Discovery of Everyday Human Activities From Long-Term Visual Behaviour Using Topic Models

Julian Steil
Perceptual User Interfaces Group
Max Planck Institute for Informatics
Saarbrücken, Germany
jsteil@mpi-inf.mpg.de

Andreas Bulling
Perceptual User Interfaces Group
Max Planck Institute for Informatics
Saarbrücken, Germany
bulling@mpi-inf.mpg.de

ABSTRACT

Human visual behaviour has significant potential for activity recognition and computational behaviour analysis, but previous works focused on supervised methods and *recognition* of predefined activity classes based on short-term eye movement recordings. We propose a fully unsupervised method to *discover* users' everyday activities from their long-term visual behaviour. Our method combines a bag-of-words representation of visual behaviour that encodes saccades, fixations, and blinks with a latent Dirichlet allocation (LDA) topic model. We further propose different methods to encode saccades for their use in the topic model. We evaluate our method on a novel long-term gaze dataset that contains full-day recordings of natural visual behaviour of 10 participants (more than 80 hours in total). We also provide annotations for eight sample activity classes (outdoor, social interaction, focused work, travel, reading, computer work, watching media, eating) and periods with no specific activity. We show the ability of our method to discover these activities with performance competitive with that of previously published supervised methods.

Author Keywords

Eye Movement Analysis; Activity Recognition; Topic Models; Bag-of-words; Latent Dirichlet Allocation (LDA)

ACM Classification Keywords

I.5.2 Pattern Recognition: Design Methodology: Pattern analysis; I.5.4 Pattern Recognition: Applications: Signal processing

INTRODUCTION

Practically everything that we do in our lives involves our eyes, and the way we move our eyes is closely linked to our goals, tasks, and intentions. These links make the eyes a particularly rich source of information about the user as demonstrated by the increasing number of works that use eye movements and closely related measures, such as pupil diameter or blink rate, for context recognition. For example, eye movement analysis has been used to recognise everyday activities, such as in the office [7] or reading in transit [5, 21]. Moreover, the close link between eye movement and cognition promises automatic analysis of covert aspects of user state that are difficult if not impossible to assess using existing sensing modalities, such as language expertise [25], visual memory recall [4], perceptual curiosity [18] or cognitive load [27, 37, 9].

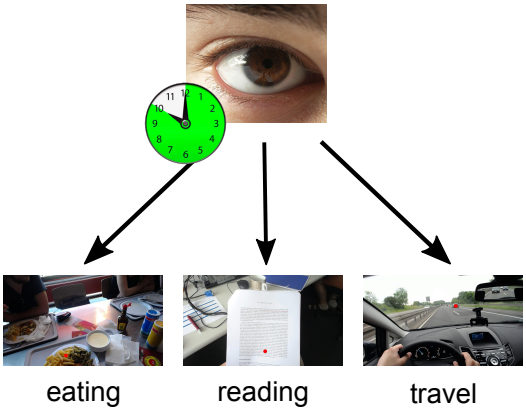


Figure 1: Our method takes long-term visual behaviour data (up to ten hours) as input and discovers everyday human activities, such as eating, reading, or being on travel, without supervision.

Despite these advances, previous works focused on short-term visual behaviour and supervised methods to *recognise* predefined activity classes. The availability of robust and affordable mobile head-mounted eye trackers points the way to a new class of context-aware systems that can *discover* activities from characteristic eye movement patterns, i.e. without any supervision. Unsupervised discovery of activities from eye movements has the potential to enable a range of novel applications, such as eye-based life logging [20], mental health monitoring [38], or the quantified self [26]. The problem setting for these applications is that of post-hoc analysis of human visual behaviour. In that setting, a full-day recording of a person's visual behaviour is available at the time of analysis. The goal of the analysis is to discover characteristic visual behaviours that can then be associated to a set of desired target activity classes. These characteristic behaviours occur at arbitrary points in time and with varying durations throughout the day. Such analysis problems commonly arise in the aforementioned application domains.

UbiComp '15, September 7-11, 2015, Osaka, Japan.

ACM 978-1-4503-3574-4/15/09...$15.00.
http://dx.doi.org/10.1145/2750858.2807520

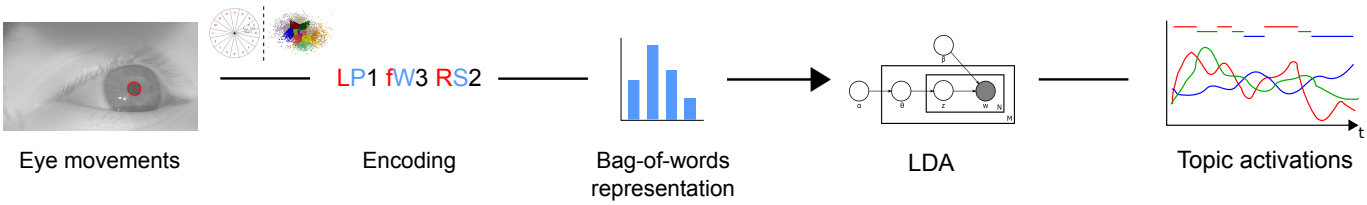


Figure 2: Input to our method consists of eye movements detected in the eye video. These movements are first encoded into a string sequence from which a bag-of-words representation is generated. The representation is used to learn a latent Dirichlet allocation (LDA) topic model. Output of the model is the set of topic activations that can be associated with different activities.

So far, however, it remains unclear how much information about daily routines is contained in long-term human visual behaviour, how this information can be extracted, encoded, and modelled efficiently, and how it can be used for unsupervised discovery of human activities. The goal of this work is to shed some light on these questions. We collected a new long-term gaze dataset that contains natural visual behaviour of 10 participants (more than 80 hours in total). The data was collected with a state-of-the-art head-mounted eye tracker that participants wore continuously for a full day of their normal life. We annotated the dataset with eight sample activity classes (outdoor, social interaction, focused work, travel, reading, computer work, watching media, and eating) and periods with no specific activity (see Figure 1). The dataset and annotations are publicly available online. We further present an approach for unsupervised activity discovery that combines a bag-of-words visual behaviour representation with a latent Dirichlet allocation (LDA) topic model (see Figure 2). In contrast to previous works, our method is fully unsupervised, i.e. does not require manual annotation of visual behaviour. It also does not only extract information from saccade sequences but learns a more holistic model of visual behaviour from saccades, fixations, and blinks.

The specific contributions of this work are three-fold. First, we present a novel ground truth annotated long-term gaze dataset of natural human visual behaviour continuously recorded using a head-mounted video-based eye tracker in the daily life of 10 participants. Second, we propose an unsupervised method for eye-based discovery of everyday activities that combines a bag-of-words visual behaviour representation with a topic model. To this end we also propose different approaches to efficiently encode saccades, fixations, and blinks for topic modelling. Third, we present an extensive performance evaluation that shows the ability of our method to discover daily activities with performance competitive with that of previously published supervised methods for selected activities.

RELATED WORK

Our method builds on previous works on eye movement analysis, eye-based activity and context recognition, as well as discovery of human activities using topic models.

Eye Movement Analysis

Eye movement analysis has a long history as a tool in experimental psychology and human vision research to better understand visual behaviour and perception. Despite its widespread use, previous works typically analysed a small set of well-known eye movement features, most notably fixation duration or fixation patterns. In an early work, Salvucci et al. described three methods based on sequence-matching and hidden Markov models for automated analysis of fixation patterns [30]. Later works used fixation analysis, for example, to identify image features that affect the perception of visual realism [13], to train novice doctors in assessing tomography images [12], or to study differences in face recognition [10]. Blink rate was shown to correlate with fatigue [32]. The analysis of the high-frequent fluctuations in pupil diameter has emerged as a robust and well-tested measure of cognitive activity, such as high cognitive load [27, 29]. All of these works demonstrated the significant influence of specific tasks on human visual behaviour, but they did not aim to analyse said behaviour to recognise the task at hand.

Eye-based Activity Recognition

Eye-based activity recognition was first explored in a series of studies by Bulling et al. They proposed a set of eye movement features, including repetitive saccade patterns, as well as a supervised method to recognise human activities from eye movements, such as reading in transit [5], office activities [7] or cognitive processes, such as visual memory recall processes [4]. A similar approach was later used by Tessendorf et al. to recognise cognitive load for context-aware hearing instruments [37], as well as by Kunze et al., who showed that different document types could be recognised from visual behaviour [24]. Ishimaru et al. used eye blink frequency and head motion patterns to recognise activities, such as reading or watching TV [21]. In human-computer interaction, recent works used specific eye movement features to recognise users' tasks, such as task transitions as well as perceptual and cognitive load [9], or cognitive abilities, such as visual working memory and perceptual speed [35]. More closely related to our work, Bulling et al. described an approach to recognise four high-level contextual cues, such as interacting with somebody vs. no interaction, from long-term visual behaviour [8]. However, their dataset was considerably smaller and, most importantly, their method was fully supervised.

Activity Discovery Using Topic Models

Topic models have been widely used to discover human activities from video (see [28] for an example) but less often from ambient and on-body sensors (see [33] for a recent analysis of different unsupervised activity discovery approaches). In

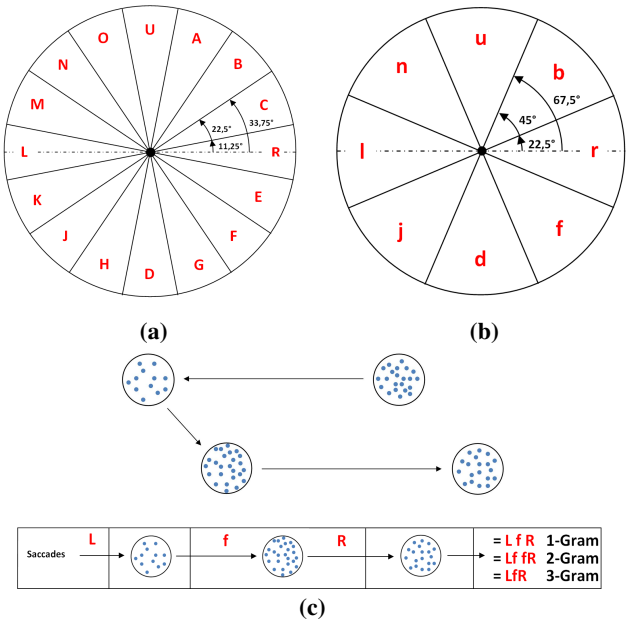


Figure 3: Encoding of large (a) and small (b) saccades according to their direction and amplitude. Example of a resulting encoding of three consecutive saccades for the 1-gram, 2-gram, and 3-gram approach (c). Blue dots indicate individual gaze samples belonging to four fixations.

an early work, Begole et al. analysed daily rhythms of computer use by clustering patterns of computer and email activity [2]. Barger et al. used mixture models to discover human behaviour patterns from statistics of sensor events in a smart home [1]. Gu et al. proposed an unsupervised approach for activity recognition based on fingerprints of object use [15]. They developed a wearable RFID system for object use detection and conducted a real-world data collection with seven participants in a smart home over two weeks. Farrahi et al. used topic models to infer daily routines from mobile phone data [14] while Huynh et al. discovered daily routines from accelerometer recordings of a single user [19]. We are not aware of any previous work that used topic models to discover activities from human visual behaviour.

ACTIVITY DISCOVERY FROM VISUAL BEHAVIOUR

We propose a method for unsupervised discovery of everyday human activities (see Figure 2 for an overview). Our method combines a bag-of-words visual behaviour representation with a latent Dirichlet allocation (LDA) topic model. Our model uses the full range of eye movements available in current head-mounted eye trackers, namely blinks, fixations (static states of the eyes), and saccades (fast simultaneous movements of both eyes to position gaze at a new location).

Eye Movement Detection

Eye movements are detected from the pupil positions provided by the eye tracker software in each eye video frame. We first identify overexposed frames and wrongly detected pupils. Specifically, we discard frames with an average grey value larger than 225, a pupil detection confidence value below 85%, or a pupil diameter smaller than 40 pixels. We

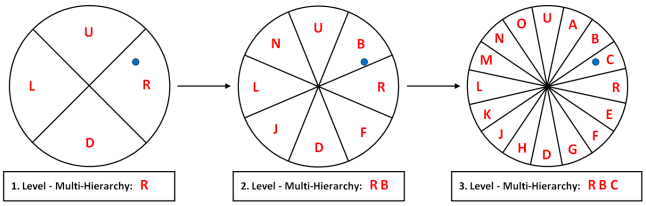


Figure 4: Sample multi-hierarchy encoding of a particular saccade direction (blue dot). The saccade is encoded across three granularity levels of the discretised saccade direction space.

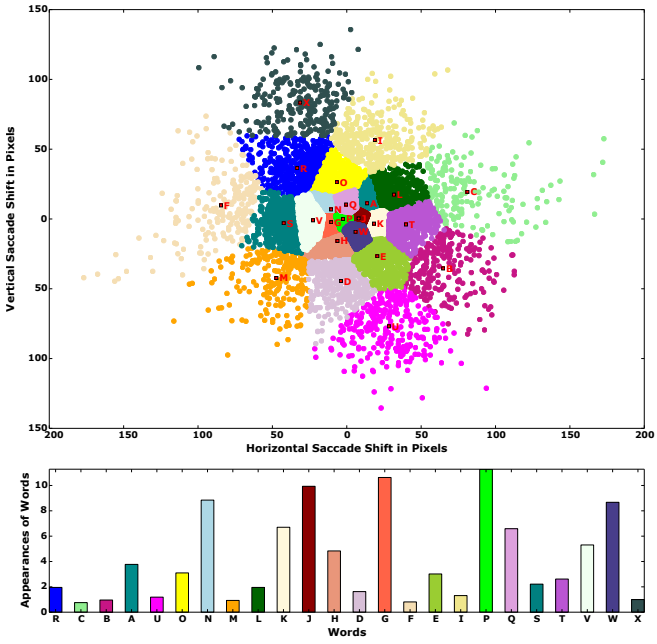


Figure 5: Sample k-means clustering of saccades based on their direction and amplitude for k = 24 of P6. Each cluster centroid is encoded with a distinct character.

found these values to work robustly in previous recordings in mobile settings with the same eye tracker.

We then detect three fundamental eye movements from the pupil positions, namely blinks, fixations, and saccades. Blinks can take place at any time and are characterised by closed eye lids. Consequently, to detect blinks, we take frames in which no pupil was detected as blink candidates. Failed pupil detections can also be caused by motion blur, e.g. during a saccade. To discriminate blinks and saccades we apply a velocity threshold of 150 pixels/sec on pupil positions. The velocity is calculated as the difference in pupil position before and after a particular blink candidate divided by the blink duration. We detect fixations using a dispersion-based algorithm [31]. Frames are assumed to belong to a fixation if the dispersion of the corresponding pupil positions is within a maximum radius of 7.5 pixels, which we determined empirically. In addition, a fixation had to last at least for 200ms [17].

Eye Movement Encoding

We propose four different approaches for encoding saccades into a sequence of characters. In the 1-gram approach we con-

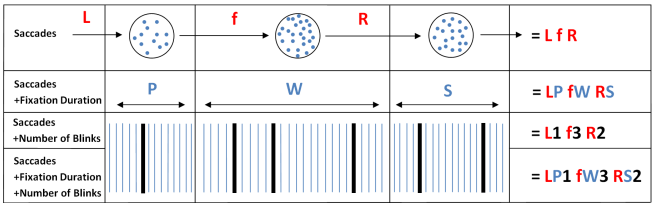


Figure 6: Fixation duration is binned into a person-specific histogram and each bin is encoded with a distinct character. The number of blinks is directly encoded in the character sequence.

sider individual saccades that we encode according to their direction and amplitude. Similar to [7], the n-gram approach generalises the 1-gram approach by considering n consecutive 1-gram encodings, thereby retaining information about pre- and succeeding saccades (see Figure 3). In the multi-level approach we discretise the saccade direction space with three granularity levels and encode saccades across these levels (see Figure 4). For the fourth approach we use k-means to cluster saccades into k clusters based on their direction and amplitude and encode each cluster centroid individually (see Figure 5). This approach is data-driven and only requires a single parameter, the number of clusters k, instead of predefined thresholds for saccade amplitudes and directions.

We further encode fixation duration and blink rate (see Figure 6). Fixation duration is a well-established measure in experimental psychology and commonly used for studies on visual perception and cognition [22]. We encode fixation duration by first finding the person-specific minimum and maximum durations and then splitting this range into 10 equally-sized bins. Each bin, and consequently all fixation durations that fall into that bin, is then encoded with a distinct character. The number of blinks during a fixation we directly encode in the string sequence. Finally, we encode the combined character sequence – that still contains temporal information – into a bag-of-words representation by generating histograms of word occurrence counts (see Figure 5).

Topic Modelling

The bag-of-words visual behaviour representation serves as input to an LDA topic model [3]. We opted for an LDA model given that it recently proved most robust among popular topic models [33]. Topic models were originally proposed in the text processing community [16] but subsequently became influential also in other domains, most notably computer vision [11] and human activity recognition [19]. As introduced by Blei et al. [3], topic models regard a corpus of text documents as a collection of words belonging to different topics, the so-called bag-of-words (BoW) representation. Topic models learn probability distributions of words belonging to these topics but, more importantly, also make it possible to infer the underlying topics from a corpus of documents.

Expressed mathematically, a document is defined as a collection of N words denoted by $\mathbf{w} = (w_1, w_2, ..., w_n)$, where w_n is the n^{th} word in the document. The document corpus C contains M documents denoted by $C = \{\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_M\}$. In addition to the document corpus, the number of topics K and the Dirichlet prior $p(\theta_d|\alpha)$ with parameter α on the topic-document distributions $p(t|\theta_d)$ have to be determined to derive θ, which describes the topic-document distribution. By defining the number of topics, the dimensionality of the topic variable t is assumed to be known and fixed. The word probabilities are parametrised by a $K \times V$ matrix β, where $\beta_{ij} = p(w^j = 1|t^i = 1)$. To calculate the probability of a corpus $p(C|\alpha, \beta)$, the parameters α for the Dirichlet distribution and parameter β for the word distribution $p(w|t, \beta)$ have to be found to maximise the likelihood $\mathcal{L}$ over all documents $d = 1, ..., M$. The formula is given by

$$p(C|\alpha, \beta) = \mathcal{L}(\alpha, \beta) =$$

$$\prod_{d=1}^{M} \int p(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} \sum_{t=1}^{K} p(w_n^d|t_n^d, \beta) p(t_n^d|\theta_d) \right) \theta_d, \quad (1)$$

where each document consists of the words w_n^d with $n = 1, ..., N_d$. With α and β, θ can be derived and the corpus C can be decomposed into the following form:

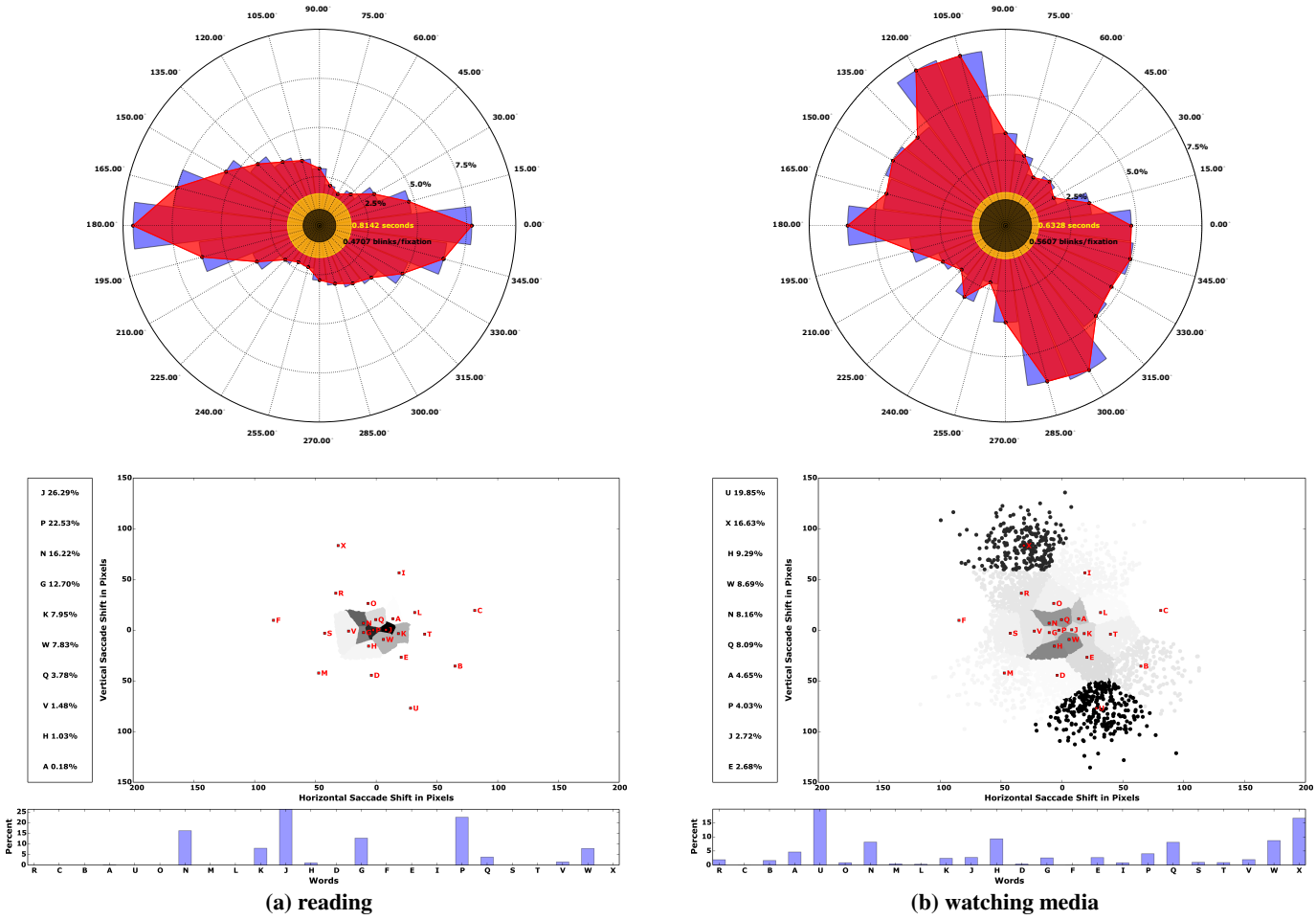
$$C = \phi \cdot \theta \quad (2)$$

DATA COLLECTION

In this representation the word-topic distribution ϕ and the topic-document distribution θ are key to discovering activities. Following the same terminology as [3], we propose to encode eye movement characteristics as words and to regard long-term visual behaviour as a corpus of text documents composed of these words, from which activities (topics) are automatically inferred. Consequently, we split the encoded visual behaviour sequence into a corpus of documents using a sliding window with a window size of five minutes and a stepsize of 30 seconds. These values were, again, determined empirically. We then run the LDA topic model with $K = 4, 6, 8, 10$ topics using a Dirichlet prior α of $50/K$, as recommended by Griffths and Steyvers [36]. The topic model generates two outputs: 1) the word-topic distribution ϕ that describes the visual behaviour for a specific topic or, as in our our case, during a specific activity, and 2) the topic-document distribution θ that indicates if and when a topic is active in a particular document. These topic activations are then associated with the different ground truth activities.

Figure 7 shows sample saccade direction distributions for "reading" and "watching media" as well as the corresponding word-topic distributions. The corresponding topic-document distributions are shown in Figure 8b while Figure 8a shows the topic activations. The active topics can then be compared to the annotated ground truth activities (see Figure 8c). In this example, topic 2 seems to represent "reading" while topic 3 matches best with "watching media".

To the best of our knowledge, the only long-term dataset of human visual behaviour recorded in daily life so far is the one presented in [8]. However, that dataset is not publicly available and, as mentioned before, it only contains relative eye movements of four participants recorded using a wearable electrooculography device. We therefore collected our own long-term visual behaviour dataset using a state-of-the-art head-mounted video-based eye tracker.



(a) reading (b) watching media

Figure 7: Sample saccade direction distributions for "reading" and "watching media" (top), as well as the corresponding 24-means saccade encoding and (middle) and word-topic distributions (bottom) for P6.

Apparatus

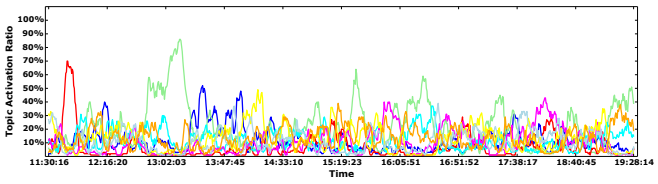
The recording system consisted of a Lenovo Thinkpad X220 laptop, an additional 1TB hard drive and battery pack, as well as an external USB hub. Gaze data was collected using a PUPIL head-mounted eye tracker connected to the laptop via USB [23] (see Figure 9). The eye tracker features two cameras: one eye camera with a resolution of 640×360 pixels recording a video of the right eye from close proximity, as well as an egocentric (scene) camera with a resolution of 1280×720 pixels. Both cameras record at 30 Hz. The battery lifetime of the system was four hours. We implemented custom recording software with a particular focus on ease of use as well as the ability to easily restart a recording if needed.

Procedure

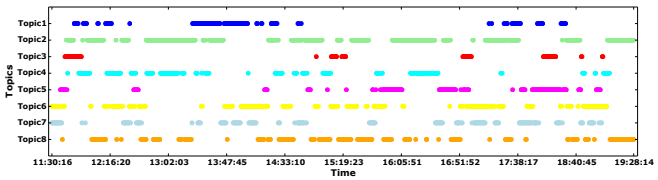
We recruited 10 participants (three female) aged between 17 and 25 years through university mailing lists and adverts in university buildings. Most participants were bachelor's and master's students in computer science and chemistry. None of them had previous experience with eye tracking. After arriving in the lab, participants were first introduced to the purpose and goals of the study and could familiarise themselves with the recording system. In particular, we showed them how to start and stop the recording software, how to run the calibration procedure, and how to restart the recording. We then asked them to take the system home and wear it continuously for a full day from morning to evening. We asked participants to plug in and recharge the laptop during prolonged stationary activities, such as at their work desk. We did not impose any other restrictions on these recordings, such as which day of the week to record or which activities to perform, etc.

Ground Truth Annotation

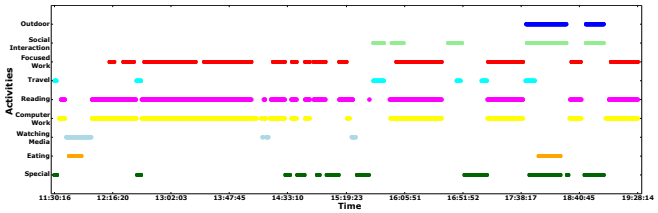
For evaluation purposes, the full dataset was annotated post-hoc from the scene videos by a paid human annotator with a set of nine non-mutually-exclusive ground truth activity labels (see Table 1 and Figure 8c). Specifically, we included labels for whether the participant was inside or outside (outdoor), took part in social interaction, did focused work, travelled (such as by walking or driving), read, worked on the computer, watched media (such as a movie) or ate. We further included a label for special events, such as tying shoes or packing a backpack. This selection of labels was inspired by previous works and includes a subset of activities from [8, 7, 34].



(a) topic activation



(b) active topics



(c) ground truth

Figure 8: Result of the topic modeling approach applied with eight topics on the 24-means encoding and the ground truth annotation of P6.

(a) (b)

Figure 9: Recording setup consisting of a laptop with an additional external hard drive and battery pack, as well as a PUPIL head-mounted eye tracker (a). Recording hardware worn by a participant (b).

Dataset

We were able to record a dataset of more than 80 hours of eye tracking data (see Table 1 for an overview and Figure 10 for sample images). The dataset comprises 7.8 hours of outdoor activities, 14.3 hours of social interaction, 31.3 hours of focused work, 8.3 hours of travel, 39.5 hours of reading, 28.7 hours of computer work, 18.3 hours of watching media, 7 hours of eating, and 11.4 hours of other (special) activities. Note that annotations are not mutually exclusive, i.e. these durations should be seen independently and sum up to more than the actual dataset size.

Most of our participants were students and wore the eye tracker through one day of their normal university life. This is reflected in the overall predominant activities, namely focused work, reading, and computer work. Otherwise, as can also be seen from the table, our dataset contains significant variability with respect to participants' daily routines and consequently the number, type, and distribution of activities that they performed. For example, while P1 wore the eye tracker during a normal working day at the university, P7 and P9 recorded at a weekend and stayed at home all day mainly reading and working on the computer (P7) or watching movies (P9) with little or no social interactions.

RESULTS

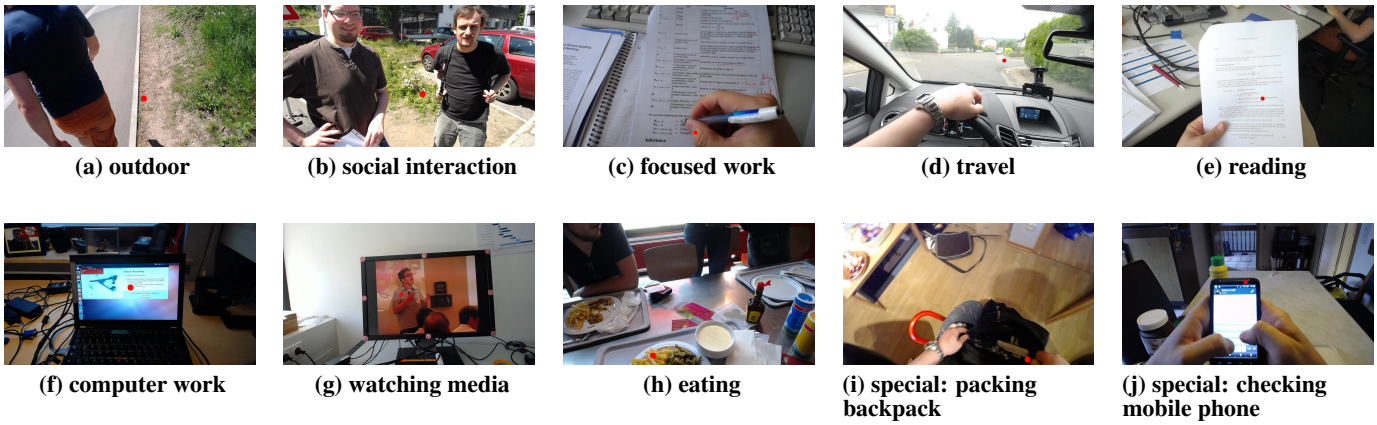
Huynh et al. used topic models to discover daily routines that consisted of re-occurring activities of a single person over several days [19]. Although participants' activities varied across days, their overall daily routines were still rather similar. In contrast, we deal with full-day recordings of multiple participants and a large variability with respect to the number, type, and distribution of activities that they performed, as well as their visual behaviour. In consequence, the best-performing model – specifically the best-performing saccade encoding, eye movement characteristics, as well as topic model parameters – is highly person-specific. We therefore opted to first show the best performance for each participant irrespective of the particular parameters used. In subsequent analyses we then focus on one representative participant to show the influence of different parameters on performance. In all analyses that follow, performance was calculated using the F1 score $F_1 = 2 * \frac{precision*recall}{precision+recall}$, which is the harmonic mean of precision $\frac{TP}{TP+FP}$ and recall $\frac{TP}{TP+FN}$, where TP, FP, and FN represent frame-based true positive, false positive, and false negative counts, respectively.

Performance for Each Participant

We first calculated the performance for each participant while optimising all free parameters of our method, i.e. saccade encoding, eye movement characteristics, as well as the number of topics in the topic model. Figure 11 shows the top mean F1 score for each participant with error bars visualising the range of performances for the particular subset of activities performed by the participant. As can be seen from the figure, our method achieves robust performance for discovering everyday activities across all participants independent of the particular type and distribution of activities. However, the figure also shows the considerable variability in performance for individual activities depending on the duration with which these activities were performed (cf. Table 1). For example, the minimum F1 score was achieved for P1 for watching media (8.34%) and P7 for social interaction (7.58%). As can be seen from Table 1, in both cases the respective activity was performed over considerably shorter durations than all other activities. The top F1 scores were achieved by P2 (93.83%)

Activity Class	Description	P1 (m)	P2 (m)	P3 (f)	P4 (m)	P5 (m)	P6 (m)	P7 (m)	P8 (f)	P9 (m)	P10 (f)	Total
outdoor	Person is outside	134	48	6	27	6	62	0	33	0	150	466
social interaction	Person is interacting with somebody else	173	69	127	77	81	95	5	59	0	169	855
focused work	Person is doing focused work	313	34	114	170	221	221	275	214	72	243	1877
travel	Person is travelling, e.g. walking or driving	156	70	40	47	33	47	18	32	23	30	496
reading	Person is reading	347	39	182	278	282	266	350	288	83	256	2371
computer work	Person is working on the computer	189	30	135	267	277	263	327	121	81	30	1720
watching media	Person is watching media	9	280	115	114	46	37	90	36	308	62	1097
eating	Person is eating	44	43	49	34	34	32	55	47	28	56	422
special	Special events, e.g. tying shoes	49	45	97	32	95	124	52	67	79	45	685

Table 1: Overview of the dataset showing the amount of ground truth annotated data for each activity class and participant in minutes. Participants' gender is given in brackets (f: female, m: male). Note that annotations are non-mutually exclusive, i.e. they sum up to more than the actual dataset size.



(a) outdoor **(b) social interaction** **(c) focused work** **(d) travel** **(e) reading**

(f) computer work **(g) watching media** **(h) eating** **(i) special: packing backpack** **(j) special: checking mobile phone**

Figure 10: Sample scene images for each activity class annotated in our dataset showing the considerable variability in terms of place and time of recording. The red dot indicates the gaze location in that particular image.

and P9 (91.33%) for watching media. These were also the activities performed the most among all activities.

Performance Across Participants

We then studied performance across all participants. As before, we optimised all free parameters of our method and calculated the mean, minimum, and maximum F1 scores for each activity. Figure 12 shows the top mean F1 score averaged over all participants performing the activity with error bars visualising the range of individual performances. The best performance was achieved for reading (74.75%), focused work (70.01%), and computer work (64.18%), while all other activities could be discovered with a mean F1 score of around 50%. These findings are in line with results reported in previous works that showed that reading and focused work could be recognised well using supervised learning methods [6, 37]. Table 1 further shows that the good performance correlates with the duration with which these three activities were performed, i.e. the more data is available, the better the activity can be discovered by our LDA topic model.

Impact of Different Saccade Encodings

We then evaluated the different saccade encodings because of their fundamental importance for our activity discovery method. For each encoding (1-gram, n-gram, multi-hierarchy, and k-means) we calculated the best average performance across activities and participants using all eye movement characteristics. We also swept the the number of topics $K = 4, 6, 8, 10$ in our topic model. Although not shown here, the k-means encoding with $k = 24$ and $K = 10$ topics performed best overall. Thus, we decided to use k-means encoding with k = 24 in all following evaluations.

As mentioned before, both the activities that participants performed and their visual behaviour was highly person-specific. Evaluating all parameters for all participants was therefore deemed infeasible. To select one representative participant, we calculated histograms over the activity durations for each participant as well as the total, and calculated the binary distances between these using the χ^2 distance metric. Based on these distance comparisons, we selected P6 for further investigation, as his activity distribution most closely resembled the distribution of the full dataset.

Impact of Eye Movement Characteristics

We were further interested in the impact of different eye movement characteristics on performance for individual activities. Figure 13 provides an overview of the performance for P6 for different eye movement characteristics using the 24-means saccade encoding for each activity. The figure shows that the best-performing eye movement characteristic is indeed activity-specific. For this specific participant,

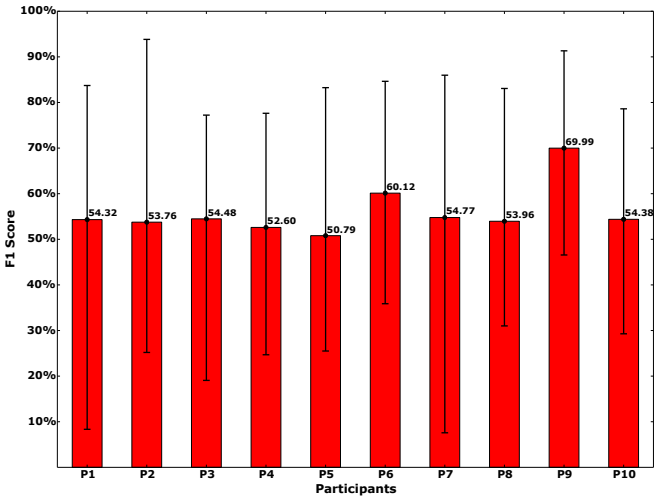


Figure 11: Top mean F1 scores for each participant with error bars visualising the range of performances for the particular set of activities performed by the participant irrespective of the particular saccade encoding, eye movement characteristics, or topic model parameters used.

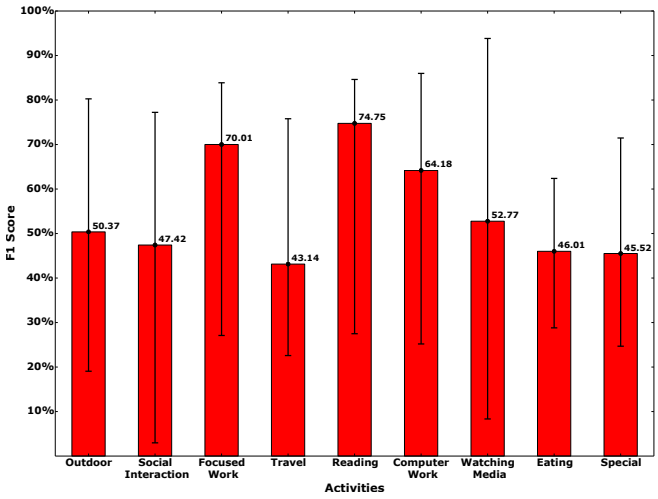


Figure 12: Top mean F1 score for each activity across all participants with error bars visualising the range of performances results for the participants performing the corresponding activity irrespective of the particular saccade encoding, eye movement characteristics, or topic model parameters used.

only using information about saccades achieved the best performance for four out of the nine activity classes, namely outdoor (45.7%), social interaction (53.6%), eating (41.5%), and special (56.9%). Additional information on fixation duration achieved the best performance only for focused work (73.7%) while adding information on blinks achieved best performance for travel (35.9%) and watching media (33.4%). Finally, using information about all three eye movement characteristics achieved best performance for reading (73.2%) as well as computer work (69.9%).

Impact of Number of Topics

The previous evaluation showed that additional eye movement characteristics can improve performance for particular activities. We further analysed the impact of different number of topics $K = 4, 6, 8, 10$ on performance. Figure 14 shows a performance comparison for different numbers of topics for the 24-means saccade encoding with blinks for P6. The figure shows that, similar to the different eye movement features, the number of topics affects individual activities differently. These performance differences are also linked to the duration of the activities performed by the participant (cf. Table 1). Generally speaking, the lower the number of topics the better the dominating activities – focused work, reading, computer work, and special – can be discovered from visual behaviour. The higher the number of topics, the more activities can be discovered, but with decreased F1 scores. This can be seen in Figure 14 where the F1 scores are generally higher for eight topics than for ten topics. If there are many activities, the smaller the number of topics, the worse the results given that one topic will encode multiple activities.

Comparison with Supervised Methods

Supervised methods have previously been used to recognise reading and different office activities from eye movement [7, 5]. We were therefore finally interested in comparing the performance achieved for discovering reading, computer work,

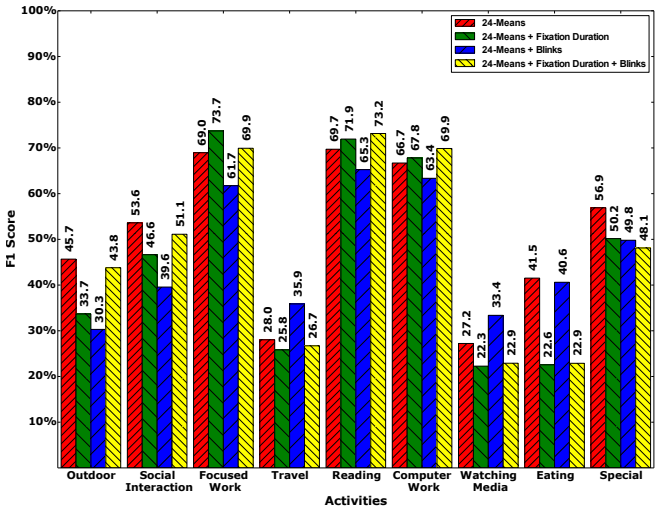


Figure 13: Performance comparison for different eye movement characteristics for the 24-means saccade encoding with 10 topics for P6.

and watching media with our unsupervised method with those used in prior work (see Figures 15-17). As shown in Figure 15 we were able to recognise reading with a top F1 score of 74.75% compared to the F1 score of about 70% achieved using a linear support vector machine as reported in [7]. For computer work we achieved a maximum mean F1 score of 64.18%, which is a bit lower than the 70% for browsing reported in [7]. For watching media we achieved a maximum mean F1 score of only 52.77%, while the corresponding performance for recognising watching video in [7] was about 83%. It is important to note, however, that performance for discovering computer work and watching media is reduced because not every participant performed these activities for a sufficient amount of time. For individual participants we

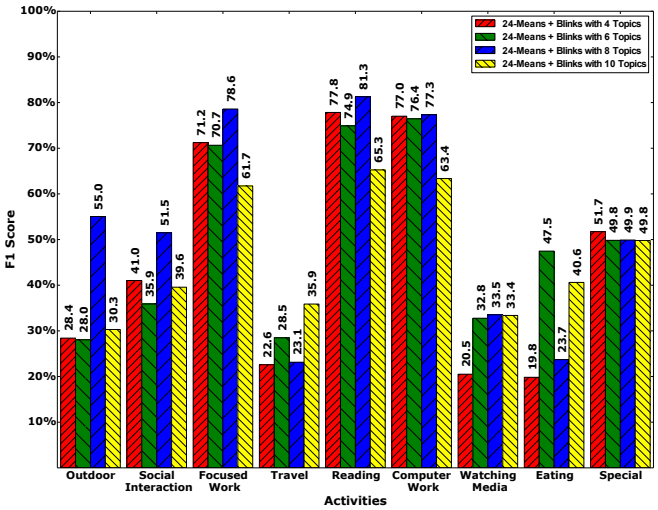


Figure 14: Performance comparison for different number of topics for the 24-means saccade encoding with blinks for P6.

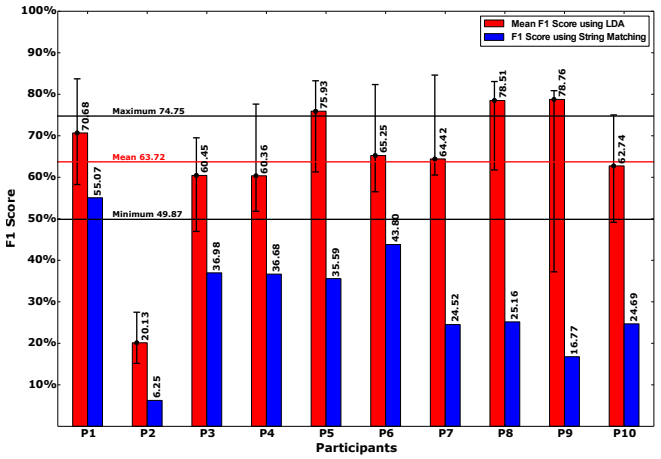


Figure 15: Performance comparison for "reading" for each participant using the 24-means saccade encoding with blinks and 10 topics. The blue bars show the F1 scores achieved using the string matching approach described in [6] which moves the reading template "Rlll" over the encoded 1-gram saccade sequence and thresholds on the Levensthein distances.

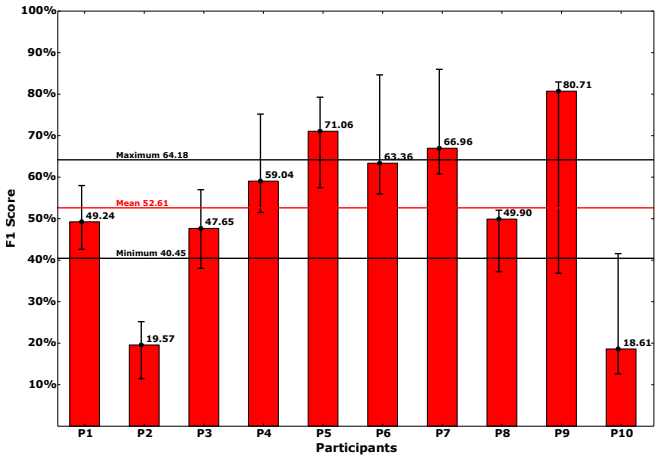


Figure 16: Performance comparison for "computer work" for each participant using the 24-means saccade encoding with blinks and 10 topics.

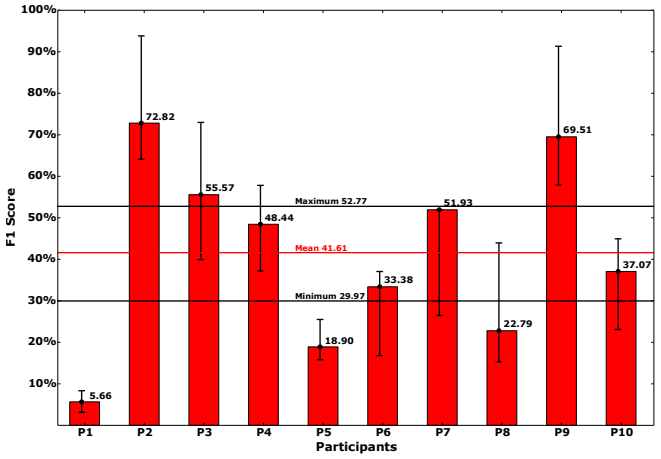


Figure 17: Performance comparison for "watching media" for each participant using the 24-means saccade encoding with blinks and 10 topics.

were able to achieve a maximum performance of over 90% F1 score for watching media.

To establish baseline performance results and directly compare the different methods on our new dataset, we reimplemented the string matching algorithm for reading recognition as described in [5]. We also trained our own linear support vector machines and naïve Bayes classifiers for binary activity classification, the former of which was used in [7]. In a nutshell, the string matching algorithm moves a predefined reading template "Rlll" over the encoded 1-gram saccade sequence. Intuitively, the template describes the characteristic sequence of small saccades to the left while scanning a line of text, followed by the large "carriage return" saccade to the right to jump to the beginning of the next line. To detect reading, the algorithm calculates the Levensthein distance and applies a distance threshold of $T_{ed} = 3$ in each step of moving the template over the sequence and finally performs majority voting in a window of string length $W_{str} = 30$. As can be seen from Figure 15, our LDA topic model outperforms the string matching approach for all participants.

For the SVM algorithm we fixed the two main parameters, the cost C and the tolerance of termination criterion ϵ, to C = 1 and ϵ = 0.001. Every feature vector consists of 56 of the 62 features described in [7] and was computed for a time window $W_{fe} = 120s$ and a step size $S_{fe} = 1s$. Table 2 provides an overview of this comparison for P6. As can be seen from the table, our method shows competitive performance to the SVM in terms of F1 score, accuracy and correlation and even outperforms SVM in terms of recall. Both always outperforms the naïve Bayes classifier.

DISCUSSION

Referring to the open questions from the introduction, results on our new 10-participant dataset demonstrate that long-term human visual behaviour does indeed contain a significant amount of information about our daily routines. We demonstrated that this information can be extracted from key eye

	Precision			Recall			F1 Score			Accuracy			Correlation		
Activity Class	*LDA*	*SVM*	*NB*	*LDA*	*SVM*	*NB*	*LDA*	*SVM*	*NB*	*LDA*	*SVM*	*NB*	*LDA*	*SVM*	*NB*
outdoor	38.0	68.7	29.6	100.0	85.0	69.9	55.0	76.0	41.6	81.2	94.4	89.3	0.55	0.73	0.41
social interaction	43.4	75.2	18.1	80.8	30.7	57.9	56.5	43.6	27.6	76.9	62.2	81.5	0.46	0.27	0.25
focused work	74.9	81.8	97.5	84.5	78.3	58.8	79.4	80.0	73.4	79.3	81.3	67.5	0.59	0.62	0.46
travel	23.7	69.7	93.5	73.3	23.3	11.0	35.9	35.0	19.7	78.9	78.9	38.0	0.33	0.31	0.16
reading	82.4	80.8	96.4	82.3	86.5	67.3	82.3	83.6	79.3	79.8	82.4	72.2	0.58	0.65	0.47
computer work	86.2	87.9	95.8	83.1	83.0	64.7	84.6	85.4	77.2	83.3	83.6	69.2	0.66	0.67	0.42
watching media	27.0	20.0	93.7	59.3	79.2	8.8	37.1	32.0	16.1	82.8	93.9	30.4	0.32	0.38	0.12
eating	38.3	60.8	97.2	86.1	68.7	18.0	53.0	64.5	30.4	89.4	95.6	70.6	0.53	0.62	0.34
special	41.4	57.4	96.1	92.5	79.8	28.6	57.2	66.8	44.1	66.6	86.0	40.2	0.44	0.59	0.21
Average	50.6	66.9	79.8	82.4	68.3	42.8	60.1	63.0	45.5	79.8	84.3	62.1	0.50	0.54	0.3

Table 2: Performance comparison for the LDA topic model, a support vector machine (SVM), and a naïve Bayes (NB) classifier in terms of precision, recall, F1 score, accuracy, and Matthews correlation coefficient for P6.

movements that can be readily tracked with head-mounted eye trackers, namely saccades, fixations, and blinks. We further proposed and evaluated different methods to efficiently encode the extracted information into a joint bag-of-words representation. Building on this representation, we introduced LDA topic models as a versatile method to model a wide variety of human visual behaviours. We demonstrated the suitability of this whole approach for unsupervised discovery of everyday activities. Specifically, we are able to recognise reading with a top average performance of 74.75%, which is competitive with results reported in previous works using fully supervised methods [5].

Our evaluations also revealed that the best combination of methods and parameters – and consequently the performance – depend considerably on the particular user and his specific visual behaviour as well as the type, number, and distribution of activities that he performed throughout the day. Consequently, to achieve good performance, both the specific eye movement characteristics as well as the number of topics (activities) modelled in the topic model have to be optimised to the particular set of activities relevant for a particular application. While this may seem a severe limitation, supervised methods pose even stricter requirements, as the set of activity classes recognised by the system has to be defined and trained up front. In contrast, the proposed method can deal with an arbitrary number of activity classes as long as the target activities are performed sufficiently long relative to all other activities. This requirement directly stems from the fact that topic models rely on word-topic and topic-document distributions and require sufficient statistics about individual topics.

CONCLUSION

In this work we proposed a new dataset as well as a fully unsupervised approach to discover human activities from long-term visual behaviour. Our approach efficiently encodes the full range of eye movements available in current head-mounted eye trackers, namely blinks, saccades, and fixations. Our results show the significant information content available in human visual behaviour for unsupervised discovery of activities, opening up new venues for research on eye-based behavioural monitoring and life logging.

ACKNOWLEDGEMENTS

We would like to thank Sonja Forderer for her help with the dataset annotation, Sabrina Hoppe for providing code for the evaluations of the supervised baseline methods, and all participants for their help with the data collection. This work was funded, in part, by the Cluster of Excellence on Multimodal Computing and Interaction (MMCI) at Saarland University, as well as a JST CREST research grant.

REFERENCES

1. Barger, T. S., Brown, D. E., and Alwan, M. Health-status monitoring through analysis of behavioral patterns. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans 35*, 1 (2005), 22–27.
2. Begole, J. B., Tang, J. C., and Hill, R. Rhythm modeling, visualizations and applications. In *Proc. UIST* (2003), 11–20.
3. Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *The Journal of Machine Learning Research 3* (2003), 993–1022.
4. Bulling, A., and Roggen, D. Recognition of Visual Memory Recall Processes Using Eye Movement Analysis. In *Proc. UbiComp* (2011), 455–464.
5. Bulling, A., Ward, J. A., and Gellersen, H. Multimodal Recognition of Reading Activity in Transit Using Body-Worn Sensors. *ACM Transactions on Applied Perception 9*, 1 (2012), 2:1–2:21.
6. Bulling, A., Ward, J. A., Gellersen, H., and Tröster, G. Robust recognition of reading activity in transit using wearable electrooculography. In *Proc. Pervasive 2008* (2008), 19–37.
7. Bulling, A., Ward, J. A., Gellersen, H., and Tröster, G. Eye Movement Analysis for Activity Recognition Using Electrooculography. *IEEE Transactions on Pattern Analysis and Machine Intelligence 33*, 4 (2011), 741–753.
8. Bulling, A., Weichel, C., and Gellersen, H. EyeContext: Recognition of High-level Contextual Cues from Human Visual Behaviour. In *Proc. CHI 2013* (2013), 305–308.

9. Chen, S., Epps, J., and Chen, F. Automatic and continuous user task analysis via eye activity. In *Proc. IUI* (2013), 57–66.
10. Chuk, T., Chan, A. B., and Hsiao, J. H. Understanding eye movements in face recognition using hidden markov models. *Journal of Vision 14*, 11 (2014).
11. Csurka, G., Dance, C., Fan, L., Willamowski, J., and Bray, C. Visual categorization with bags of keypoints. In *Proc. ECCVW*, vol. 1 (2004), 1–2.
12. Dempere-Marco, L., Hu, X.-P., MacDonald, S. L., Ellis, S. M., Hansell, D. M., and Yang, G.-Z. The use of visual search for knowledge gathering in image decision support. *IEEE Transactions on Medical Imaging 21*, 7 (2002), 741–754.
13. Elhelw, M., Nicolaou, M., Chung, A., Yang, G.-Z., and Atkins, M. S. A gaze-based study for investigating the perception of visual realism in simulated scenes. *ACM Transactions on Applied Perception 5*, 1 (2008), 3.
14. Farrahi, K., and Gatica-Perez, D. What did you do today?: Discovering daily routines from large-scale mobile data. In *Proc. MM* (2008), 849–852.
15. Gu, T., Chen, S., Tao, X., and Lu, J. An unsupervised approach to activity recognition and segmentation based on object-use fingerprints. *Data & Knowledge Engineering 69*, 6 (2010), 533–544.
16. Hofmann, T. Unsupervised learning by probabilistic latent semantic analysis. *Machine learning 42*, 1-2 (2001), 177–196.
17. Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., and Van de Weijer, J. *Eye tracking: A comprehensive guide to methods and measures*. Oxford University Press, 2011.
18. Hoppe, S., Morey, S., Loetscher, T., and Bulling, A. Recognition of curiosity using eye movement analysis. In *Adj. Proc. UbiComp* (2015).
19. Huynh, T., Fritz, M., and Schiele, B. Discovery of activity patterns using topic models. In *Proc. UbiComp* (2008), 10–19.
20. Ishiguro, Y., Mujibiya, A., Miyaki, T., and Rekimoto, J. Aided eyes: eye activity sensing for daily life. In *Proc. AH* (2010), 25.
21. Ishimaru, S., Weppner, J., Kunze, K., Kise, K., Dengel, A., Lukowicz, P., and Bulling, A. In the Blink of an Eye - Combining Head Motion and Eye Blink Frequency for Activity Recognition with Google Glass. In *Proc. AH* (2014).
22. Just, M. A., and Carpenter, P. A. Eye fixations and cognitive processes. *Cognitive psychology 8*, 4 (1976), 441–480.
23. Kassner, M., Patera, W., and Bulling, A. Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction. In *Adj. Proc. UbiComp* (2014), 1151–1160.
24. Kunze, K., Bulling, A., Utsumi, Y., Yuki, S., and Kise, K. I know what you are reading – recognition of document types using mobile eye tracking. In *Proc. ISWC* (2013), 113–116.
25. Kunze, K., Kawaichi, H., Yoshimura, K., and Kise, K. Towards inferring language expertise using eye tracking. In *Ext. Abstr. CHI* (2013), 217–222.
26. Kunze, K., Kawaichi, H., Yoshimura, K., and Kise, K. The wordometer–estimating the number of words read using document image retrieval and mobile eye tracking. In *Proc. ICDAR* (2013), 25–29.
27. Marshall, S. P. The index of cognitive activity: Measuring cognitive workload. In *Proc. Human factors and power plants* (2002), 7–5.
28. Niebles, J. C., Wang, H., and Fei-Fei, L. Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision 79*, 3 (2008), 299–318.
29. Palinko, O., Kun, A. L., Shyrokov, A., and Heeman, P. Estimating cognitive load using remote eye tracking in a driving simulator. In *Proc. ETRA* (2010), 141–144.
30. Salvucci, D. D., and Anderson, J. R. Automated eye-movement protocol analysis. *Human-Computer Interaction 16*, 1 (2001), 39–86.
31. Salvucci, D. D., and Goldberg, J. H. Identifying fixations and saccades in eye-tracking protocols. In *Proc. ETRA* (2000), 71–78.
32. Schleicher, R., Galley, N., Briest, S., and Galley, L. Blinks and saccades as indicators of fatigue in sleepiness warnings: looking tired? *Ergonomics 51*, 7 (July 2008), 982 – 1010.
33. Seiter, J., Amft, O., Rossi, M., and Tröster, G. Discovery of activity composites using topic models: An analysis of unsupervised methods. *Pervasive and Mobile Computing 15* (2014), 215 – 227.
34. Shiga, Y., Toyama, T., Utsumi, Y., Kise, K., and Dengel, A. Daily Activity Recognition Combining Gaze Motion and Visual Features. In *Adj. Proc. UbiComp* (2014), 1103–1111.
35. Steichen, B., Carenini, G., and Conati, C. User-adaptive information visualization: using eye gaze data to infer visualization tasks and user cognitive abilities. In *Proc. IUI* (2013), 317–328.
36. Steyvers, M., and Griffiths, T. Probabilistic topic models. *Handbook of latent semantic analysis 427*, 7 (2007), 424–440.
37. Tessendorf, B., Bulling, A., Roggen, D., Stiefmeier, T., Feilner, M., Derleth, P., and Tröster, G. Recognition of Hearing Needs From Body and Eye Movements to Improve Hearing Instruments. In *Proc. Pervasive* (2011), 314–331.
38. Vidal, M., Turner, J., Bulling, A., and Gellersen, H. Wearable Eye Tracking for Mental Health Monitoring. *Computer Communications 35*, 11 (2012), 1306–1311.