*. Vol. 1, 243–248.

CRAMMER, K. AND SINGER, Y. 2003. Ultraconservative online algorithms for multiclass problems. *J. Mach. Learn. Res. 3*, 951–991.

DAVIES, N., SIEWIOREK, D. P., AND SUKTHANKAR, R. 2008. Special issue: Activity-based computing. *IEEE Pervasive Comput. 7*, 2.

Dempere-Marco, L., Hu, X., MacDonald, S. L. S., Ellis, S. M., Hansell, D. M., and Yang, G.-Z. 2002. The use of visual search for knowledge gathering in image decision support. *IEEE Trans. Syst. Man. Cybern. 22*, 3, 741–754.

Ding, Q., Tong, K., and Li, G. 2005. Development of an EOG (electro-oculography) based human-computer interface. In *Proceedings of the 27th Annual International Conference of the Engineering in Medicine and Biology Society*. 6829–6831.

Duchowski, A. T. 2007. *Eye-Tracking Methodology: Theory and Practice*. Springer, Berlin.

Elhelw, M., Nicolaou, M., Chung, A., Yang, G.-Z., and Atkins, M. S. 2008. A gaze-based study for investigating the perception of visual realism in simulated scenes. *ACM Trans. Appl. Percept. 5*, 1, 1–20.

Hacisalihzade, S. S., Stark, L. W., and Allen, J. S. 1992. Visual perception and sequences of eye movement fixations: A stochastic modeling approach. *IEEE Trans. Syst. Man Cybern. 22*, 3, 474–481.

Hayhoe, M. M. and Ballard, D. H. 2005. Eye movements in natural behavior. *Trends Cognitive Sci. 9*, 188–194.

Henderson, J. M. 2003. Human gaze control during real-world scene perception. *Trends Cognitive Sci. 7*, 11, 498–504.

Huynh, T., Fritz, M., and Schiele, B. 2008. Discovery of activity patterns using topic models. In *Proceedings of the 10th International Conference on Ubiquitous Computing*. ACM, New York, 10–19.

Ji, Q. and Yang, X. 2002. Real-time eye, gaze, and face pose tracking for monitoring driver vigilance. *Real-Time Imaging 8*, 5, 357–377.

Johansson, R. S., Westling, G., Backstrom, A., and Flanagan, J. R. 2001. Eye-hand coordination in object manipulation. *J. Neurosci. 21*, 17, 6917–6932.

Karson, C. N., Berman, K. F., Donnelly, E. F., Mendelson, W. B., Kleinman, J. E., and Wyatt, R. J. 1981. Speaking, thinking, and blinking. *Psych. Res. 5*, 3, 243–246.

Keat, F. T., Ranganath, S., and Venkatesh, Y. V. 2003. Eye gaze based reading detection. In *Proceedings of the Conference on Convergent Technologies for Asia-Pacific Region.* Vol. 2. 825–828.

Kern, N., Schiele, B., and Schmidt, A. 2007. Recognizing context for annotating a live life recording. *Personal Ubiquit. Comput. 11*, 4, 251–263.

Levenshtein, V. I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Phys. Doklady 10*, 8, 707–710.

Lin, C.-J. 2008. LIBLINEAR - A library for large linear classification. http://www.csie.ntu.edu.tw/∼cjlin/liblinear/.

Liversedge, S. P. and Findlay, J. M. 2000. Saccadic eye movements and cognition. *Trends Cognitive Sci. 4*, 1, 6–14.

Logan, B., Healey, J., Philipose, M., Tapia, E., and Intille, S. S. 2007. A long-term evaluation of sensing modalities for activity recognition. In *Proceedings of the 9th International Conference on Ubiquitous Computing.* ACM, New York, 483–500.

Maglio, P. P., Matlock, T., Campbell, C. S., Zhai, S., and Smith, B. 2000. Gaze and speech in attentive user interfaces. In *Proceedings of the 3rd International Conference on Advances in Multimodal Interfaces*. Vol. 1948, 1–7.

Manabe, H. and Fukumoto, M. 2006. Full-time wearable headphone-type gaze detector. In *Extended Abstracts of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, 1073–1078.

Manor, B. R. and Gordon, E. 2003. Defining the temporal threshold for ocular fixation in free-viewing visuo-cognitive tasks. *J. Neurosci. Meth. 128*, 1–2, 85–93.

Mitra, S. and Acharya, T. 2007. Gesture recognition: A survey. *IEEE Trans. Syst. Man Cybern. Part C: Appl. Rev. 37*, 3, 311–324.

Najafi, B., Aminian, K., Paraschiv-Ionescu, A., Loew, F., Bula, C. J., and Robert, P. 2003. Ambulatory system for human motion analysis using a kinematic sensor: monitoring of daily physical activity in the elderly. *IEEE Trans. Biomed. Eng. 50*, 6, 711–723.

Pelz, J. B., Hayhoe, M. M., and Loeber, R. 2001. The coordination of eye, head, and hand movements in a natural task. *Experim. Brain Res. 139*, 3, 266–277.

Peng, H. 2008. mRMR feature selection toolbox for MATLAB. http://penglab.janelia.org/proj/mRMR/.

Peng, H., Long, F., and Ding, C. 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Patt. Anal. Mach. Intell. 27*, 8, 1226–1238.

Rayner, K. 1998. Eye movements in reading and information processing: 20 years of research. *Psychol. Bull. 124*, 3, 372–422.

Sailer, U., Flanagan, J. R., and Johansson, R. S. 2005. Eye-hand coordination during learning of a novel visuomotor task. *J. Neuroscience 25*, 39, 8833–8842.

Salvucci, D. D. and Anderson, J. R. 2001. Automated eye-movement protocol analysis. *Human-Comput. Interact. 16*, 1, 39–86.

Schiffman, H. R. 2001. *Sensation and Perception: An Integrated Approach* 5th Ed., Wiley, New York.

Schleicher, R., Galley, N., Briest, S., and Galley, L. 2008. Blinks and saccades as indicators of fatigue in sleepiness warnings: Looking tired? *Ergonomics 51*, 7, 982–1010.

Sibert, J. L., Gokturk, M., and Lavine, R. A. 2000. The reading assistant: Eye gaze triggered auditory prompting for reading remediation. In *Proceedings of the 13th Symposium on User Interface Software and Technology.* ACM, New York, 101–107.

TINATI, M. A. AND MOZAFFARY, B. 2006. A wavelet packets approach to electrocardiograph baseline drift cancellation. *Int. J. Biomed. Imaging* Article ID 97157.

TURAGA, P., CHELLAPPA, R., SUBRAHMANIAN, V. S., AND UDREA, O. 2008. Machine recognition of human activities: A survey. *IEEE Trans. Circuits Syst. Video Technol. 18*, 11, 1473–1488.

VEHKAOJA, A. T., VERHO, J. A., PUURTINEN, M. M., NOJD, N. M., LEKKALA, J. O., AND HYTTINEN, J. A. 2005. Wireless head cap for EOG and facial EMG measurements. In *Proceedings of the 27th Annual International Conference of the Engineering in Medicine and Biology Society*. 5865–5868.

WARD, J. A., LUKOWICZ, P., AND GELLERSEN, H. 2011. Performance metrics for activity recognition. *ACM Trans. Intell. Syst. Technol. 2*, 1, Article 6.

WARD, J. A., LUKOWICZ, P., TRÖSTER, G., AND STARNER, T. E. 2006. Activity recognition of assembly tasks using body-worn microphones and accelerometers. *IEEE Trans. Patt. Anal. Mach. Intell. 28*, 10, 1553–1567.

WIDDEL, H. 1984., Operational problems in analysing eye movements. In *Theoretical and Applied Aspects of Eye Movement Research*, Elsevier, Amsterdam, 21–29.

WIJESOMA, W. S., WEE, K. S., WEE, O. C., BALASURIYA, A. P., SAN, K. T., AND SOON, K. K. 2005. EOG based control of mobile assistive platforms for the severely disabled. In *Proceedings of the International Conference on Robotics and Biomimetics*. 490–494.

YOUNG, S. 2010. Cognitive user interfaces. *IEEE Signal Process. 27*, 3, 128–140.

Received October 2010; revised January 2011; accepted January 2011