EyeContext: Recognition of High-level Contextual Cues from Human Visual Behaviour

Andreas Bulling
Max Planck Institute for
Informatics
Saarbrücken, Germany
andreas.bulling@acm.org

Christian Weichel
Lancaster University
Lancaster, United Kingdom
c.weichel@lancaster.ac.uk

Hans Gellersen
Lancaster University
Lancaster, United Kingdom
hwg@comp.lancs.ac.uk

ABSTRACT

In this work we present *EyeContext*, a system to infer high-level contextual cues from human visual behaviour. We conducted a user study to record eye movements of four participants over a full day of their daily life, totalling 42.5 hours of eye movement data. Participants were asked to self-annotate four non-mutually exclusive cues: *social* (interacting with somebody vs. no interaction), *cognitive* (concentrated work vs. leisure), *physical* (physically active vs. not active), and *spatial* (inside vs. outside a building). We evaluate a proof-of-concept EyeContext system that combines encoding of eye movements into strings and a spectrum string kernel support vector machine (SVM) classifier. Our results demonstrate the large information content available in long-term human visual behaviour and opens up new venues for research on eye-based behavioural monitoring and life logging.

Author Keywords

Context Recognition; Eye Movement Analysis; Visual Behaviour; Electrooculography (EOG)

ACM Classification Keywords

H.5.m Information Interfaces and Presentation (e.g., HCI): Miscellaneous; I.5.4 Pattern Recognition: Applications — *Signal processing*

INTRODUCTION

Practically everything we do in our lives involves our eyes, and the way we move our eyes is linked to our goals and tasks. This makes the eyes a particularly rich source of information: one that, as we will show in this work, can provide basic cues on very different aspects of what we do, at any point in time. Figure 1 illustrates our idea: to provide a system that is able to produce diverse inferences, about social, cognitive, physical and spatial aspects, all from eye movement as single source of information.

Our contribution is two-fold. First, we introduce the *EyeContext* system for cue inference from eye movement. The system takes continuous eye movement as input, and produces a vector of cues as output. The cues are binary descriptors of daily life situations at any given time. In our proof-of-concept system, the cues describe whether or not we: socially interact; concentrate on a mental task; engage in physical activity; are inside or outside. At the core of our system, we introduce a novel method for inferring such cues from eye movement. We encode eye movement as string of symbols that represent movements in different directions. Patterns of successive movements thus become represented by words of different lengths, which we use as basis for our binary classification problems. The underlying hypothesis is that we find different patterns for when we interact or not; are inside or outside; etc. We use a string kernel support vector machine (SVM) for classification, inspired by their original use in bioinformatics for efficient large-scale protein sequence classification [7].

Figure 1. The *EyeContext* system infers contextual cues about different aspects of what we do, by analysing eye movement patterns over time (from left to right): *social* (interacting with somebody vs. no interaction), *cognitive* (concentrated work vs. leisure), *physical* (physically active vs. not active), and *spatial* (inside vs. outside a building).

Our second contribution is an evaluation of the system for which we collected eye movements of four participants over a typical daily life from morning to evening. Participants self-annotated the four cues of interest as our ground truth reference. Using person-dependent training, we assessed the recognition performance for each of our cues. The results validate the EyeContext system but moreover provide evidence that eye movement holds contextual information about very diverse aspects of our daily life. Each individual cue, for example whether we are engaging in physical activity, might be approached with other means (e.g. body motion [2]), while eye movement can provide cues ranging from social to cognitive to physical and spatial.

THE EYECONTEXT SYSTEM

The EyeContext system takes continuous eye movement signals as its input. In this work, we used electrooculography

CHI 2013, April 27–May 2, 2013, Paris, France.
Copyright 2013 ACM 978-1-4503-1899-0/13/04...$15.00.

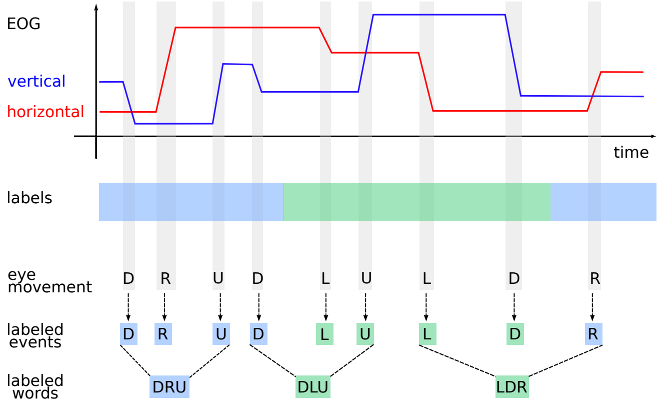

Figure 2. For classification using a spectrum string kernel SVM, saccades are first detected in the vertical and horizontal EOG signal. Saccades are encoded into characters and combined into a string of eye movements. The resulting string is split up into non-overlapping words of fixed length s (here $s = 3$), each labelled using majority voting.

(EOG) signals but EyeContext is independent of the specific measurement technique and works as well on eye movement signals recorded using a video-based eye tracker. The system processes these signals offline, detects and encodes saccades, models saccadic behaviour using machine learning, and generates continuous context labels using this model as its output.

The system removes noise from the recorded eye movement signals using a median filter as well as baseline drift using a wavelet packets approach. The system then detects and removes blinks using the *Continuous Wavelet Transform – Blink Detection* algorithm and saccades using the *Continuous Wavelet Transform – Saccade Detection* (CWT-SD) algorithm [4]. Briefly, CWT-SD detects saccades by thresholding on the continuous 1-D wavelet coefficient vector computed from the processed eye movement signals.

Detected saccades are encoded into eye movements using an alphabet $\mathcal{A}$ of eight distinct characters based on the saccades' direction. For example, a saccade to the left is encoded as “L” while a saccade to the diagonal right is encoded as “B”. These characters are merged into a single string that represents the sequence of consecutive eye movements for each participant. For classification, EyeContext splits this string into words S_i of fixed length s using majority voting for words that cover several ground truth labels (see Figure 2). These words are used as input to a spectrum string kernel SVM classifier.

A spectrum string kernel $K(S_i, S_j)$ of power k is defined as

$$\begin{aligned} K(S_i, S_j) &= \langle \Phi(S_i, k), \Phi(S_j, k) \rangle \\ \Phi(S_i, k) &= \left[\phi_{a_1}(S_i) \; \dots \; \phi_{a_{|\mathcal{A}|^k}}(S_i) \right]^T \end{aligned}$$

Intuitively, $\phi_a(S_i)$ counts the position-independent occurrences of all contiguous eye movement patterns $a \in \mathcal{A}^k$ of length k (also called k-mers or k-grams) contained in S_i. The set of all k-mers is also called the k-spectrum of a string with each k-mer representing one dimension of the feature space (for details see [7]). The two main parameters of a spectrum string kernel are the word length s, i.e. the length of each input string to the SVM, and the power k of the kernel, i.e. the length of the eye movement patterns.

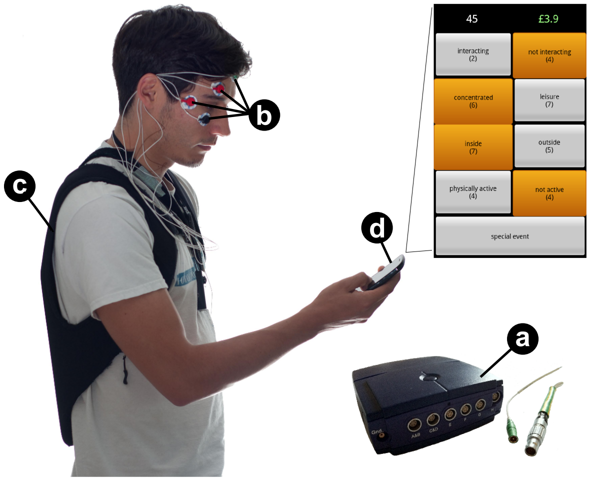

Figure 3. Experimental equipment used for data collection consisting of the Mobi (a), five EOG electrodes (b), a laptop for data recording (c), as well as a smartphone for self-annotation (d).

STUDY

We designed a user study to answer the fundamental question of whether analysing human visual behaviour can be used to automatically recognise four high-level contextual cues. The example cues that we explore in this work are:

1. Social cues: Social (face to face) interactions with another person versus no interaction.
2. Cognitive cues: Concentrated work, such as reading, versus leisure, which includes all forms of passive consumption (e.g. watching a video) or non-goal driven activities (e.g. a night out with friends).
3. Physical cues: Physically active, such as walking, versus resting, such as standing or sitting.
4. Spatial cues: Being inside or outside, e.g. a building, a car, or a train.

Apparatus

We used EOG to be able to record continuously for more than 12 hours. For EOG data recording we used a Mobi system by Twente Medical Systems International. The Mobi was worn by the participants and transmitted data sampled at 2 kHz over Bluetooth to a laptop carried in a shoulder bag. EOG signals were picked up using five Ag/AgCl wet electrodes. Pairs of electrodes were attached to the outer edge of the left and right eye, above the right eyebrow and below the right eye, and an additional reference electrode was placed on the forehead. Ground truth annotation was performed using a custom software running on an Android smartphone (see Figure 3).

Procedure

For data collection we recruited four full-time researchers from the lab. Their typical work day is characterised by lots of work with only little leisure. Specifically, their everyday life includes commuting to the university, working concentrated

	P1	P2	P3	P4	Total
concentrated	454	341	268	374	1,437
leisure	268	286	354	210	1,118
inside	698	600	525	521	2,344
outside	24	28	97	63	211
social interaction	116	215	238	299	868
no interaction	606	413	384	285	1,688
physically active	114	76	278	99	566
resting	608	551	345	485	1,989

Table 1. Overview of the recorded dataset. The table shows the amount of data labelled for each contextual cue in minutes.

at the desk, interacting with colleagues throughout the day, occasionally leaving the building, e.g. for lunch, and commuting back home in the evening. We asked them to self-annotate such non-mutually exclusive transitions as accurately as possible, while still retaining their daily routine. In an initial physical meeting in the lab participants were introduced to the recording system and shown how to attach the electrodes and start the recording. Participants were instructed to start the recording at home in the morning and stop it in the evening.

Data Analysis

All parameters of the signal processing and saccade detection algorithms were fixed to values common to all participants. We considered four individual binary classification problems, one for each contextual cue. As class distributions were considerably skewed (see Table 1), similar to [1], we used a discrete HMM model to oversample the smaller class until both classes were of the same size. The predictions returned by the string kernel SVM were compared to the annotated ground truth using a person-dependent evaluation scheme: the dataset for each participant was split using 70% for training and 30% for testing. Classification was only performed on the test set. During classification the values of k and s were optimised with respect to recognition accuracy. This evaluation was run five times (5-fold cross-validation) and the following performance measures averaged. Precision was calculated as $\frac{TP}{TP+FP}$, recall (true positive rate) as $\frac{TP}{TP+FN}$, and false positive rate (FPR) as $\frac{FP}{FP+TN}$, where TP, FP, TN and FN represent true positive, false positive, true negative and false negative counts, respectively.

RESULTS

We were able to record a dataset of more than 42.5 hours of eye movement data (see Table 1). The dataset comprises nearly 24 hours of concentrated work (18.5 hours of leisure), 39 hours were spent inside (3.5 hours outside), 14.5 hours of social interactions (28 hours of no interaction), as well as 9.3 hours of physically active periods (33.2 hours of resting).

Figure 4 plots the recognition performance for each contextual cue and participant, as well as the means over all participants. The best mean result is for recognising social interactions for which the system performed well for all participants (85.3% precision, 98.0% recall on average). The worst result is for recognising physical activities (75.3% precision, 74.4% recall) with a notably lower recall than for the cognitive cue (73.2% precision, 83.1% recall) and the spatial cue (74.0% precision, 85.0% recall). These results show that while the system has similar performance in correctly recognising actual activity, cognitive and spatial cue instances, it has a harder time in spotting all activity instances. Figure 4 also indicates tendencies of particular participants to perform consistently worse than others. Table 2 confirms this finding. The highest performance is achieved for P4 (84.0% precision, 86.0% recall, 15.8% FPR), while the worst result is for P3 (66.3% precision, 79.8% recall, 38.4% FPR). On closer inspection of the raw EOG data, it turned out that the signal quality for P3 was much worse compared to the other participants and saccades could not be robustly detected. Dry skin or poor electrode placement are the most likely culprits.

	P1	P2	P3	P4	mean
precision	81.4%	75.5%	*66.3%*	**84.0%**	76.8%
recall	**88.6%**	87.4%	*79.8%*	86.0%	85.5%
FPR	19.8%	27.7%	*38.4%*	**15.8%**	25.4%

Table 2. Precision, recall and false positive rate (FPR) for each participant averaged over all four contextual cue as well as the mean over all participants. Best results are indicated in bold; worst results in italic.

DISCUSSION

Our EyeContext system demonstrates a novel way in which eye movements can be embraced for human-computer interaction. At the human-computer interface, eye movements were previously studied mostly for explicit control or specific diagnostics. Recent related work demonstrated automated recognition of particular activities, such as reading and writing, from eye movements [3, 4]. The current work differs fundamentally, as it demonstrates the feasibility of inferring contextual cues that are not limited to particular activities but broadly descriptive of our situation at any point in time.

Previous works on eye-based activity recognition used a computationally complex feature-based recognition approach. In contrast, the proof-of-concept system described in this work focuses on eye movement patterns that are first encoded and then classified using a string spectrum kernel method. This approach is computationally simple – as it does not require to extract a large number of low-level eye movement features – and it also implements the assumption that high-level contextual cues are characterised by differences in repetitive visual behaviours. It will be interesting to see how this approach compares to other methods geared to processing large amounts of sequences of symbols, such as networks of motifs [9].

Limited recording time and bulky equipment still prevent current video-based eye trackers from being used for long-term recordings in daily life. We thus opted to use EOG, which is light-weight and can be implemented as a low-power wearable system. It is important to note that neither the eye movement encoding and recognition approach, nor EyeContext in general are limited to the specific measurement technique used in this work. The system can also be used with other

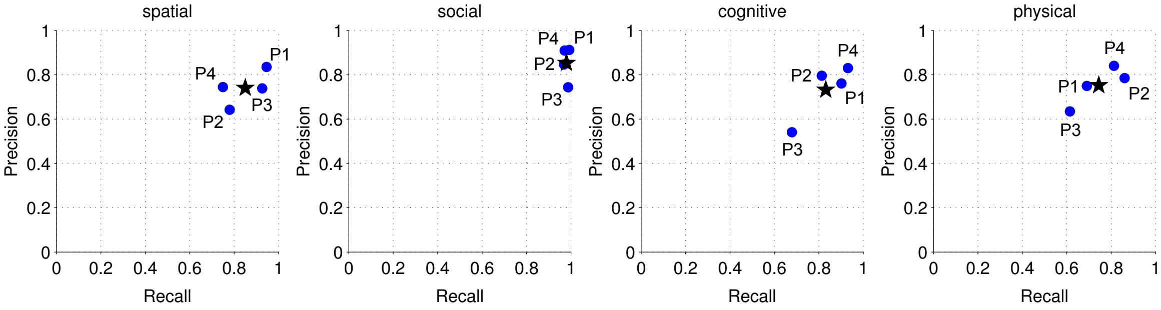

Figure 4. Overall best performance in terms of precision and recall for each cue and participant. Black stars mark the mean performance.

portable and ambient eye trackers and we strongly believe that even totally unobtrusive tracking will become feasible.

As any dataset, the one we presented here has limitations in terms of the number of days and participants that we recorded. Although larger variability is always desirable, it has to be stressed that there was a high degree of variability of activities within each class. For example, every instance of a conversation recorded by a user will have varied considerably, as it occurs in different daily life situations. The class distributions within the dataset were considerably skewed (unbalanced) for three of the four contextual cues (see Table 1). It is deliberate that we did not constrain the participants in any way (e.g. by scripting their entire working day) but opted for a data collection that is as imbalanced as the real world is.

The contextual cues investigated in this work are only exemplary but potentially useful for a number of applications. For example, logging one's life in digital form has a long held fascination and research has shown that recordings in everyday life can support memory, sharing, and behaviour analysis [8, 10, 5]. While capture technology is well explored [6], automatic annotation and filtering of long-term life logging data is still a significant challenge. Recognition of when a person is physically active or indoor/outdoor is promising for filtering because it is less narrow than annotation of specific activities and useful for breaking down the search space. Recognition of social cues may allow caregivers to automatically measure how socially active elderly or people with autism spectrum disorders are. Information on cognitive load may provide valuable insights into cognitive abilities relevant for medical or behaviour monitoring. EyeContext, however, is not restricted to these cues or the specific experimental settings investigated here but can be extended to other cues and everyday situations by using additional binary classifiers.

CONCLUSION

In this work we described EyeContext, a system that infers high-level contextual cues about different aspects of our daily life by analysing visual behaviour over time. Based on a proof-of-concept implementation and four long-term eye movement datasets we showed that we could robustly recognise four example binary cues. While previous work demonstrated the rich information content available in low-level eye movement characteristics, these results show that additional and equally valuable information is contained in the general eye movement patterns that we perform throughout a day.

REFERENCES

1. Beigi, M., Zell, A. Synthetic protein sequence oversampling method for classification and remote homology detection in imbalanced protein data. *Proc. of the 1st Int. Conf. on Bioinformatics Research and Development* (2007), 263–277.

2. Blum, M., Pentland, A., Tröster, G. InSense: Interest-Based Life Logging. *IEEE Multimedia 13*, 4 (2006), 40–48.

3. Bulling, A., Ward, J. A., Gellersen, H. Multimodal Recognition of Reading Activity in Transit Using Body-Worn Sensors. *ACM Trans. on Applied Perception 9*, 1 (2012), 2:1–2:21.

4. Bulling, A., Ward, J. A., Gellersen, H., Tröster, G. Eye Movement Analysis for Activity Recognition Using Electrooculography. *IEEE Trans. on Pattern Analysis and Machine Intelligence 33*, 4 (2011), 741–753.

5. Doherty, A. R., Caprani, N., O Conaire, C., Kalnikaite, V., Gurrin, C., O'Connor, N. E., Smeaton, A. F. Passively recognising human activities through lifelogging. *Computers in Human Behavior 27* (2011), 1948–1958.

6. Hodges, S., Williams, L., Berry, E., Izadi, S., Srinivasan, J., Butler, A., Smyth, G., Kapur, N., Wood, K. Sensecam: a retrospective memory aid. *Proc. of the 8th Int. Conf. on Ubiquitous Computing* (2006), 177–193.

7. Leslie, C., Eskin, E., Noble, W. S. The spectrum kernel: a string kernel for SVM protein classification. *Proc. of the Pacific Symp. on Biocomputing* (2002), 564–575.

8. Sellen, A. J., Fogg, A., Aitken, M., Hodges, S., Rother, C., Wood, K. Do life-logging technologies support memory for the past?: an experimental study using sensecam. *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems* (2007), 81–90.

9. Sinatra, R., Condorelli, D., Latora, V. Networks of motifs from sequences of symbols. *Physical Review Letters 105*, 17 (2010), 178702.

10. Truong, K. N., Hayes, G. R. Ubiquitous computing for capture and access. *Foundations and Trends in Human-Computer Interaction 2*, 2 (2009), 95–171.