Using Eye Movements to Recognize Activities on Cartographic Maps

Peter Kiefer
Institute of Cartography and Geoinformation
ETH Zurich
Wolfgang-Pauli-Str. 15
CH-8093 Zurich, Switzerland
pekiefer@ethz.ch

Ioannis Giannopoulos
Institute of Cartography and Geoinformation
ETH Zurich
Wolfgang-Pauli-Str. 15
CH-8093 Zurich, Switzerland
igiannopoulos@ethz.ch

Martin Raubal
Institute of Cartography and Geoinformation
ETH Zurich
Wolfgang-Pauli-Str. 15
CH-8093 Zurich, Switzerland
mraubal@ethz.ch

ABSTRACT
The spatio-temporal characteristics of eye movements vary according to the activity the user of a cartographic map is performing. In this paper, we use these eye movement characteristics to automatically detect the map user's activity, an approach with great potential in gaze-assistive map interfaces. A dataset of 587 eye movement recordings from 17 participants was used to train and cross-validate a support vector machine (SVM) classifier over 229 features. The classifier can distinguish 6 common map activities with an accuracy of approx. 78%.

Categories and Subject Descriptors
I.2.1 [**Artificial Intelligence**]: Applications and Expert Systems – *cartography*

General Terms
Algorithms, Experimentation, Human Factors

Keywords
Activity recognition, gaze-based assistance, eye tracking, support vector machine, geographic human-computer interaction

1. GAZE-BASED ASSISTANCE ON MAPS
In this paper we elaborate on the idea of using eye tracking as a modality for the interaction with cartographic maps. By processing the user's gaze position in real-time we can design intelligent and efficient gaze-based user interfaces [1; 2].

Research on map perception has demonstrated that the way people visually explore maps is influenced by several factors: the stimulus [3], the user's cognitive state [4], and group differences [5]. Technically, these differences in visual map exploration are reflected in the spatio-temporal characteristics of eye movement patterns. Here, we are specifically interested in the user's activity (as indication for her cognitive state). Bulling et al. have shown that eye movement patterns can be used to recognize general office activities [6]. However, activities on maps are different.

Permission to make digital or hard copies of part or all 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 third-party components of this work must be honored. For all other uses, contact the Owner/Author.

Copyright is held by the owner/author(s).

SIGSPATIAL'13, Nov 05-08 2013, Orlando, FL, USA
ACM 978-1-4503-2521-9/13/11.
http://dx.doi.org/10.1145/2525314.2525467

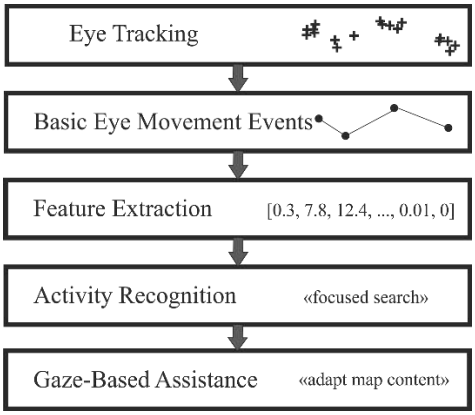

Figure 1. From eye movements to gaze-based assistance.

The processing hierarchy depicted in Figure 1 illustrates the required steps for activity recognition from eye movements on maps. We track a user's gaze while she is pursuing a map activity. Raw eye movements are preprocessed and classified into basic events (blinks, fixations, saccades). A set of features is computed and used in the fourth step as input for the classification.

In the remainder of this paper we describe our approach to *gaze-based activity recognition on maps*. We report on an eye-tracking experiment in which 587 gaze recordings of 6 different tasks were collected from 17 participants. The results show that a support vector machine (SVM) classifier is able to distinguish between 6 common map activities with an accuracy of approx. 78%.

2. RELATED WORK
Various approaches and application scenarios for activity recognition have been proposed, including activity recognition from movement in geographic space [7]. Similar to gaze-based activity recognition, the raw data here are a sequence of spatial positions. However, with a velocity of 30-500°/s [8] (p. 23), the saccadic movements of the eye are very different to movements in geographic space. In addition, activities related to motion tracks can be assumed to occur on a road network, which is in general not possible for gaze.

With current eye tracking technology it is possible to access the stream of gaze data in real-time, which in turn allows for using gaze as an input modality. These gaze-based interfaces can be designed with explicit or implicit interaction [9]. During explicit interaction, the user intentionally gazes at a certain position with the goal of triggering an interaction, such as selecting a zoom-in

position by gaze [1]. Implicit interaction, on the other hand, records the user's gaze during regular interaction and, at some later time, uses this information to adapt to the user's needs. An example for an implicit gaze-based geo-interface is the GeoGazemarks approach [2]. In this paper, we address what we believe is the biggest current challenge for implicit gaze-based interfaces: the correct interpretation of gaze in terms of activities.

For general office activities, such as writing, reading or copying, gaze-based activity recognition has already been proposed by Bulling et al. [6]. Similar to our work, they chose a machine learning approach and achieved recall values between 62% and 83%, depending on the activity. In contrast to our research, they used electrooculography instead of video-based eye tracking and trained the classifier with recordings of 5 minutes length. The most obvious difference to our work, however, are the types of activities used. It is not clear whether a classifier with acceptable accuracy can be learned for map activities.

Gaze map matching was investigated as the problem of matching gaze to the road a person is inspecting [10]. This is a purely geometric problem and thus on a lower semantic level than activity recognition. On a higher semantic level, eye tracking is also used in spatial cognition research, for instance for explaining the cognitive processes involved in wayfinding [11]. A recognizer for activities could help to automate the analyses of the data collected in such empirical studies.

3. DATA COLLECTION

3.1 Hardware and Software Setup

The hardware utilized for the experiment consisted of the SMI head-mounted eye tracking glasses with a gaze capture rate of 30 Hz[1]. This relatively low frequency was chosen with the goal of mobile systems in mind. The data were transmitted via a USB cable to a laptop, designated only for gaze recording. A chin rest was placed at a distance of 65 cm to the stimulus in order to provide stability. The stimulus was presented on a 24" widescreen monitor (1920x1200 pixels). The experiment was controlled through our own software framework that is able to choose a random set of test cases and present the stimuli.

3.2 Participants and Procedure

19 participants took part, 2 were excluded due to calibration errors. From the remaining 17 participants, 10 were female. The average age was 28 years (±8.7), and all were university students, or already holding a university degree. None of them could be considered an expert map user.

Each participant had 36 trials in total, taken from the 6 tasks (see section 3.3) and presented in randomized order, where no two successive trials were from the same task. After a trial, a recalibration was performed if necessary. Each trial consisted of three phases:

1) *Instruction phase*: the participant was presented a textual description of the task (in German) and could ask questions.
2) *Preview phase*: a preview showing small parts of the stimulus was shown. The goal of this phase was to clearly separate the activity to be analyzed from an orientation activity beforehand. At the end of the preview phase the participant was asked to fixate a certain point.
3) *Task phase*: the stimulus was shown, and the eye movements were recorded while the participant solved the task.

[1] http://www.smivision.com/en html

3.3 Tasks and Map Material

The following principles guided the selection of map material:

a) All maps must be taken from the *same cartographic product*. Our classifier should distinguish activities, not map designs.
b) Participants should be *familiar with the cartographic product.*
c) For tasks involving search, the relevant feature type should be distributed over the whole map extent.
d) Participants should be *unfamiliar with the geographic area* shown in the stimulus, but *familiar with the language and cultural context* of the area.

Based on these principles, we selected map material from Google Maps™ in the classical style[2]. All maps were chosen from Germany or Austria, since all participants were from Switzerland and native German speakers. The following six tasks were selected to provoke six activities:

Task 1. *Free exploration*: "You have 20 seconds for exploring the map. You can look at whatever you want". 6 stimuli (3 urban, 3 rural areas).

Task 2. *(Global) search*: "On the following map, please search for X", where X was a point of interest. 9 stimuli (urban areas containing at least 30 labeled points). Only trials taking at least 20 seconds were used for the analysis. In 6 stimuli, the object to search for was missing on the map.

Task 3. *Route planning*: "Do you see X and Y? Please, plan the shortest route from X to Y". 6 stimuli, randomly selected from 8 prepared stimuli, each covering one direction of search[3]. This should ensure that the classifier abstracts from the direction. In the preview phase, the labels of X and Y were shown on a white screen at the exact position where they would later appear in the stimulus.

Task 4. *Focused search*: "Do you see your position (the blue dot)? Please, search for the three closest Z", where Z is an object type. 5 stimuli, urban areas. As preview, the blue dot was shown on a white screen.

Task 5. *Line following*: "Do you see X? Please, follow X from North to South and count the number of intersections", where X is a street name, and the direction of the street was different for each stimulus (like in task 3). Each participant was shown 6 out of 8 available stimuli. The label of the street at the start position was shown as preview.

Task 6. *Polygon comparison*: "Do you see X and Y? Please compare the areas of these two lakes and name the bigger one". 4 stimuli, each covering one direction of comparison. The preview consisted of a white screen with two labels X and Y at the true position of the lakes.

4. METHODOLOGY

Preprocessing

Three basic eye movement events were computed by the SMI software: saccades, fixations, and blinks. Saccades are characterized by rapid eye movements, during which visual perception is reduced. New information can be obtained during

[2] http://maps.google.com/. In 2013, Google introduced a new Google Maps design. We used the classical design for our study to ensure participants were familiar with it.

[3] From North-West to South-East, North to South, North-East to South-West, East to West, plus the inverse direction for each. The randomizer balanced the distribution of trials for the 8 stimuli over all participants.

fixations, which occur in between saccades when the eyes remain relatively still for a short period of time [12]. Blinks occur when the eyes close and then quickly open again.

Blink-, fixation-, and saccade-based features

In order to capture the spatio-temporal characteristics of eye movements, a total of 229 features were computed for each of the eye tracking recordings (see Table 1). The first three types are eye movement features based on blinks, fixations, and saccades. These are standard measures for which definitions can be found in the eye tracking literature [8].

The saccade classification scheme

The remaining types of features are advanced, some inspired by [6] and some new. These measures try to capture the geometry of the scanpath by analyzing succeeding saccades w r.t their directions and amplitude. Two classification schemes for saccades are used in this context, where small and large amplitude are distinguished based on a threshold of 1.1° (as suggested by [13]):

- 16-classification (c16 in Table 1): eight cardinal directions, each for small and large amplitude
- 8-classification (c8 in Table 1): four cardinal directions, each for small and large amplitude.

Saccadic direction-based features

Based on c8, eight sub-sets of saccades were created. For each of these subsets, a number of saccade-based features were computed (based on amplitude, duration, skewness, and frequency).

In order to handle the cases where saccades occurred alternately in opposite directions we compared the angles of every two sequential saccades, leading to an *inversity* measure ($\in [0,1]$). Mean, minimum, maximum, and variance were considered, once taking all saccades into account, and once only for large saccades.

Similarly, the *category inversity* measure counts how often two succeeding saccades have an opposite direction (w.r.t. c8 or c16), normalized to the sequence length. We applied this measure once taking all saccades into account and once only for long saccades, and for both classification schemes.

Neighboring direction counts the number of occurrences of saccade sub-sequences where each two sequential saccades fall in the same or in neighboring direction categories (based on c8 or c16 respectively). The sub-sequences have lengths between 3 and 6 (sliding window). Neighboring direction is then normalized to the sequence length.

String sequence-based features

Two string sequences were created, one for c8 and one for c16, where each category was represented by one letter. Both sequences were analyzed with a sliding window algorithm (for all lengths between 1 and 4). The algorithm sequentially moves from left to right and creates sub-strings, based on the window size. The number of occurrences of sub-strings in the sequence is counted.

For all four window sizes we computed the number of created string patterns (c16w1-4 size), the number of minimum (c16w1-4 min occ) and maximum occurrences of a string pattern (c16w1-4 max occ), the difference between minimum and maximum occurrences (c16w1-4 min diff) as well as the variance of the occurrences (c16w1-4 var occ). These features were computed once taking all saccades into account and once only for long saccades. The procedure for deriving string sequence-based features was inspired by [6].

Blink-based features			
mean, min, max, var	duration	blinks	4
rate			1
Fixation-based features			
mean, min, max, var	duration, dispersion, dispersion X, dispersion Y	fixations	16
frequency			1
Saccade-based features			
mean, min, max, var	amplitude, duration	saccades	8
skewness	amplitude		1
frequency			1
g-l ratio	amplitude		1
Saccadic direction-based features			
mean, min, max, var	amplitude, duration	category (c8)	64
skewness	amplitude		8
frequency			8
mean, min, max, var	inversity	saccades, large-saccades	8
category inversity	c16w2 occ, c8w2 occ		4
neighboring direction	c16w3-6 occ, c8w3-6 occ		8
String sequence-based features			
size	c16w1-4, c8w1-4	saccades, large-saccades	16
min, max	c16w1-4 occ, c8w1-4 occ		32
mean, var, |max-min|	c16w1-4 occ, c8w1-4 occ		48
Total features			**229**

Table 1. Features used to train the classifier. Rows are interpreted as {cell 1} {cell 2} of {cell 3}. First row, for instance: "mean duration of blinks"

5. RESULTS

612 datasets were collected in total. 22 successful trials of task 2 were excluded because they were shorter than 20 seconds. 3 trials had to be excluded due to calibration issues or problems in understanding the task. This yields in a total of 587 eye movement recordings (102 task 1, 131 task 2, 100 task 3, 85 task 4, 101 task 5, 68 task 6). All eye movement recordings longer than 20 seconds were cut to the first 20 seconds.

The 229 features described in section 4 were computed for each of the 587 datasets. Features were linearly scaled to [0;1] over all trials. We used the LibSVM package for learning the classifier [14]. In our case, a C-support vector classificator (C-SVC) with RBF kernel revealed the best performance. The optimal values for γ and C were found with an iterative gridded-search using stratified 10-fold cross-validation (C=226.23; γ=0.0). Results of the classification are listed in Table 2: in total, 456 trials were classified correctly (accuracy of 77.7%).

The results look promising: recall is between 63.5% and 98%, with a recall of less than 75% for only two out of six activities. For office activities, previous work on activity recognition from gaze [6] has returned recall values which are lower on average (NULL activity 82%, reading 67%, browsing 62%, writing 73%, video watching 83%, copying 68%). These results are surprising, because we had expected map activities to be less distinguishable than office activities.

The best recall is achieved for polygon comparison. This is probably due to the very characteristic large inverse saccades between the compared objects. Free exploration and (global) search have high recall values (80.4%, 85.5%). The most incorrect classifications for both of them happen with the respective other. We had expected similarity between them, as both activities are not restrained to certain areas of the map.

Focused search has the lowest recall of all (63.5%), which is mainly due to confusion with global search and free exploration. Not surprisingly, route planning seems difficult to distinguish from line following, because planning a route implies following linear features.

		true activity						**precision**
		1	**2**	**3**	**4**	**5**	**6**	
predicted activity	**1**	82	10	3	11	2	0	75.9
	2	14	112	5	18	3	0	73.7
	3	1	2	65	1	19	0	73.9
	4	4	7	5	54	1	0	76.1
	5	1	0	21	1	76	1	76.0
	6	0	0	1	0	0	67	98.5
	∑	102	131	100	85	101	68	
recall		80.4	85.5	65.0	63.5	75.3	98.5	

accuracy = 77.7%

1: free exploration, 2: search, 3: route planning,
4: focused search, 5: line following, 6: polygon comparison.

Table 2. Confusion matrix for SVM classifier (10-fold cross-validation), precision and recall in %.

6. CONCLUSION AND OUTLOOK

We began this research with the idea of recognizing map activities from gaze. We approached the problem with an SVM classifier and achieved a recognition accuracy of 77.7%. Though our results on activity recognition look promising, a number of open issues for future research remain:

Does the approach work for other cartographic products? How can we solve the segmentation problem if different activities occur unseparated? What is the best time threshold for activity recognition? How can higher-level cognitive states (intentions) be inferred? Can we use the classifier learned from one user group for other users? Can map activities be distinguished from other activities? Do users accept gaze-assistive map user interfaces?

7. REFERENCES

[1] Stellmach, S. and Dachselt, R., 2012. Look & touch: gaze-supported target acquisition. In Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems (Austin, Texas, USA2012), ACM, pp. 2981-2990. DOI= http://dx.doi.org/10.1145/2208636.2208709.

[2] Giannopoulos, I., Kiefer, P., and Raubal, M., 2012. GeoGazemarks: Providing gaze history for the orientation on small display maps Proceedings of the 14th International Conference on Multimodal Interaction (ICMI '12), ACM, New York, NY, USA, pp. 165-172.

[3] Lloyd, R., 1997. Visual search processes used in map reading. Cartographica: The International Journal for Geographic Information and Geovisualization 34, 1, pp. 11-32.

[4] Steinke, T.R., 1987. Eye movement studies in cartography and related fields. Cartographica: The International Journal for Geographic Information and Geovisualization 24, 2 (Summer 1987), pp. 40-73. DOI= http://dx.doi.org/10.3138/J166-635U-7R56-X2L1.

[5] Ooms, K., De Maeyer, P., Fack, V., Van Assche, E., and Witlox, F., 2012. Interpreting maps through the eyes of expert and novice users. International Journal of Geographical Information Science 26, 10, pp. 1773-1788. DOI= http://dx.doi.org/10.1080/13658816.2011.642801.

[6] Bulling, A., Ward, J.A., Gellersen, H., and Tröster, G., 2011. Eye Movement Analysis for Activity Recognition Using Electrooculography. IEEE Transactions on Pattern Analysis and Machine Intelligence 33, 4, pp. 741--753.

[7] Stenneth, L., Wolfson, O., Yu, P.S., and Xu, B., 2011. Transportation mode detection using mobile phones and GIS information. In Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (Chicago, Illinois2011), ACM, pp. 54-63. DOI= http://dx.doi.org/10.1145/2093973.2093982.

[8] Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., and Van De Weijer, J., 2011. Eye Tracking - A Comprehensive Guide To Methods And Measures. Oxford University Press, New York.

[9] Schmidt, A., 2000. Implicit human computer interaction through context. Personal Technologies 4, 2-3, pp. 191-199.

[10] Kiefer, P. and Giannopoulos, I., 2012. Gaze map matching: mapping eye tracking data to geographic vector features Proceedings of the 20th International Conference on Advances in Geographic Information Systems, ACM, New York, NY, USA, pp. 359-368. DOI= http://dx.doi.org/10.1145/2424321.2424367.

[11] Kiefer, P., Giannopoulos, I., and Raubal, M., 2013. Where am I? Investigating map matching during self-localization with mobile eye tracking in an urban environment. Transactions in GIS (in print)

[12] Rayner, K., 1998. Eye movements in reading and information processing: 20 years of research. Psychological Bulletin 124, 3 (Nov), pp. 372-422.

[13] Zangemeister, W., Sherman, K., and Stark, L., 1995. Evidence for a global scanpath strategy in viewing abstract compared with realistic images. Neuropsychologia 33, 8, pp. 1009-1025.

[14] Chang, C.-C. and Lin, C.-J., 2011. LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) 2, 3, pp. 1-27.