Eye Movement Analysis for Activity Recognition

Andreas Bulling[†], Jamie A. Ward[‡], Hans Gellersen[‡] and Gerhard Tröster[†]

[†] ETH Zurich
Wearable Computing Laboratory
andreas.bulling@acm.org

[‡] Lancaster University
Computing Department

ABSTRACT

In this work we investigate eye movement analysis as a new modality for recognising human activity. We devise 90 different features based on the main eye movement characteristics: saccades, fixations and blinks. The features are derived from eye movement data recorded using a wearable electrooculographic (EOG) system. We describe a recognition methodology that combines minimum redundancy maximum relevance feature selection (mRMR) with a support vector machine (SVM) classifier. We validate the method in an eight participant study in an office environment using five activity classes: copying a text, reading a printed paper, taking hand-written notes, watching a video and browsing the web. In addition, we include periods with no specific activity. Using a person-independent (leave-one-out) training scheme, we obtain an average precision of 76.1% and recall of 70.5% over all classes and participants. We discuss the most relevant features and show that eye movement analysis is a rich and thus promising modality for activity recognition.

Author Keywords

Eye Movement Analysis, Activity Recognition, Electrooculography (EOG), Wearable Computing

ACM Classification Keywords

I.5.2 Pattern Recognition: Design Methodology–Feature evaluation and selection; I.5.4 Pattern Recognition: Applications–Signal processing

General Terms

Algorithms, Experimentation, Measurement

INTRODUCTION

Activity recognition has been studied by many researchers. A variety of indoor physical activities can be recognised using ambient sensors [21, 12]. Body-worn sensors have been extensively used to recognise activities in mobile and daily life situations [16, 18]. A rich source of information about a person's context yet under-investigated are the movements of the eyes. Eye movement characteristics have the potential to reveal a lot about our daily life. This includes information on visual tasks, such as reading [4], but also on cognitive processes of visual perception, such as attention [20] or saliency determination [15]. Because we use our eyes in almost everything that we do, it is conceivable that eye movements may provide a useful information source for activity recognition.

The aim of this work is to assess the feasibility of recognising human physical activity using eye movement analysis. We see two basic approaches for linking eye movements to physical activity. The first is to define activities using a grammar built upon an alphabet of basic eye movement atoms (e.g. left, right, up and down). However, it is not yet clear how alphabets and grammars of eye movements should be defined - if this is possible at all. The second approach, and the one taken in this work, is to define a set of physical activities and then to attempt to infer these from eye movement data using machine learning techniques.

We first develop a set of 90 features that best describe the eye movement data; some based directly on fundamental eye movement characteristics, others devised to capture particular eye movement dynamics. We then rank and evaluate these features using minimum redundancy maximum relevance feature selection (mRMR) and a support vector machine (SVM) classifier. To evaluate both algorithms on a real-world example we devise an experiment involving a continuous sequence of five physical office activities, plus a period without any specific activity (the NULL class). The activities we investigate are: copying a text, reading a printed paper, taking hand-written notes, watching a video, and browsing the web. We choose these activities for two reasons. Firstly, they are all commonly performed during a typical working day. Secondly, they exhibit interesting eye movement patterns that are both structurally diverse, and that have varying levels of complexity. We believe these activities thus well represent the broad range of activities observable in daily life. We record and annotate an eight participant dataset using wearable electrooculography (EOG). In contrast to commonly used video-based systems, EOG is a cheap method for mobile eye movement recordings; it is both computationally light-weight (no demanding video processing) and relatively unobtrusive (no bulky equipment) [3].

The specific contributions of our work are: (1) an annotated dataset of participants performing five different real-world office activities in a continuous sequence; (2) a new method

UbiComp 2009, Sep 30 – Oct 3, 2009, Orlando, Florida, USA.

for analysing repetitive eye movements using a wordbook encoding scheme; (3) the development of 90 features extracted from eye movement characteristics and geared towards eye-based activity recognition; and (4) an evaluation of these features and their application to activity recognition using mRMR feature selection and SVM classification.

In the remainder of the paper we first survey related work, introduce EOG and describe the main eye movement characteristics. We then detail the signal processing required to detect these characteristics in EOG signals and describe the extracted features and the algorithms for feature selection and classification. Afterwards, we elaborate on the experiment, discuss the results and give an outlook to future work.

RELATED WORK

In recent work, Logan *et al.* aimed at recognising common activities in an indoor setting using a large variety and number of ambient sensors [21]. They found that activities typically performed in the same location could be recognised using only one sensor. They also found that mobile activities such as reading or using the phone were much harder to detect. They concluded that for these activities additional sensor modalities would be needed. Bao *et al.* used body-worn accelerometers to detect physical activities under real-world conditions [2]. Using a decision tree classifier they reported accuracies of over 80%. They discovered that although some activities (e.g. stretching) require person-specific training, most (such as walking and running) could be recognised using independent training. Huynh *et al.* introduced a novel approach for modelling daily routines using data recorded by on-body sensors [16]. They converted sensor data into a series of documents that were mined for activity patterns. In a single-person study using seven days of real-world data they showed that the detected patterns are highly correlated to behaviour and daily routine.

Eye movement analysis has a long history as a tool to investigate visual behaviour. In an early study, Hacisalihzade *et al.* used Markov processes to model visual fixations of observers recognising an object [14]. They transformed fixation sequences into character strings and used the string edit distance to quantify the similarity of eye movement sequences. Elhelw *et al.* used discrete time Markov chains to investigate the sequence of temporal fixations [11]. The goal was to identify salient image features that affect the perception of visual realism. They found that fixation clusters 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 [8]. They modelled the assessment behaviour of two domain experts based on the dynamics of their saccadic eye movements. Salvucci *et al.* evaluated means for automated analysis of eye movement protocols [25]. They described three methods based on sequence-matching and Hidden Markov Models. Their methods were able to interpret eye movements as accurately as human experts but in significantly less time.

All of these studies analysed eye movement characteristics and successfully modelled visual behaviour during specific tasks in an automatic manner. However, eye movement analysis has so far only rarely been used as a modality for activity recognition. In an earlier study, we investigated the recognition of reading activity of people in transit in a variety of daily-life settings [4]. Using a string matching algorithm applied to recorded EOG signals and person-independent training, we were able to achieve recognition rates of up to 80.2%. However, the methodology developed in that study was geared towards recognising only one specific activity.

EYE MOVEMENT ANALYSIS

Wearable Electrooculography

The human eye can be modelled as a dipole with its positive pole at the cornea and its negative pole at the retina. Assuming a stable corneo-retinal potential difference (CRP), the eye is the origin of a steady electric potential field. The corresponding electrical signal, the so-called Electrooculogram (EOG), can be measured using a pair of electrodes placed on the skin at opposite sides of the eye. If the eye moves from the centre position towards the periphery, 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 analysing these changes, eye movements can be tracked. Using two pairs of electrodes appropriately positioned around the eye, two signal components (EOG_h and EOG_v) corresponding to two movement components - a horizontal and a vertical - can be identified.

Baseline drift

Baseline drift is a slow signal change superposing the EOG signal but mostly unrelated to eye movements. It has many possible sources, e.g. interfering background signals or electrode polarisation [13]. Baseline drift only marginally influences the EOG signal during fast eye movements; all other movements are subject to baseline drift. In a five electrode setup, as used in this work, baseline drift may differ between the horizontal and vertical EOG signal components.

Eye Movement Characteristics

To be able to use eye movement analysis for activity recognition, it is important to understand the three main eye movement types: saccades, fixations and blinks (see Figure 1).

Saccades

The eyes do not remain still when viewing a visual scene; they have to move constantly to build up a mental "map" from interesting parts of the 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 simultaneous movement of both 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.

Fixations

A fixation is the static state of the eyes during which gaze is held upon a specific location. Humans typically alternate saccadic eye movements and fixations. The term "fixation" can also be referred to as the time between two saccades during which the eyes are relatively stationary.

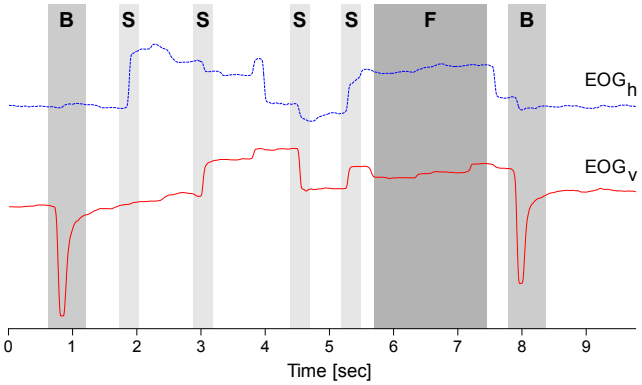

Figure 1. Denoised and baseline drift removed EOG with horizontal (EOG_h) and vertical (EOG_v) signal components. The highlighted segments show the three main types of eye movement characteristic: saccades (S), fixations (F) and blinks (B).

Blinks

The frontal part of the cornea is coated with a thin liquid film, the "precornial tear film". To spread this lacrimal 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 [17]; it is influenced by environmental factors (e.g. relative humidity, temperature, brightness), but also by physical activity, cognitive workload or fatigue [26].

METHOD

The methods in this work were all implemented offline using MATLAB and C. In this section, we first describe the signal processing required for removing noise and baseline drift, and for detecting saccades, fixations and blinks. We then describe the features calculated from these eye movement characteristics during feature extraction. Finally, we briefly introduce the algorithms minimum redundancy maximum relevance (mRMR) for feature selection, and support vector machines (SVM) for classification.

Noise and Baseline Drift Removal

EOG_h and EOG_v were first stripped of high frequency noise using a median filter. For baseline drift removal, we then performed an approximated multilevel 1-D wavelet decomposition at level nine using Daubechies wavelets on each signal component. The reconstructed decomposition coefficients gave a baseline drift estimation. Subtracting this estimation from the original signals yielded the corrected signals with reduced drift offset (see [27] for further details).

Saccade Detection

In an earlier work we introduced the *Continuous Wavelet Transform - Saccade Detection* (CWT-SD) algorithm [4]. CWT-SD detects saccades by thresholding on the continuous 1-D wavelet coefficient vector computed from the denoised and baseline drift removed EOG_h and 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, we extended the original CWT-SD algorithm; an additional step removed all saccade candidates that did not comply to typical physiological saccade characteristics described in literature [9].

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 [29]. In a first step, EOG_h and EOG_v were divided into saccadic and non-saccadic segments using the output from saccade detection. For each non-saccadic segment, the algorithm calculated the corresponding dispersion and duration values. If the dispersion was below a maximum threshold, and the duration above a minimum threshold, a fixation was detected (see [29] for typical values).

Blink Detection

Similar to the algorithm for saccade detection, the so-called *Continuous Wavelet Transform - Blink Detection* (CWT-BD) algorithm used thresholding of wavelet coefficients to detect blinks in EOG_v. In contrast to saccades, a blink is characterised 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 were distinguished from saccades by applying a maximum threshold on this time difference.

Feature Extraction

Three groups of features were extracted based on the detected saccades, fixations and blinks. We developed a fourth feature group that captures sequence information from eye movement patterns using workbooks. Table 1 details the naming scheme used for all features. The features were calculated using a sliding window (window size W_{fe} and step size S_{fe}) on both EOG_h and EOG_v. From a pilot study, we were able to fix W_{fe} at 30 seconds and S_{fe} at 0.25 seconds.

Saccade Features

Ehrlichman *et al.* showed that changes in the saccade rate correlate with task requirements and the type of memory access required to perform these tasks [10]. In this work, features calculated from saccadic eye movements made up the biggest portion of all features extracted. In total, we extracted 62 saccadic features: the mean, variance and maximum signal amplitudes of saccades and normalised saccade rates. All of these features were calculated for both EOG_h and EOG_v, for small and large saccades, for saccades in positive or negative direction, and for all combinations of these.

Fixation Features

Canosa *et al.* showed that for different tasks such as reading, counting, talking, sorting and walking noticeable differences can be identified in the mean fixation duration and the mean saccade amplitude [6]. For each fixation, we calculated five different features: the mean and the variance of the signal amplitude within the fixation; the mean and the variance of the fixation duration, and the fixation rate in the window.

Group	Features
saccade (*S*-)	mean (mean), variance (var) or maximum (max) EOG signal amplitudes (Amp) or rate (rate) of small (S) or large (L), positive (P) or negative (N) saccades in horizontal (Hor) or vertical (Ver) direction
fixation (*F*-)	mean (mean) and/or variance (var) of the horizontal (Hor) or vertical (Ver) EOG signal amplitude (Amp) within or length (Length) of a fixation or rate of fixations
blink (*B*-)	mean (mean) or variance (var) of the blink duration or blink rate (rate)
wordbook (*W*-)	wordbook size (size) or maximum (max), difference (diff) between maximum and minimum, mean (mean) or variance (var) of all occurrence counts (Count) in the wordbook of length (-lx)

Table 1. Naming scheme for the features used in this work. For a particular feature, e.g. *S-rateSPHor*, the capital letter represents the group - saccadic (S), blink (B), fixation (F) or wordbook(W) - and the combination of abbreviations after the dash describes the particular type of feature and the characteristics it covers.

Blink Features

Blink rate inhibition was shown to be a good measure of attentional disposition towards a visual stimuli and thus may reflect visual engagement [22]. Caffier *et al.* found that parameters of spontaneous eye blinks such as the blink duration are influenced by cognitive efforts and can thus serve as a drowsiness measure [5]. We extracted three blink features: blink rate, and the mean and variance of blink duration.

Wordbook Features

To assess repetitive patterns of eye movements, we propose the following wordbook analysis (see Figure 2). First, the sequence of horizontal and vertical saccades in the window is encoded into a combined character stream in which each character represents one eye movement. Our implementation distinguishes between 24 eye movements of different direction and distance. In a second step, a sliding window is used to scan the character stream for repetitive eye movement patterns. A pattern is defined as a sequence of l successive characters. For example, the pattern "LrBd" translates to large left (L) $\rightarrow$ small right (r) $\rightarrow$ large diagonal right (B) $\rightarrow$ small down (d). Each newly found pattern is added to the corresponding workbook Wb_l. For a pattern that is already included in Wb_l, its occurrence count is increased by one. In this work, we analysed eye movement patterns up to a length of four ($l = 4$). Thus, the output of the algorithm were four wordbooks each containing the type and number of all patterns found for a particular length. For each of these wordbooks we extracted 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.

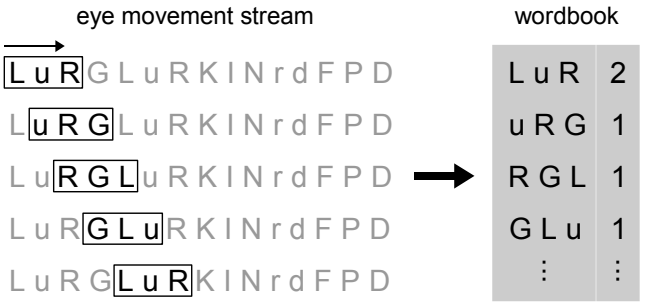

Figure 2. Example wordbook encoding for eye movement patterns of length $l = 3$. A sliding window scans a stream of eye movements encoded as characters for repetitive patterns. Newly found patterns are added to the corresponding wordbook Wb_3; otherwise only the occurrence count is increased by one.

Feature Selection and Classification

For feature selection, we preferred a filter scheme over commonly used wrapper approaches because of the lower computational costs and thus shorter runtime given the large dataset. In this work, we chose minimum redundancy maximum relevance feature selection (mRMR) as described by Peng *et al.* [24]. Briefly, the mRMR algorithm selects a feature subset of arbitrary size S that best characterises the statistical properties of the given target classes based on the ground truth labelling. Amongst the possible underlying statistical measures described in literature, mutual information was shown to yield the most promising results and was thus selected in this work (see [24] for details on the algorithm, and [23] for the MATLAB implementation we used).

For classification, we chose a linear support vector machine. Our SVM implementation used a fast sequential dual method for dealing with multiple classes [7]. Compared to common schemes such as one-versus-all, the sequential approach reduced training time considerably (see [19] for the C implementation we used).

Evaluation and Parameter Selection

For evaluation, we followed a leave-one-person-out scheme: the datasets of all but one participant were combined and used for training; both datasets of the remaining participant were combined and used for testing. This was repeated for each participant. Feature selection was always performed solely on the training set. During the classification process the size of the feature set for each leave-one-person-out iteration was optimised with respect to recognition accuracy by sweeping over S and the SVM cost parameter. In addition, the prediction vector returned by the classifier was smoothed using a sliding majority window. Its main parameter, the window size W_{sm}, was also obtained using a parameter sweep and fixed at 2.4 seconds. All parameters of the saccade, fixation and blink detection algorithms were fixed to values common to all participants.

EXPERIMENT

The experimental setup was designed with two objectives in mind: (1) to unobtrusively record eye movements of people performing a set of activities in a real-world environment,

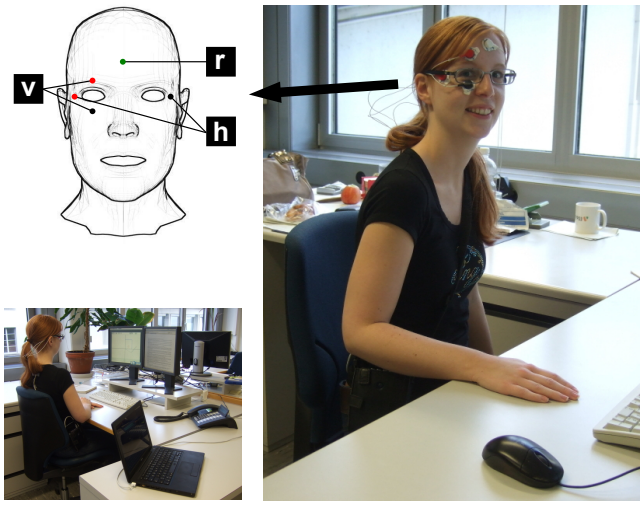

Figure 3. Experimental setup consisting of five electrodes for EOG data collection (h: horizontal, v: vertical, r: reference). The participants' eye movements were recorded while seated at a desk in a real office environment during normal working hours. No constraints with respect to movements of the head and upper body were imposed.

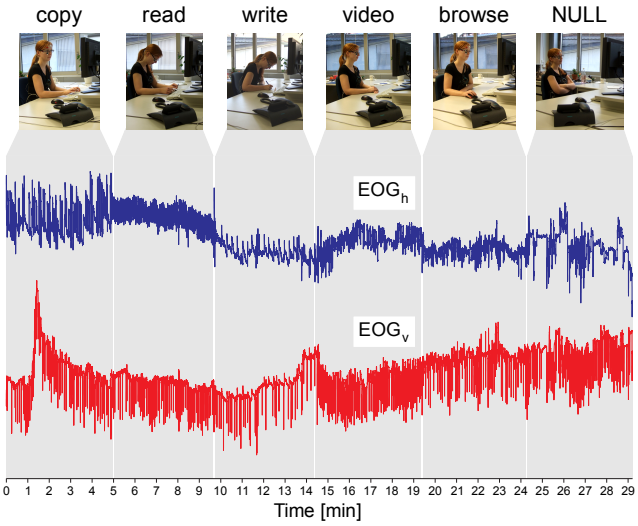

Figure 4. Experimental procedure and corresponding horizontal (EOG_h) and vertical (EOG_v) EOG signals for a continuous sequence of office activities: copying a text, reading a printed paper, taking hand-written notes, watching a video, browsing the web, and a period of no specific activity (the NULL class).

and (2) to use these recordings to analyse and evaluate if and how well such activities can be recognised using eye movement analysis. We chose a scenario with five activities typically performed at a desk during an office working day: copying a text, reading a printed paper, taking hand-written notes, watching a video and browsing the web. An additional NULL class was comprised of any time participants were distracted from their task, and a fixed period in which they took a break.

Participants

We collected data from eight paid participants - six male and two female - recruited from the lab and the authors' friends. Participants were between 23 years and 31 years old (*mean* = 26.1, *sd* = 2.4); all were daily computer users, reporting 6 to 14 hours of use per day (*mean* = 9.5, *sd* = 2.7).

Apparatus

For EOG data collection, we used a commercial system, the Mobi from Twente Medical Systems International (TMSI). The device recorded a four-channel EOG with a sampling rate of 128 Hz. It was worn on a belt around each participant's waist and transmitted aggregated data via Bluetooth. Each participant was observed by an assistant who annotated changes in activity. For annotation, we used the Nintendo Wii controller based on good experiences in an earlier study [4]. Data recording and synchronisation was handled by the Context Recognition Network (CRN) Toolbox [1].

EOG signals were picked up using an array of five 24 mm Ag/AgCl wet electrodes from Tyco Healthcare placed around the right eye (see Figure 3). The horizontal signal was collected using one electrode on the nose and another directly across from this on the edge of the right eye socket. The vertical signal was collected using one electrode above the right eyebrow and another on the lower edge of the right eye socket. The fifth electrode, the signal reference, was placed away from the other electrodes in the middle of the forehead. Five participants (three male, two female) had to wear spectacles during the experiment. For these participants, the nose electrode was moved to the edge of the left eye socket to not interfere with the glasses frame.

The experiment was performed in a real, well-lit office during normal working hours. The participants were seated in front of two 17" flat screens with a resolution of 1280x1024 pixels on which a browser, a video player, a word processor and a text for copying were already opened and ready for use. No head stand was used; free movement of the head and upper body were possible throughout the experiment.

Procedure

Participants were asked to follow two sequences each composed of five different, randomly ordered activities plus a period of no specific activity, the NULL class (see Figure 4 for an example). Each activity lasted about five minutes. Overall, this resulted in eight hours of EOG data comprised of similarly sized fractions for each activity and a NULL class of about one fifth of the total dataset.

For the text copying task, the original document was shown on the right screen while the word processor was opened on the left. The participants were free in the way they copied the text. Some touch typed and only checked for errors in the text from time to time; others continuously switched attention between the screens or the keyboard while typing. Because the screens were more than half a meter from the participants' faces, the video was shown full screen to elicit more distinct eye movements. For the browsing task, no constraints were imposed concerning the type of website or the manner of interaction. For the reading and writing tasks, a book (12 pt, one column with pictures) and a pad with a pen

read	browse	write	video	copy
W-maxCount-l2	S-rateSPHor	W-varCount-l4	F-meanVarVertAmp	S-varAmp
W-meanCount-l4	W-varCount-l4	F-meanVarVertAmp	F-meanVarHorAmp	S-meanAmpSNHor
W-varCount-l2	W-varCount-l3	F-varLength	B-rate	S-meanAmpLPHor
F-varLength	W-varCount-l2	F-meanLength	S-varAmpNHor	S-rateS
B-rate	W-meanCount-l1	S-rateLPVer	S-meanAmpSPHor	F-meanVarHorAmp

Table 2. Top five features selected by mRMR for each activity over all training sets (see Table 1 for a description of each feature).

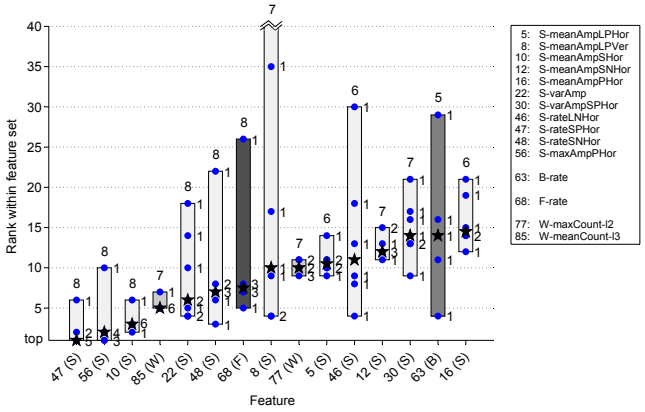

Figure 5. Top 15 features selected by mRMR for all eight training sets. X-axis shows feature number and group; the key on the right shows the corresponding feature names as described in Table 1; Y-axis shows the rank (top = 1). For each feature, the bars show: the total number of training sets for which the feature was chosen (bold number at the top), the rank of the feature within each set (dots, with a number representing the set count), and the median rank over all sets (black star). For example, a useful feature is 47 (S) - a saccadic feature selected for all sets, in 7 of which it is ranked 1 or 2; less useful is 63 (B) - a blink feature used in only 5 sets and ranked between 4 and 29.

were provided. Data was also gathered from a period during which the participants took a break (included in the NULL class). No activity was required of them but they were asked not to engage in any of the other studied activities.

RESULTS

Eye Movement Features

We first analysed how mRMR ranked the features on each of the eight leave-one-person-out training sets. The rank of a feature is the position at which mRMR selected it within a set. The position corresponds to the importance with which mRMR assesses the feature's ability to discriminate between classes in combination with the features ranked before it. Figure 5 shows the top 15 features according to the median rank over all sets (see Table 1 for a description of the type and name of the features). Each vertical bar represents the spread of mRMR ranks: for each feature there is one rank per training set. The most useful features are those found with the highest rank (close to one) for most training sets, indicated by shorter bars. Our classification scheme chose the best recognition accuracy based on a sweep of the number of features for each set. Thus, some features are not included in the final result (e.g. feature 63 only appears in five sets). Equally, a useful feature that is ranked lowly by mRMR might be the one that improves a classification (e.g. feature 68 is spread between rank five and 26, but is included in all eight sets).

This analysis reveals that the top three features, as judged by high ranks for all sets, are all based on horizontal saccades: 47 (*S-rateSPHor*), 56 (*S-maxAmpPHor*) and 10 (*S-meanAmpSHor*). Feature 68 (fixation rate) is used by all sets, seven of which rank it highly. Feature 63 (blink rate) is selected for five out of the eight sets, only one of which gives it a high rank. And although wordbook features 77 (*W-maxCount-l2*) and 85 (*W-maxCount-l3*) are not used in one of the sets, they are highly ranked by the other seven.

We performed an additional study into the effect of optimising mRMR for each activity class. For this, we combined all training sets and performed a one-versus-many mRMR for each non-NULL activity. The top five features selected during this evaluation are shown in Table 2. For example, the table reveals that reading and browsing can be described using wordbook features. Writing additionally requires fixation features. Video is described by a mixture of fixations and saccades in all directions and - as reading - the blink rate, while the copy task involves mainly horizontal saccades.

Classification Performance

The SVM classification was compared to the annotated ground truth. Classification performance was then scored in two ways: using a time-based confusion matrix, and using the additional error categories introduced in [28]. For specific results on each participant, or on each activity, class-relative precision and recall metrics were used. Precision was defined as $\frac{correct_c}{output_c}$ and recall as $\frac{correct_c}{total_c}$ for each class c.

Table 3 shows the average precision and recall for each participant. Here we see a range of differences in recognition performance. The highest performance was achieved for participants six (89.2% precision, 86.9% recall) and seven (93.0% precision, 81.9% recall). The worst results were for participants four (46.6% precision, 47.9% recall) and five (59.5% precision, 46.0% recall). The mean performance over all participants is 76.1% precision and 70.5% recall.

Figure 6 plots the classification results in terms of precision and recall for each activity and participant. The best results approach the top right corner while worst results are close to the lower left. For most activities, precision and recall fall within the top right corner. Participant four, however, completely fails for the reading and copying task and also shows

	P1 (m)	P2 (m)	P3 (m)	P4 (m)	P5(m)	P6 (f)	P7 (f)	P8 (m)	Mean
Precision	76.6	88.3	83.0	**46.6**	59.5	89.2	**93.0**	72.9	76.1
Recall	69.4	77.8	72.2	47.9	**46.0**	**86.9**	81.9	81.9	70.5

Table 3. Precision and recall for each participant using SVM. The gender is given in brackets; best and worst results are indicated in bold.

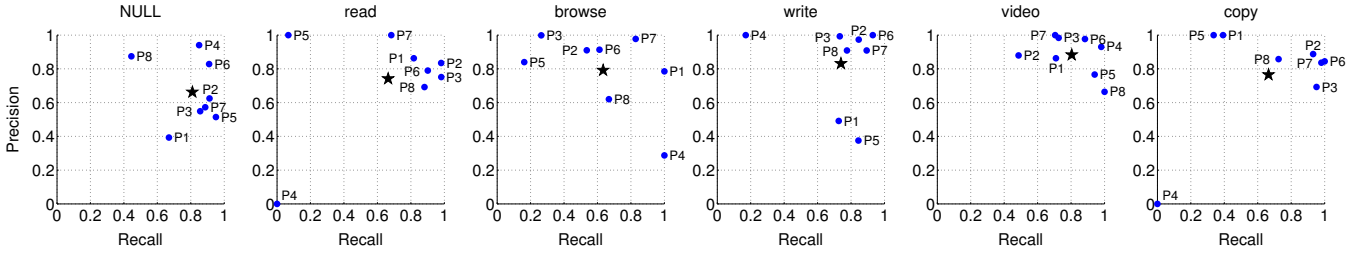

Figure 6. Precision and recall for each activity and participant using SVM. Black stars mark the mean performance over all participants.

noticeably lower precision for browsing. For participant five, similar but less strong characteristics apply for the reading, writing and browsing task.

The summed confusion matrix from all participant sets, normalised across ground truth rows, is given in Figure 7. Correct recognition is shown on the diagonal; substitution errors are off-diagonal. The largest between-class substitution errors, those not involving NULL, fall between 12% and 13% of their respective class times. Most of these errors involve the browse activity, which is falsely returned during 13% each of read, write and copy activities. A similar amount is substituted by read during browse time.

The error division diagram (EDD) of Figure 8 shows an alternative representation of these results. The EDD treats an activity as "positive" and NULL as "negative". The substitution errors from the confusion matrix are summed together, with a more detailed breakdown shown for the errors relating to NULL: false negatives (FN) and false positives (FP). The FN errors include: deletion (activity not detected), fragmenting (fragments of NULL within a correct activity) and underfill (the false NULL time at the beginning and end of a correct activity). The FP errors include: insertion (an activity falsely returned during NULL), merge (the NULL time lost when two activities are detected as one), and overfill (the NULL time lost by a correct activity spilling over its boundaries). We see that 7.1% of the total time is underfill and overfill. These are cases where the fault may be slightly offset labelling, or delays in the recognition system. The errors that might be regarded as more serious - merge, insertion, fragmentation, deletion and substitution - account for 20.1% of the total experiment time.

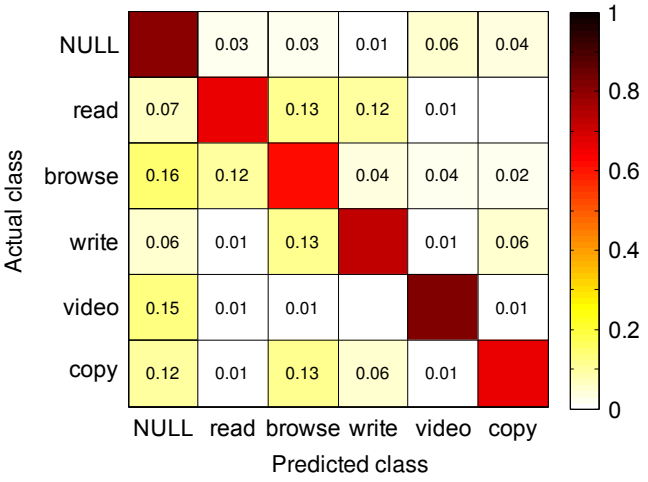

Figure 7. Summed confusion matrix from all participant sets for SVM, normalised across ground truth rows.

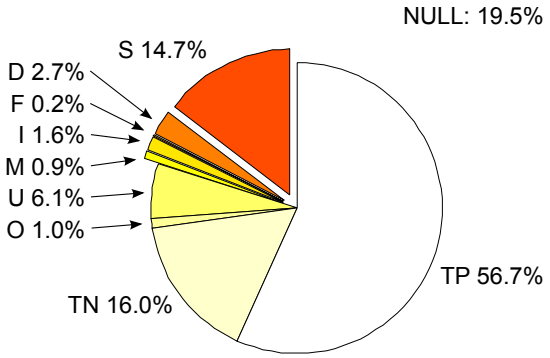

Figure 8. Error division diagram (EDD) for SVM over all participants showing the proportion of the total dataset comprising: true positives (TP), true negatives (TN), overfills (O), underfills (U), merges (M), insertions (I), deletions (D), fragmentations (F), and substitutions (S).

DISCUSSION

Feature Groups

The mRMR-based feature selection presented here provides a snapshot of the types of eye movement features that might be used for activity recognition. Features from three of the four proposed groups - saccade, fixation and wordbook - were all prominently represented in our study. It has to be noted that no-one feature type performs well alone. The best performance results were always obtained using a mixture of different features. Among these, the fixation rate was always selected. This result is akin to that of Canosa *et al.* who

found that both fixation duration and saccade amplitude are strong indicators of certain activities [6].

Features derived from blinks are less well represented in the top ranks. One explanation for this might be that for the short activity duration of only five minutes the participants did not become fully engaged in the tasks, and were thus less likely to show the characteristic blink rate variations suggested by Palomba *et al.* [22]. These features may be found to be more discriminative for longer duration activities. Coupled with the ease by which they were calculated, we believe blink features are still very promising for future work.

The wordbook encoding scheme introduced in this work produced two top ranking features for all but one of the participants. Both features describe short sequences of only two and three successive eye movements. This indicates the existence of underlying eye movement atoms, and might be further explored as a basis for a grammar-based approach to eye-based activity recognition.

Features for Each Activity

The analysis of the best features for each activity class is particularly revealing.

Reading is a regular pattern characterised by a very specific sequence of saccades and short fixations of similar duration. Consequently, mRMR chose mostly wordbook features describing eye movement sequencing in its top ranks, as well as a feature describing the fixation length variance. The fifth feature, the blink rate, reflects that for reading as an activity of high visual engagement people tend to blink less [22].

Browsing is a highly unstructured activity that - depending on the website being viewed - may be comprised of different activities, e.g. watching a video, typing or looking at a picture. In addition to the small, horizontal saccade rate, mRMR also selected several workbook features of varying lengths. This is probably due to our participants' browsing activities containing mostly sequences of variable length reading such as scanning headlines or searching for a product in a list.

The writing activity is similar to reading, but requires greater fixation duration (it takes longer to write a word than to read it) and greater variance. mRMR correspondingly selected average fixation length and its variance as well as a wordbook feature. However, this activity is also characterised by short thinking pauses, during which people invariably look up. This corresponds extremely well to the choice of the fixation feature that captures variance in vertical position.

Watching a video is a completely unstructured activity, but is carried out within a narrow field of view. The lack of a wordbook feature reflects this, as does the mixed selection of features based on all three types: variance of both horizontal and vertical fixation positions, small positive and negative saccadic movements, and blink rate. The use of blink rate likely reflects the tendency towards blink inhibition when performing an engaging yet sedentary task [22].

Finally, the copy task involves many back and forth saccades between screens. mRMR reflects this by choosing as its top selection a mixture of small and large horizontal saccade features, as well as variance in horizontal fixation positions.

Robustness Across Participants

All parameters of the saccade, fixation and blink detection algorithms were fixed to values common to all participants; the same applies to the parameters of the feature selection and classification algorithms. Despite person-independent training, six out of the eight participants returned best average precision and recall values of between 69% and 93% using the SVM classifier. However, two participants returned results that were lower than 50%. Participant four had zero correct classifications for both reading and copying, and close to zero recall for writing; participant five had close to zero recall for reading and browsing. On closer inspection of the raw EOG data, it turned out that in both cases the signal quality was much worse compared to the others. The signal amplitude changes for saccades and blinks - upon which feature extraction and thus classification performance heavily depend - were not distinctive enough to be reliably detected. As was found in an earlier study [4], dry skin or poor electrode placement are the most likely culprits.

Results for Each Activity

As might have been expected, reading is detected with comparable accuracy to that reported previously [4]. However, the methods used are quite different. The string matching approach used in the earlier study makes use of a specific "reading pattern". That approach is not suited for activities involving less homogeneous eye movement patterns. For example, one could not expect to find a similarly unique pattern for browsing or watching a video as there exists for reading. This is because eye movements show much more variability during these activities as they are driven by an ever-changing stimulus. As shown in this work, the feature-based approach is much more flexible and scales better with the number and type of activities that are to be recognised.

Accordingly, we are now able to recognise four additional activities - web browsing, writing on paper, watching video, and copying text - with almost, or above, 70% precision and 70% recall. Particularly impressive is video, which is recognised with an average precision of 88% and recall of 80%. This is indicative of a task where the user might be concentrated on a relatively small field of view (like reading), but follows a more or less unstructured path (unlike reading). Similar examples outside the current study might include interacting with a graphical user interface or watching television at home. Writing is very similar to reading in that the eyes follow a structured path, albeit at a slower rate. It involves more eye "distractions" - when the person looks up to think, for example. Browsing is recognised less well over all participants (average precision 79%, recall 63%) - but with a large spread between people. A likely reason for this is that it is not only unstructured, but that it involves a variety of sub-activities - including reading - that may need to be modelled. The copy activity, with an average precision of 76% and a recall of 66%, is representative of activities with

a small field of view that include regular shifts in attention (in this case to another screen). A comparable activity outside the chosen office scenario might be driving, where the eyes are on the road ahead with occasional checks to the side mirrors. Three of these activities, writing, copy and browse, all include sections of reading. From quick checks over what has been written or copied, to longer perusals of online text, reading is a pervasive sub-activity in this scenario. This is confirmed by the relatively high rate of substitution errors involving reading shown in Figure 7.

Finally, the NULL class returns a high recall of 81%. However, there are many false returns (activity false negatives) for half of the participants, resulting in a precision of only 66%. Closer inspection of Figure 8 indicates that many of the false NULL results (6.1% of the total dataset) are in fact attributed to underfill - an error that is recorded when an activity class is correctly detected, but is detected after the start and before the end of its corresponding annotation. With a more accurate annotation scheme that removes most of these underfill errors - for example based on video recordings - we believe a NULL precision of around 80% is possible.

Limitations and Considerations for Future Work

Although the experimental scenario in this work only considered a handful of specific activities within an office setting, the study does reveal a number of very useful findings for the general problem of activity and context recognition using eye movement analysis.

Firstly, that eye movement analysis alone, i.e. without any information on gaze, can serve as an alternative sensing modality for recognising human activity. So far, only reading activity was studied in this way [4]. One of the results from the current work is the successful recognition of four additional activities. We view this as an initial proof of concept towards recognising more general activities. We believe that the feature set and recognition methodology developed here are equally applicable to a more general recognition problem. For recording eye movements, we chose EOG over common video-based eye trackers because of its simpler signal processing and thus potentially longer runtime. This is crucial with a view to long-term recordings in mobile settings. It is important to note, however, that the current approach is not limited to EOG. All of the features described previously can be extracted equally well from eye movement data recorded using a video-based eye tracker.

This leads to the second main finding. Good recognition results are achieved by using a mixture of features based on the fundamentals of eye movements. Sequence information on eye movement patterns, in the form of a wordbook encoding, also proved very useful and can probably be extended to capture additional statistical properties. Different recognition tasks will likely require different combinations of these features. For this reason, we recommend that a large number of features based on a mixture of all of these feature types be considered initially for each new task. Additional features such as pupil diameter may lead to further improved recognition performance.

Finally, the study reveals some of the complexity one might face in using the eyes as a source of information on user context. The ubiquity of the eyes' involvement in everything a person does means that it is challenging to annotate precisely what is being "done" at any one time. It is also a challenge to define a single identifiable activity. The reading task is perhaps one of the easiest to capture because of the intensity of eye focus that is required and the well defined paths that the eyes follow. A task such as web browsing is more difficult because of the wide variety of different eye movements involved. It is challenging, too, to separate relevant eye movements from momentary distractions.

These problems may be solved, in part, by an annotation process that uses video and precise gaze tracking. Activities from the current scenario could be redefined at a smaller time scale, breaking web-browsing into smaller activities such as "use scrollbar", "read", "look at image", "type", and so on. This would also allow us to investigate more complicated activities outside the office. An alternative route would be to study activities at larger time scales, to perform situation analysis rather than recognition of specific activities. Longer term eye movement features, for example the average eye movement velocity and blink rate over one hour, might be useful in revealing whether a person is walking along an empty or busy street, whether they are at their desk working, or whether they are at home watching television. Again, annotation will be an issue, but one that may be alleviated using unsupervised or self-labelling methods [16, 2].

The ways our eyes move in daily life are an indicator for what we do. Moreover, they are linked to cognitive processes of visual perception. Thus, if it were possible to infer cognitive behaviour from eye movement - such as memory, learning or attention - this might add a new cognitive dimension to the common understanding of context-awareness.

CONCLUSION

In this work we proposed the use of eye movement analysis as a novel modality for the recognition of physical activity. We devised 90 features specifically geared towards capturing a wide variety of eye movement characteristics. Using wearable EOG recordings from an eight participant study, we showed that we can recognise five different physical office activities from a continuous sequence.

The importance of these findings lies in their fundamental significance for eye movement analysis to become a general tool for the recognition of human activity. The developed feature set and recognition methodology are not limited to the chosen setting, activities or eye tracking equipment. Instead, the current work shows that eye movement analysis has the potential to be successfully applied to many other activity recognition problems in a variety of different settings and for a broad range of visual and physical activities.

REFERENCES

1. D. Bannach, P. Lukowicz, and O. Amft. Rapid Prototyping of Activity Recognition Applications. *IEEE Pervasive Computing*, 7(2):22–31, 2008.

2. L. Bao and S. S. Intille. Activity Recognition from User-Annotated Acceleration Data. In *Proc. Pervasive 2004*, pages 1–17. Springer, 2004.

3. A. Bulling, D. Roggen, and G. Tröster. Wearable EOG goggles: Seamless sensing and context-awareness in everyday environments. *Journal of Ambient Intelligence and Smart Environments (JAISE)*, 1(2):157–171, 2009.

4. A. Bulling, J. A. Ward, H. Gellersen, and G. Tröster. Robust Recognition of Reading Activity in Transit Using Wearable Electrooculography. In *Proc. Pervasive 2008*, pages 19–37. Springer, 2008.

5. P. P. Caffier, U. Erdmann, and P. Ullsperger. Experimental evaluation of eye-blink parameters as a drowsiness measure. *European Journal of Applied Physiology*, 89(3-4):319–325, 2003.

6. R. L. Canosa. Real-world vision: Selective perception and task. *ACM Transactions on Applied Perception*, 6(2):1–34, 2009.

7. K. Crammer and Y. Singer. Ultraconservative online algorithms for multiclass problems. *Journal of Machine Learning Research*, 3:951–991, 2003.

8. L. Dempere-Marco, X. Hu, S. L. S. MacDonald, S. M. Ellis, D. M. Hansell, and G.-Z. Yang. The use of visual search for knowledge gathering in image decision support. *IEEE Transactions on Medical Imaging*, 21(7):741–754, 2002.

9. A. T. Duchowski. *Eye Tracking Methodology: Theory and Practice*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2007.

10. H. Ehrlichman, D. Micic, A. Sousa, and J. Zhu. Looking for answers: Eye movements in non-visual cognitive tasks. *Brain and Cognition*, 64(1):7 – 20, 2007.

11. M. Elhelw, M. Nicolaou, A. Chung, G.-Z. Yang, and M. S. Atkins. A gaze-based study for investigating the perception of visual realism in simulated scenes. *ACM Transactions on Applied Perception*, 5(1):1–20, 2008.

12. J. Fogarty, C. Au, and S. E. Hudson. Sensing from the basement: a feasibility study of unobtrusive and low-cost home activity recognition. In *Proc. UIST 2006*, pages 91–100. ACM Press, 2006.

13. J. J. Gu, M. Meng, A. Cook, and G. Faulkner. A study of natural eye movement detection and ocular implant movement control using processed EOG signals. In *Proc. ICRA 2001*, volume 2, pages 1555–1560, 2001.

14. S. S. Hacisalihzade, L. W. Stark, and J. S. Allen. Visual perception and sequences of eye movement fixations: a stochastic modeling approach. *IEEE Transactions on Systems, Man and Cybernetics*, 22(3):474–481, 1992.

15. J. M. Henderson. Human gaze control during real-world scene perception. *Trends in Cognitive Sciences*, 7(11):498–504, 2003.

16. T. Huynh, M. Fritz, and B. Schiele. Discovery of Activity Patterns using Topic Models. In *Proc. UbiComp 2008*, pages 10–19. ACM Press, 2008.

17. C. N. Karson, K. F. Berman, E. F. Donnelly, W. B. Mendelson, J. E. Kleinman, and R. J. Wyatt. Speaking, thinking, and blinking. *Psychiatry Research*, 5(3):243–246, 1981.

18. N. Kern, B. Schiele, and A. Schmidt. Recognizing context for annotating a live life recording. *Personal and Ubiquitous Computing*, 11(4):251–263, 2007.

19. C.-J. Lin. LIBLINEAR - a library for large linear classification, 2008. `http://www.csie.ntu.edu.tw/~cjlin/liblinear/`.

20. S. P. Liversedge and J. M. Findlay. Saccadic eye movements and cognition. *Trends in Cognitive Sciences*, 4:6–14, 2000.

21. B. Logan, J. Healey, M. Philipose, E. Tapia, and S. S. Intille. A Long-Term Evaluation of Sensing Modalities for Activity Recognition. In *Proc. UbiComp 2007*, pages 483–500. ACM Press, 2007.

22. D. Palomba, M. Sarlo, A. Angrilli, A. Mini, and L. Stegagno. Cardiac responses associated with affective processing of unpleasant film stimuli. *International Journal of Psychophysiology*, 36(1):45 – 57, 2000.

23. H. Peng. mRMR Feature Selection Toolbox for MATLAB, 2007. `http://research.janelia.org/peng/proj/mRMR/`.

24. H. Peng, F. Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, 2005.

25. D. D. Salvucci and J. R. Anderson. Automated eye-movement protocol analysis. *Human-Computer Interaction*, 16(1):39–86, 2001.

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

27. M. A. Tinati and B. Mozaffary. A wavelet packets approach to electrocardiograph baseline drift cancellation. *International Journal of Biomedical Imaging*, Article ID 97157, 2006.

28. J. A. Ward, P. Lukowicz, and G. Tröster. Evaluating performance in continuous context recognition using event-driven error characterisation. In *Proc. LoCA 2006*, pages 239–255, 2006.

29. H. Widdel. *Theoretical and Applied Aspects of Eye Movement Research*, chapter Operational problems in analysing eye movements, pages 21–29. Elsevier, 1984.