I know what you are reading – Recognition of Document Types Using Mobile Eye Tracking

Kai Kunze, Yuzuko Utsumi, Yuki Shiga, Koichi Kise
Osaka Prefecture University, Japan
lastname@m.cs.osakafu-u.ac.jp

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

ABSTRACT
Reading is a ubiquitous activity that many people even perform in transit, such as while on the bus or while walking. Tracking reading enables us to gain more insights about expertise level and potential knowledge of users – towards a reading log tracking and improve knowledge acquisition. As a first step towards this vision, in this work we investigate whether different document types can be automatically detected from visual behaviour recorded using a mobile eye tracker. We present an initial recognition approach that combines special purpose eye movement features as well as machine learning for document type detection. We evaluate our approach in a user study with eight participants and five Japanese document types and achieve a recognition performance of 74% using user-independent training.

Author Keywords
Eye tracking, document classification, reading behaviour

ACM Classification Keywords
H.5.2 Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION
People increasingly track physical fitness, from simple step counting over recording sports exercises to monitoring sleep. Tracking physical fitness increases the time people stay active, decreasing the risk for any obesity related diseases [4]. Our research goal is to investigate whether we can analyse knowledge acquisition activities in a similar fashion and motivate people to improve their mental fitness. We focus on reading as a particularly promising means to measure mental fitness. Although an increase in reading volume improves language comprehension and critical thinking skills, only few previous works evaluated reading activities in situ [10, 6, 7].

Information on how often a user reads specific document types is particularly valuable as it provides insights into interests (e.g. belletristic versus newspaper) or language expertise and skills (e.g. comics versus English literature). Inferring the type of document is challenging and cannot be easily done using standard wearable sensors. State of the art computer vision systems are able to spot text in a natural scene but there exists no general-purpose method for detecting the type of document.

We investigate for the first time whether the type of document a user is reading can be recognised only using information extracted from visual behaviour. The specific contributions are 1) a novel method to recognise the document type from visual behaviour and 2) an evaluation of document type recognition in a user study with 5 different Japanese reading materials (novel, manga, magazine, newspaper, textbook) and 8 users in 5 environments. We achieve a recognition rate of 99% user-dependent and 74% user-independent for approximately 1 minute of gaze data. We reach a recognition rate of 90 % applying a majority decision over the complete 10 min recordings user independent.

RELATED WORK
Bulling et al. described an approach to recognise different visual activities, including reading, solely from gaze behaviour using machine learning techniques in stationary [7] and mobile [6] settings. Kunze et al. presented Wordmeter – an approach to estimate the number of words read during the day using a mobile eye tracker and automatic document image retrieval [12]. In a related work they further investigated whether language expertise could be automatically inferred from visual behaviour [11]. They were able to spot words difficult for the user to understand. They also explore electroencephalography for detecting reading activities [13]. Biedert et al. investigated the complementary problem of determining how people read text. To this end they proposed a real-time algorithm to discriminate skimming from reading using a novel set of eye movement features [3]. They described an approach to classify text comprehensibility from visual behaviour recorded from multiple readers [2].

While several previous works focused on detecting when, how much and how a user is reading, our work is the first to explore whether the type of document can be detected from visual behaviour. While information on the document type is readily available when reading digital content, inferring the document type from reading printed text is considerably more challenging. There is a lot of related work using computer vision to perform document (type) classification (see Chen et. al. for an overview [9]). However, most researchers focus on scanned documents and do not perform in-scene document recognition. Unconstrained scene document/text detection is still considered an unsolved problem [14]. Document image retrieval methods work reliably and fast to identify the doc-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
ISWC'13, September 9–12, 2013, Zurich, Switzerland.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-2127-3/13/09...$15.00.
DOI string from ACM form confirmation

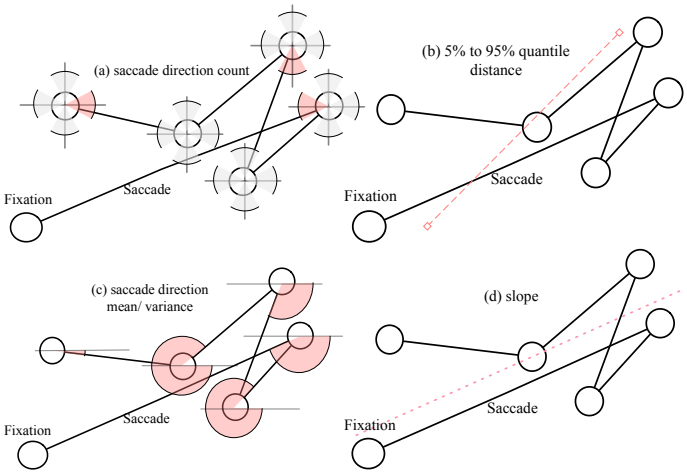


Figure 1: Features used for document type classification (circles represent fixations, lines represent saccades): saccade direction counts (a), 95% quartile distance (b), saccade direction (mean, variance) (c) and slope over the fixations (d).

ument, yet have the drawback that all documents need to be registered first with the system [15]. Using a scene camera also has other strong limitations, such as problems with ambient light and privacy issues. While computer vision methods can spot text in the field camera of an eye tracker, such methods when used alone still wouldn't allow to detect whether the user is actually reading.

DOCUMENT TYPE CLASSIFICATION

Research in experimental psychology evidenced that fixation patterns are strongly influenced by the task that elicited them. Specifically during reading, visual behaviour is influenced by a number of factors including the difficulty of the text, the document type (e.g. amount and location of text blocks and pictures), font size and type [1] the reading task or goal (e.g. reading for leisure vs. reading a task description during an exam), the number and duration of distractions, as well as the general context and reading situation of the person (e.g. reading while walking or while sitting at a desk).

Our approach to document type classification is inspired by earlier work that exploited the task-dependent nature of visual behaviour for eye-based activity recognition [7, 6] and cognitive processes [5]. Similar to that work, our key idea is to extract a number of gaze features during reading and to use machine learning to discriminate between different document types. The underlying assumption is that the type of document has a higher impact on visual behaviour than the other sources of influence described above.

In this paper, we focus on documents in Japanese. Japanese has two different reading styles: "Yokogaki" and "Tategaki". Yokogaki style writes left-to-right, top-to-bottom, as with English. Tategaki style writes first top-to-bottom, and then moves right-to-left. Most of the documents consist of a combination of text styles. These mixtures of writing styles are not only used in Japan but also in other parts of Asia, most notably Korea and China.
For document type classification we first split the raw gaze data into fixations and saccades according to Buscher et al. [8]. We then extract eight different gaze features on a sliding window with a window size of 200 and an overlap of 50 fixations. This window size corresponds to approximately one minute of gaze data depending on the user and document type. The large window size is needed to compensate for user-dependent reading habits. As features we extract counts for four different saccade directions (between 335° and 25°, 65° and 115°, 155° and 205° and 245° and 295°), the Euclidean distance between the 95% and 5% quantile of the fixation coordinates, the variance and median of the saccade directions, as well as the slope using linear regression over the fixations (see Figure 1 for details). We selected these features empirically out of a larger set of 25 gaze features using visual inspection and entropy measurements. The saccade direction count, variance and mean give a summary about the reading directions of the user. These feature differs significantly if the documents text layout is horizontal versus vertical. The features vary also for multicolumn versus single column layouts, or layouts with a lot of images versus no/little illustrations. The angle adjustments for the saccade counts gave empirically the best results as they remove crosswise eye movements of some users uncharacteristic for different document types. The quantile distance gives an indication on how the distance the eye covered. Therefore, it contains some information about the size of the document page. The quantile is used as it filters out some uncharacteristic eye movements (e.g. looking at the ceiling) from the distance measure. The slope over the fixations indicates the users general reading direction over the complete window. We use the extracted features as input to train a J48 decision tree classifier.

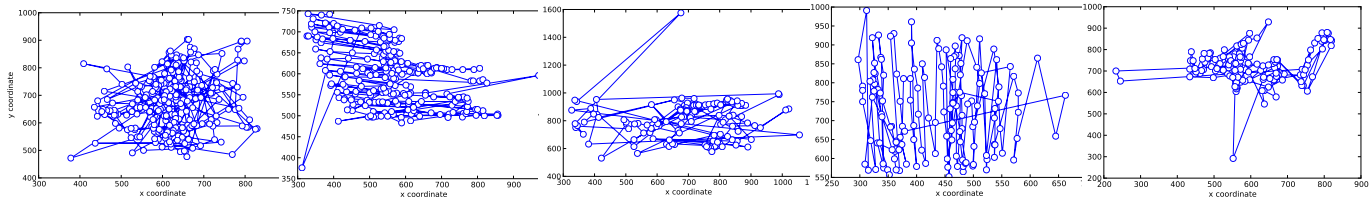
EXPERIMENT

We recorded gaze data of eight Japanese participants (four female) aged between 21 and 32 years using a mobile head-mounted eye tracker, the SMI Eye Tracking Glasses. The eye tracker provided binocular gaze estimates at a joint sampling frequency of 30Hz as well as a scene video with a resolution of 1280x960 pixels. At this sampling frequency, saccades of about 33ms and beyond can be measured. We only used the gaze data for our analysis; the scene videos were recorded solely for ground truth and documentation purposes. Gaze data was streamed and stored for later processing on a laptop.

We calibrated the eye tracker using a standard 3-point calibration prior to each recording. We chose five common reading situations: an office, a coffee shop, a home setting, a library and a lecture hall. We asked participants to read five different Japanese document types in each of these situations: a novel, a manga, a fashion magazine, a newspaper, and a text book. To achieve a more realistic coverage of documents we included four different novels, two newspapers, one manga book, two fashion magazines and one textbook. The Manga had 616 pages, contained lots of illustrations and most sections were in Tategaki (including larger ones) while only few were in Yokogaki. The magazines had 425 pages with lots of pictures, mostly in Yokogaki and only few in Tategaki. The novels had about 600 pages in total and were written in traditional Tategaki (text and reading direction). The textbook had 393 pages all in Yokogaki and only some illustrations. Fi-

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

Figure 2: Scene images of the eye tracker while a participant was reading a manga comic in the office (a), a textbook in the lecture hall (b), a fashion magazine at home (c), a novel in a coffee shop (d) and a newspaper in the library (e).



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

Figure 3: Representative fixation patterns for a manga (a), a textbook (b), a magazine (c), a novel (d) and a newspaper (e).

a	b	c	d	e	← classified as
89.26	1.97	4.50	4.27	0	a=novel
0.47	64.04	11.05	24.09	0.35	b=manga
4.61	10.60	45.16	13.94	25.69	c=magazine
6.39	16.54	12.94	63.14	0.98	d=newspaper
0.	0.35	9.87	1.70	88.08	e=textbook

Table 1: Confusion matrix in percent for user-independent frame-by-frame classification.

nally, the newspaper had 50 pages, some pictures, was written mostly in Tategaki with only little Yogaki and with a reading direction in Tategaki. All participants read 10 to 15 minutes of the assigned texts (we used 10 minutes for our analysis). On average, participants read 35 pages from the magazine (std 9), 40 pages from the manga (std 5), 8 pages from the novel (std 3), 6 pages from the textbook (std 4), and 4 pages from the newspaper (std 2). We vary the starting position for the reading materials (e.g. different chapter in the textbook). The document types, locations and starting positions were assigned using the Latin Square method. We ask participants to read naturally, i.e. according to their usual reading habits. Specifically, participants are allowed to sit and hold the reading material freely as preferred. Only the starting position and the location for the reading is defined by us. The reading behaviour was entirely up to the user. The calibration of the eye tracker is done in the pose the users selected to start their reading. The users are neither required nor told to keep their starting pose. Figures 2 and 3 show example scene images and gaze data for the different document types.

RESULTS AND DISCUSSION

In the user-independent case using leave-one-person-out evaluation we reach an average recognition performance of 74% accuracy over all users over the 1 min windows. Table 1 shows the confusion matrix for the frame-by-frame classification. "Textbook" and "Novel" show the best performance as they contain mostly structured text and users read the lines sequentially. Recognition performance for magazines is the lowest. Magazine recognition seems to be problematic only for male participants. A possible explanation is that they were not as interested in the content as women. The male participants quickly scanned the text from top to bottom while rarely looking at pictures. The "Magazine" regions misclassified as "Textbook" for P8 also contain a significant amount of such vertical text skimming. During fast reading and quick skimming, saccades under 20ms can occur and we are unable to register them using the eye tracker (only saccades over 33ms). For our analysis and experimental setup this was not an issue. Inference could be problematic, if users skimm large parts of the document etc.

Using majority voting over the whole 10 min of reading, we are able to increase performance to 90% accuracy. Misclassification occurs in 4 out of 40 cases (8 participants times 5 document types) for 3 participants. For P4 "Magazine" is misclassified as "Novel", for P6 "Newspaper" is misclassified as "Manga", and for P8 "Magazine" is classified as "Textbook" and "Manga" as "Newspaper". Regarding the error between "Newspaper" and "Manga" , the eye movement for P6 is very similar for both classes, as P6 mostly read and seldomly looked at pictures. He also encountered a larger structured text section in the "Manga" (very close to an article in the newspaper). Both classes consist of images and text (it depends on the users on how much they read or jump from one to the other).

For user-dependent evaluation (10-fold cross validation, 66% for training, 33% for testing) we achieve 98%-100% accuracy (mean performance 99%). Very few frames are not detected correctly, all between magazine, newspaper and manga. This high performance is promising for future reading applications

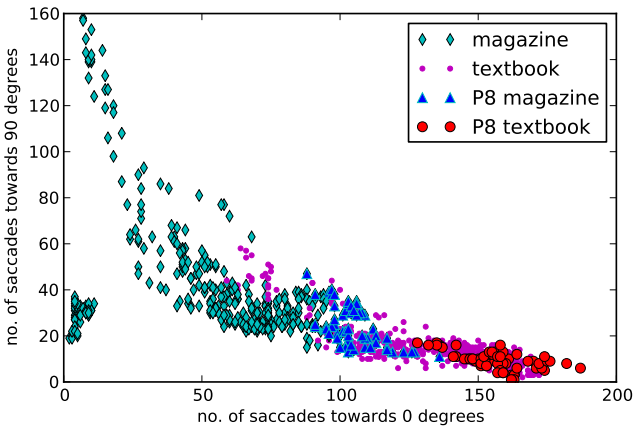


Figure 4: Feature plot to illustrate a problem with the classification of P8's reading behaviour between magazine and textbook.

that are personal, such as a reading assistant integrated in a head-mounted display that is worn by a single user.

As expected, the saccade direction features (a, c and d in Fig. 1) work best for classes with different text orientation (novel vs. textbook) and to discriminate between much versus little/no illustrations (magazine vs. textbook). The quantile distance performs best to distinguish smaller versus larger documents (manga vs. newspaper). This evaluation is based on calculating feature entropy between document types.
The misclassifications occurred for participants who showed uncharacteristic reading patterns for the specific document type. Figure 4 illustrates this and shows also why user-dependent classification still works flawlessly. The number of vertical saccades for the "Textbook" class is higher than for "Magazine". This holds also for data from P8. Yet, although both classes are clearly separable, his feature points for "Magazine" are still in "Textbook" section for the other participants. Thus, the misclassification happens.

It remains to be seen how well our approach works for languages that do not have different text/reading directions (e.g. English). As a quick test, we recorded 3 additional participants while reading a 2-column science paper for 10 minutes (the layout is very similar to the Textbook class except for the two columns). When adding this class to the existing ones the user-dependent classification rate was still at 99 %. Therefore, we believe that as long as the text/graphics layout is significantly different our approach works. Still, more experiments are needed to validate our approach for none East-Asian languages.

CONCLUSION

We investigated whether different document types can be recognised automatically from reading behaviour. We presented a recognition approach that achieved a recognition performance of 74% accuracy user-independent and 99% accuracy user-dependent. These results present a first step toward the vision of wearable reading assistance and reading logs.

REFERENCES

1. D. Beymer, D. Russell, and P. Orton. An eye tracking study of how font size and type influence online reading. In *Proc. British HCI*, pages 15–18, 2008.
2. R. Biedert, A. Dengel, M. Elshamy, and G. Buscher. Towards robust gaze-based objective quality measures for text. In *Proc. ETRA 2012*, pages 201–204, 2012.
3. R. Biedert, J. Hees, A. Dengel, and G. Buscher. A robust realtime reading-skimming classifier. In *Proc. ETRA 2012*, pages 123–130, 2012.
4. D. M. Bravata and Smith-Spangler. Using pedometers to increase physical activity and improve health. *The journal of the American Medical Association*, 298(19):2296–2304, 2007.
5. A. Bulling and D. Roggen. Recognition of visual memory recall processes using eye movement analysis. In *Proc. UbiComp 2011*, pages 455–464, 2011.
6. A. Bulling, J. A. Ward, and H. Gellersen. Multimodal Recognition of Reading Activity in Transit Using Body-Worn Sensors. *ACM Trans. on Applied Perception*, 9(1):2:1–2:21, 2012.
7. A. Bulling, J. A. Ward, H. Gellersen, and G. Tröster. Eye Movement Analysis for Activity Recognition Using Electrooculography. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 33(4):741–753, Apr. 2011.
8. G. Buscher and A. Dengel. Gaze-based filtering of relevant document segments. In *WSRSP Workshop*, volume 9, pages 20–24, 2009.
9. N. Chen and D. Blostein. A survey of document image classification. *International Journal of Document Analysis and Recognition (IJDAR)*, 10(1):1–16, 2007.
10. A. Cunningham and K. Stanovich. What reading does for the mind. *Journal of Direct Instruction*, 1(2):137–149, 2001.
11. K. Kunze, H. Kawaichi, K. Yoshimura, and K. Kise. Towards inferring language expertise using eye tracking. In *Ext. Abs. CHI 2013*, pages 4015–4021, 2013.
12. K. Kunze, H. Kawaichi, K. Yoshimura, and K. Kise. The wordmeter – estimating the number of words read using document image retrieval and mobile eye tracking. In *Proc. ICDAR 2013*, 2013.
13. K. Kunze, S. Yuki, S. Ishimaru, and K. Kise. Reading activity recognition using an off-the-shelf eeg. In *Proc. ICDAR 2013*, 2013.
14. H. Li, D. Doermann, and O. Kia. Automatic text detection and tracking in digital video. *IEEE Transactions on Image Processing*, 9(1):147–156, 2000.
15. T. Nakai, K. Kise, and M. Iwamura. Use of affine invariants in locally likely arrangement hashing for camera-based document image retrieval. *Proc. of DAS 2006*, 3872:541–552, Feb. 2006.