Inferring Human Knowledgeability from Eye Gaze in Mobile Learning Environments

Oya Celiktutan and Yiannis Demiris

Personal Robotics Laboratory,
Department of Electrical and Electronic Engineering,
Imperial College London, UK
{o.celiktutan-dikici,y.demiris}@imperial.ac.uk

Abstract. What people look at during a visual task reflects an interplay between ocular motor functions and cognitive processes. In this paper, we study the links between eye gaze and cognitive states to investigate whether eye gaze reveal information about an individual's knowledgeability. We focus on a mobile learning scenario where a user and a virtual agent play a quiz game using a hand-held mobile device. To the best of our knowledge, this is the first attempt to predict user's knowledgeability from eye gaze using a noninvasive eye tracking method on mobile devices: we perform gaze estimation using front-facing camera of mobile devices in contrast to using specialised eye tracking devices. First, we define a set of eye movement features that are discriminative for inferring user's knowledgeability. Next, we train a model to predict users' knowledgeability in the course of responding to a question. We obtain a classification performance of 59.1% achieving human performance, using eye movement features only, which has implications for (1) adapting behaviours of the virtual agent to user's needs (e.g., virtual agent can give hints); (2) personalising quiz questions to the user's perceived knowledgeability.

Keywords: assistive mobile applications, noninvasive gaze tracking, analysis of eye movements, human knowledgeability prediction

1 Introduction

Interactive intelligent systems are becoming a ubiquitous part of everyday life, motivated by numerous practical applications in web services, healthcare, education, and much more. In such applications, effective modelling of human behaviours and cognition is essential to build adaptation and personalisation mechanisms. Interaction logs are generally not adequate for genuinely interpreting human behaviours; tasks such as problem solving and reading are hard to be assessed based on verbal protocols [17, 5]. In addition, the more information exchanged between the user and the system through multiple modalities, the more versatile, efficient and natural the interaction becomes [17].

Eye gaze has been frequently studied in interactive intelligent systems as a cue for inferring user's internal states. Eye movements directly reflects what is at the centre of an individual's visual attention, and are linked to cognitive processes in the mind [19].

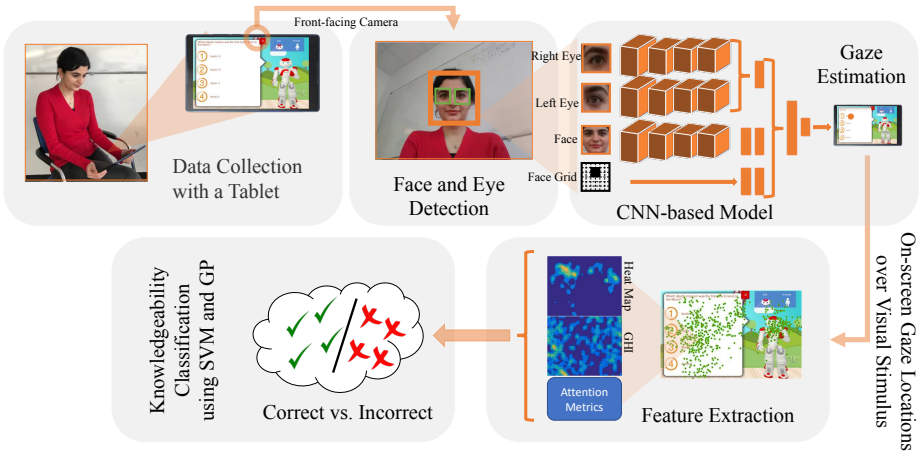


Fig. 1. Automatic inference of user's knowledgeability from eye gaze in mobile learning settings: We focus on a quiz game where we record interactions between a user and a virtual agent using the front facing camera of the tablet. We train a CNN-based system to estimate user's gaze fixations on the tablet screen. From estimated gaze fixations, we define a set of eye movement features to discriminate between users knowing the correct answer and users not knowing the correct answer. We train binary classifiers to predict users' knowledgeability from eye movement features.

A significant body of work has investigated the relationships between eye movements and cognitive processes to provide an understanding into memory recall [8], cognitive load [18], interest [13], level of domain knowledge [9], problem solving [5, 11], desire to learn [4], and strategy use in reasoning [20]. However, so far relatively little research [13] has focused on the interpretation of eye behaviours from the perspective of a camera enabled mobile device. This is due to its long list of challenges including randomness in the camera view, camera motion, drastic illumination changes and partially visible faces.

Accurate eye tracking, namely measuring eye gaze precisely to estimate where the participant is looking at on the screen, requires the use of invasive devices that are not practical to use in everyday life. Although not as precise as the eye tracking devices, recent convolutional neural network (CNN) based approaches such as iTracker [12] provide us with a positive outlook for reliably predicting user's gaze fixations from a mobile device's front-facing camera in a non-obstructive manner. Building upon the deep-learning based approaches, this paper sets out to explore the inference of user's knowledgeability from noisy and temporally sparse gaze estimations in a mobile learning setting. Intelligent tutoring systems can significantly benefit from incorporating such reasoning mechanisms to improve system adaptation and enhance learning.

As illustrated in Fig. 1, we focus on an interaction setting where a user plays a quiz game with a virtual agent while being recorded by front-facing camera of the tablet. Once we estimate on-screen gaze locations based on a deep-learning method, we make a simplifying assumption that the user's knowledgeability is binary, i.e. either the user knows correct answer of a question or not. There is a strong correlation between the

correctness of a person's answer and their Feeling of Knowing (FoK) [6]. FoK is a metacognitive state where people assess their knowledge of a subject when being posed a question. During this state, people exhibit certain nonverbal cues through face and voice such as smiles, gaze acts, which are indicator of their knowledgeability [6]. In addition, humans can predict the performance of others solving a multiple choice reasoning task by looking at their eye movements only [21]. Humans make use of high distinctive patterns and gaze dynamics to judge whether a person knows the correct answer or not. Motivated from [6, 21], we extract a multitude of features from eye movements such that gaze behaviours of a person knowing the answer and a person not knowing the answer appear dissimilar by classification schemes.

Overall, the key contributions of this paper are: (1) introducing a novel yet very challenging computer vision problem, predicting human knowledgeability from eye gaze only, and introducing the first automatic vision-based method to detect knowledgeability from eye gaze on tablets without using any additional sensor; (2) introducing the first audio-visual dataset for predicting human knowledgeability in an interactive mobile learning setting; and (3) providing baseline results for knowledgeability classification from eye movement features. Proposed features based on eye gaze only improves the performance by 7% as compared to facial features (i.e. action units).

2 Related Work

This paper focuses on the problem of inferring user's cognitive states from eye movements in mobile learning settings, which lies at the crossroad of various research areas.

Gaze Estimation. The first challenge of building an eye movement analysis framework is to develop a gaze estimation method that can reliably work during unconstrained interactions with a mobile device. There are a few prominent works that perform gaze estimation without using specialised sensors or devices. These methods can be divided into two groups: (i) methods learning a mapping from input space to 3D gaze direction (3D gaze estimation) [22, 24, 23]; and (ii) methods learning a mapping from input space to on-screen gaze location (2D gaze estimation) [10, 12].

Two recent works [10, 12] focused on 2D gaze estimation using hand-held mobile devices such as smart phones, tablets. Huang *et al.* [10] collected a gaze estimation dataset using a tablet in a systematic way, which is called TabletGaze dataset. They asked 51 subjects to gaze at a dot being displayed on the tablet screen while holding the tablet in four different body postures, namely, standing, sitting, slouching and lying. They showed that recordings from hand-held mobile devices are more characterised by partially visible faces, random camera motion and random camera angles (e.g., centred view vs. low-angle view), as compared to recordings from a laptop placed on a desk [24]. For 2D gaze estimation, they compared four generic regression approaches including Random Forests (RF), Support Vector Regression (SVR) in conjunction with commonly used appearance features (e.g., Histogram of Gradient - HoG, Local Binary Patterns - LBP) extracted from eye regions. Following the TabletGaze dataset [10], Krafka *et al.* [12] introduced the largest 2D gaze estimation dataset, called the GazeCapture dataset. The GazeCapture Dataset was collected via a crowd-sourcing service, which enabled gathering data from a large number of subjects (1474 in total)

and 15 different mobile devices (iPhones and iPads only) at four different orientations, resulting in a large variability in illumination, appearance and pose. In addition, the GazeCapture dataset enabled training a CNN model from scratch for gaze estimation. Cross-dataset evaluations showed that the iTracker outperforms the 2D gaze estimation methods in [10] by a large margin.

Eye Movement Analysis. There is a long list of prominent works focusing on eye movement analysis to provide an insight into human cognitive processing [11, 7, 8, 18, 9, 5, 4, 20, 21, 13]. Among these works, [7, 9, 4, 21] collected eye movement data using a specialised eye tracking device, and investigated internal states relevant to knowledgeability and learning. In [7], Broekens *et al.* focused on a scenario where 31 users played a card game against a desktop computer. The card game involved matching numbers, shapes and colours of objects displayed on a total of 12 cards, and selecting out of 3 cards, called Set, based on a set of rules (e.g., all 3 cards have the same shape). If user claimed a Set among 12 cards, the decision made by the user was annotated as *correct* or *incorrect* based on whether the Set identified by the user satisfied all of the rules or not. For automatically classifying user's decisions, Broekens *et al.* first extracted a set of eye movement features such as duration of fixation, attention spread, saccade length, etc., which are widely used in the literature, and then used various traditional classification methods. Using eye movement features only, Bagging resulted in the best performance of 77.9% for predicting whether the user identified a set or not, and MLP resulted in the best performance of 75.1% for classifying the Set identified by the user into *correct* or *incorrect*.

We are aware of only one work that performed gaze estimation from the front facing camera of a mobile device with the goal of inferring user's internal states. Li *et al.* [13] focused on predicting user's interest in an online store setting. They designed a task where they asked 36 users to interact with the Google Play Store and mentally pick up the items that they found interesting. They estimated the on-screen gaze locations from eye regions using a CNN-based method similar to [24, 12]. They recorded calibration data consisting of 13 points from each user prior to the task and during the task, which was further used to fine-tune the gaze estimation method, and collected users' explicit responses after the task was performed. From the estimated on-screen gaze locations, they computed two types of attention metrics: (i) gaze metrics including gaze well time, gaze dwell fraction and gaze time to first visit; and (ii) viewport metrics including viewport time, viewport dwell fraction and viewport time to first visit. They formalised the problem of inferring user's interest as a binary classification task, i.e., predicting whether a user was interested in an item or not, and used the calculated metrics both singly and jointly to compare three traditional classification methods. They obtained the best classification performance of 90.32% with nonlinear Support Vector Machine (SVM) by fusing all of the attention metrics.

Intelligent Tutoring Systems. Personalising system's actions to individual differences is compulsory for achieving good learning outcomes. To this effect, there have been some works (e.g., [2]) that focused on automatically detecting student's mental states such as satisfied, confused or bored. Alyuz *et al.* [2] collected data from 20 students (14-15 years) over the course of several months as part of a math course. Each student worked independently in the class using a laptop, and was recorded using a 3D

camera. From the recordings, they extracted two types of features, namely, appearance features and contextual features. While appearance features were composed of face location, head pose, facial gestures and seven basic facial emotions (e.g., happiness, sadness, etc.), contextual features were extracted from (i) user profiles including age, gender; (ii) session information including video duration, time within a session; and (iii) performance features including number of trials until success, number of hints used, grade. For inferring mental states, they trained Random Forest classifiers for each feature type, where the contextual features yielded better performance in overall (90.89% for assessment sessions).

The work most relevant to ours was proposed by Bourai *et al.* [6]: they developed an automatic method for predicting human knowledgeability from facial features such as head pose, gaze direction, facial action units and audio features such as speaking rate, voice pitch. They collated a dataset of 198 clips from the British Broadcasting Service's University Challenge trivia show. Each clip contained a participant answering a question posed by the moderator, and was further annotated either with *correct* or *incorrect*. Classification results with SVM showed that facial features alone can be informative enough, yielding a performance above chance (56.1%), while combining facial features with audio features significantly improved the performance (67.5%), and outperformed the human performance by a margin of 4%.

Our Work. Similarly to [10, 12], this paper is concerned with 2D gaze estimation. We propose an automatic method for predicting human knowledgeability from estimated on-screen gaze locations. We design a study where we use a tablet-based quiz game. We record users on video while they are playing the quiz game with a virtual agent to answer a set of general knowledge questions. Following [6], we formalise the problem of human knowledgeability prediction as a binary problem, and use a multitude of eye movement features in conjunction with a classification scheme to predict whether a user gives the *correct* answer to a question or not.

Taken together, to the best of our knowledge, we introduce the first dataset for predicting human knowledgeability in mobile learning settings, which we name the PAL M-Learning corpus, and we introduce the first automatic method to detect human knowledgeability from eye movements on tablets without using any additional sensor. Considering the widespread use of mobile devices and the potential of technology-enhanced learning and e-health applications, our work has significant implications for adapting system behaviours to user's individual profiles and needs beyond just clicks.

3 The PAL M-Learning Corpus

We are not aware of an existing dataset that supports research in predicting user's internal states from eye gaze in a mobile learning setting. This section therefore introduces a novel corpus, called PAL Mobile Learning (M-Learning) Corpus. The PAL M-Learning Corpus comprises two datasets: (i) PAL M-Learning Interaction dataset; and (ii) PAL M-Learning Gaze dataset. The PAL M-Learning Interaction dataset consists of audio-visual recordings of 31 users using a tablet to play Quiz Game with a virtual agent, which is further used to develop automatic models for predicting knowledgeability. The PAL M-Learning Gaze dataset is built by following a similar approach as in [12, 10],

namely, it is collected by showing users a sequence of dots one at a time on the screen and recording their gaze using the front-facing camera. We use the PAL M-Learning Gaze dataset to fine-tune gaze estimation models trained using larger datasets [12, 10].

3.1 Data Collection

For collecting data, we used the mobile interactive platform of our PAL (Personal Assistant for healthy Lifestyle) EU H2020 Project[1], which is based on an Android app. This Android app offers a wide range of tools including a virtual diary and several games, from which we focused on a quiz game as an educational tool. The quiz game is played with a virtual agent, and involves user and the virtual agent asking each other multiple choice questions about a topic. As illustrated in Fig. 1, quiz game interface consists three components, namely, a virtual agent, a question box and an information box.

We designed a study using the quiz game with the goal of developing an automatic method for inferring user's internal states. For the quiz game, we generated a question database by randomly selecting questions at different levels of difficulty (i.e., easy, medium, hard) from a trivia question database, called Open Trivia Database [14]. There is a total of 24 question categories in the Open Trivia Database. Since our goal was to observe different internal states from our target users (e.g., confident vs. uncertain), we conducted a preliminary study where we asked a total of 16 participants to assess whether they found each question category easy or hard, and whether they found them boring or interesting. Using the results of the preliminary study, we hand-picked 9 question categories ranging from boring (e.g., Celebrities) to interesting (e.g., Books, Science & Nature) and from easy (e.g., Computers, Mathematics) to hard (e.g., Mythology, Art), and generated a question database for the quiz game by randomly selecting 5 questions from each category, resulting in 45 questions in total.

We used a Lenovo TB2-X30F tablet during the data collection procedure. 31 users (10 of which were females; age ranging from 18 to 37 with a mean of 27) were recorded, and gave written informed consent for their participation. Each user were guided into the experimental room that had natural lighting. The users were asked to sit on a chair, and hold the tablet in the landscape orientation as they felt comfortable. Apart from these, no further instructions were given to the users. We collected interaction videos by asking each user to play the quiz game with the virtual agent for a duration of 12 mins. Each interaction video contained alternately user and the virtual agent asking questions (20 questions in total, half of which were asked by the virtual agent). These questions were randomly selected from the question database generated as explained before. Before we started to record the interaction, we allowed users to play the quiz game for a couple of questions to familiarise themselves with the Android app. This is a common practice widely used to reduce the novelty effect of a new technology. In both sessions, we recorded users on video from the front-facing camera of the tablet at a rate of 30 fps with an image resolution of 1280×800 pixels. Prior to collecting interaction videos, we collected gaze estimation data by asking each user to look at a sequence of red circles - one at a time - shown on the tablet screen. The red circles were appeared at 35 fixed

[1] http://www.pal4u.eu/index.php/project/about/

Fig. 2. Example snapshots from the PAL M-Learning Interaction dataset, which are taken from the perspective of the front-facing camera while users were answering a question posed by the virtual agent using the tablet.

locations and 25 random locations on the tablet screen, each for a duration of 3 s. Following [10], the fixed locations were equally distributed on the tablet screen, arranged in 5 rows and 7 columns and spaced 2.71 cm vertically and 3.10 cm horizontally. This procedure resulted in gaze estimation data for a duration of 3 mins and a total of 60 on-screen locations per user.

From the conducted study, we built two datasets, namely, the PAL M-Learning Interaction dataset and the PAL M-Learning Gaze dataset. The interaction dataset was created by segmenting each interaction video into a set of short clips. Each clip contained either the user or the virtual agent asking a question and the other one responding. More explicitly, a short clip started when a new quiz question together with four choices appeared on the tablet screen, and finished when a response was given by the user / the virtual agent. For further automatic analysis, in this paper, we only focused on the clips where the virtual agent was asking a question and the user was answering. Example snapshots from these clips are shown in Fig. 2. The gaze dataset was generated by segmenting gaze estimation data into a set of short clips, where each clip contained user gazing at an on-screen location for 2 s (i.e., we removed the first 0.5 s and the last 0.5 s to obtain a clean dataset).

3.2 Summary of the Datasets and Statistics

PAL M-Learning Interaction dataset. We removed 5 users and some of the clips from the interaction dataset as they were not usable, for example, users' eyes were not visible. The resulting dataset consists of 27 participants and a total of 242 clips. On average there are 9 short clips per participant, and each clip has a duration ranging from 10.67 s to 39.43 s with a mean duration of 22.6 s.

After the study took place, we conducted a post-study where we collected assessments from a subset of users (9 in total) for each quiz question with respect to three difficulty levels, namely, easy, medium and hard. We labelled each quiz question as easy, medium or hard by taking average of the collected assessments over all the users, where medium appears in the largest number - 106 of the 242 clips contain questions at the medium level of difficulty. For predicting user's knowledgeability, we used interaction logs to annotate each clip based on the user's response. If the user's response to a question is correct, we annotated the corresponding clip with "correct". Otherwise, we annotated the clip with "incorrect". This resulted in 136 incorrect clips and 106 correct clips, which were used for automatically predicting knowledgeability (see Section 6).

In addition to users' responses, we used a crowdsourcing service to collect perceived knowledgeability annotations from external observers. Each clip was viewed and annotated by a total of 5 external observers, according to whether the user in the clip knows the correct answer to the question being asked or not. Observers provided their responses on a 6-point scale ranging from strongly disagree to strongly agree (we did not include neutral as a response). We also included two trapping questions about the question's topic and the person in the clip to filter spam responses, also known as Honeypot technique [15]. We approved responses submitted by each external observer based on their responses to trapping questions before reimbursing them for their time. We computed the inter-observer reliability in terms of Intra-Class Correlation (ICC) [16], where we used ICC(1,k) as in our experiments each target user was rated by a different set of k observers (k = 5), randomly sampled from a larger population of observers (K = 80). We obtained a significant correlation ($ICC(1,k) = 0.69$, $p < 0.0001$), indicating a high-level agreement among observers. Motivated from this, we evaluated human performance for knowledgeability prediction task. We aggregated the responses from multiple external observers by computing the mean, and assigned the aggregated annotations to either "correct" or "incorrect" class. For this challenging task, humans achieved a classification accuracy of 59.2%. However, we observed a bias towards the incorrect class, namely, the respective F-scores ($\times 100$) were 45.2% and 73.2% for the correct and incorrect classes. We further compared these results with machine performance in Section 6.

PAL M-Learning Gaze dataset. The gaze dataset is composed of $103,911$ images captured from 31 users, 6 of which were wearing eye glasses, with corresponding gaze fixation locations. In our experiments, we benefit from the GazeCapture dataset to build a gaze estimation model, and use the PAL M-Learning Gaze dataset to fine-tune the trained model to the Lenovo tablet in landscape mode (see Section 4). As demonstrated in [12], different mobile device brands have different screen sizes and camera-screen configurations, and therefore fine-tuning the generic gaze estimation model to the target device (Lenovo in our case) is compulsory to obtain reliable results.

4 Gaze Estimation

Deep learning has proven to be successful in many end-to-end learning tasks, yielding previously unattainable performances in various challenging computer vision problems. Its performance has been validated for gaze estimation in [24, 12]. For example, the iTracker outperformed the best method using multilevel HoG and RF in [10], reducing the estimation error by 0.59 cm on the TabletGaze dataset. We therefore adopted the iTracker for gaze estimation in this work.

The architecture of the iTracker [12] is illustrated in Fig. 1. The iTracker has four networks in parallel, and takes as an input (1) the image of the eyes; (2) the image of the face; and (3) a binary mask indicating the location of the face in the image (namely, face grid), and outputs the estimated x and y positions of a gaze fixation on the device screen. The eyes are included as individual inputs into the network to allow the network to identify subtle changes, and the weights are shared between the eye networks. In addition to eye images, face image and face grid are given as inputs into the network to

Table 1. Gaze estimation error in cm as Euclidean distance between the true fixation and the estimated position on the Gaze Capture dataset and the PAL M-Learning Gaze dataset.

Gaze Capture Dataset					PAL M-Learning Gaze		
Model	Mobile Phone		Tablet		Model	Tablet	
	error	dot error	error	dot error		error	dot error
Baseline	2.99	2.40	5.13	4.54	iTrackerTF*	5.31	5.23
iTrackerTF	2.91	2.33	4.40	3.95	fc1 features + SVR	6.63	6.56
iTrackerTF*	**2.37**	**1.97**	**3.86**	**3.55**	fc1 + fc2 finetuning	**3.66**	**3.40**

incorporate head pose variations without explicitly estimating the head pose relative to the camera in contrast to [24]. In [12], it was also demonstrated that training the CNN model with sufficient and variable training examples (>1M) can remedy appearance changes due to head pose. The iTracker maps the input data onto a unified prediction space that enables training a single joint model using all the data from 15 different devices, and the model is trained using a Euclidean loss on the x and y gaze positions in the unified prediction space.

In our experiments, as in [12], we used 1,251,983 images (1271 subjects) for training, 59,480 images (50 subjects) for validation and 179,496 images (150 subjects) for testing. Firstly, we followed the implementation described in [12], and trained the model for 150,000 iterations with an initial learning rate of 0.001 and a reduced learning rate of 0.0001 after 75,000 iterations, with a batch size of 256. Similar to [12], we used stochastic gradient descent optimisation method with a momentum of 0.9. However, the trained model yielded worse results as compared to the results presented in [12]. Secondly, we repeated a similar training procedure by (1) applying batch normalisation and (2) using ADAM optimizer with an initial learning rate of 0.001 for 50,000 iterations and a reduced learning rate of 0.0001 for another 50,000 iterations. We called our first and second implementations as iTrackerTF and iTrackerTF*, respectively.

We evaluated the estimation error in terms of Euclidean distance (in centimeters) between the estimated location and the location of the true gaze fixation. We also computed dot error, where we took average of gaze estimations of the model for all the consecutive images corresponding to the same gaze fixation at a certain location. As it was done in the original paper [12], we compared our gaze estimation results with a baseline method that applies support vector regression on features extracted from a pre-trained ImageNet network (called baseline, hereafter). Our gaze estimation results are presented in Table 1: our both implementations of iTracker outperforms baseline, and using batch normalisation and ADAM optimizer further improves the accuracy by a margin of 0.62 cm in mobile phones and 0.54 cm in tablets, respectively (see iTrackerTF vs. iTrackerTF*).

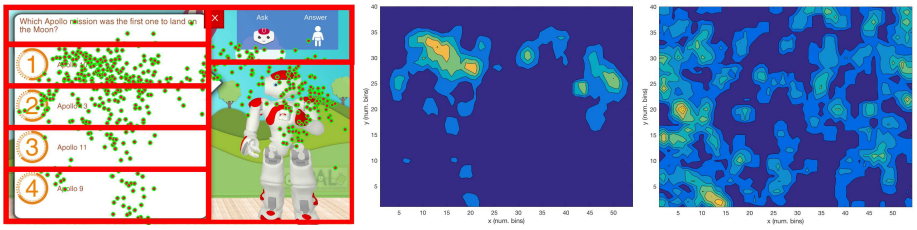
Since our goal was to apply gaze estimation on the PAL M-Learning Interaction dataset in which the full face was not always available, we experimented with different network architectures. We obtained an average error of 4.21 cm using eye images only and 3.45 cm using eye images together with face grid (without face) on the GazeCapture

dataset. We obtained the minimum error when we took into account the face as well (3.11 cm). When the full face was not available, we filled the face by repeating the last row of the face image.

Finetuning iTrackerTF* with the PAL M-Learning Gaze Dataset. The GazeCapture dataset was captured using iPhones and iPads only. On the other hand, in the PAL M-Learning experiments we used a Lenovo tablet that has different camera-screen configuration from an iPad. In addition, out of 1249 subjects, only 225 subjects used iPads during the data collection, resulting in unbalanced data distribution between mobile phones and tablets. In order to obtain on-screen gaze estimations on the PAL M-Learning Gaze and Interaction datasets, we followed [12], and extracted features from the penultimate fully connected layer of the iTrackerTF*. We used these features to train Support Vector Regression (SVR) models using 6-fold cross validation in a subject-independent manner. More explicitly, we divided the dataset into 3 sets: train (20 subjects), validation (5 subjects) and test (6 subjects), and each time we selected the optimum parameters (C and γ) on the validation set, and tested the best model on the unseen subset of test subjects. Using fc1 features in conjunction with SVR resulted in a decrease in the accuracy as compared to the jointly trained network, namely, the error increased from 5.31 cm to 6.59 cm (see fc1 features+SVR results in Table 1). In [12], this approach reduced the error on the TabletGaze dataset [10]. However, only 35 fixed on screen gaze locations were considered in [10], whereas the PAL M-Learning Gaze dataset comprises 25 random locations in addition to 35 fixed locations. We conjecture that the SVR models were not able to generalise due to the randomness in the locations of dots.

We therefore finetuned the iTrackerTF* network using the PAL M-Learning Tablet Gaze Dataset. Using the cropped faces and eyes from OpenFace [3] as inputs, we only updated the weights in the original-dataset-specific layers, namely, in the last two fully-connected layers (fc1 and fc2). As explained above, we followed the same 6-fold cross validation strategy, where we used the validation set to select a learning rate and a stopping criteria, and tested the finetuned model using a non-overlapping set of unseen test subjects each time. This resulted in 6 networks finetuned with a learning rate of 0.0001 for 10,000 iterations, and reduced the error from 5.31 cm to 3.66 cm on the PAL M-Learning Gaze dataset. We further used these networks for gaze estimation on the PAL M-Learning Interaction dataset (see Section 5).

In summary, our implementation has three necessary differences from [12]: (i) we applied batch normalisation; (2) we used ADAM optimizer; and (3) we mirrored right eye images before feeding into the network as it resulted in smaller estimation error. We also showed that fine-tuning the last two layers is a better approach for cross-dataset generalization in our case. Finally, we optimised the trained model for deployment on mobile devices using Tensorflow Mobile. Without face and eye detection, inference using the trained model on a GPU enabled mobile device (namely, Google Pixel C) takes approximately 400-450 ms per frame. As proposed in [12], we then trained a smaller version of the network, namely, reduced the size of the input images from 224×224 to 80×80. The smaller version of the network achieved 80-100 ms per frame, with a slight increase in the mean error on the GazeCapture dataset (i.e., 0.15 cm).



(a) (b) (c)

Fig. 3. (a) Visual stimulus divided into seven region of interests (ROIs) overlaid with sample estimated on-screen locations that are shown with circles in red and green; (b) A heat map computed from the estimated on-screen locations in (a); (c) A gaze history image (GHI) computed from the estimated on-screen locations in (a).

5 Feature Extraction

Once we finetuned the iTracker to the Lenovo tablet using the PAL M-Learning Gaze dataset, we applied it onto the PAL M-Learning Interaction dataset to obtain estimated on-screen gaze locations, and defined a set of eye movement features to model gaze behaviour for knowledgeability classification. Prior to feature extraction, we used OpenFace [3] to detect faces and eyes. In addition, similarly to [6], we used OpenFace [3] to detect and estimate the intensity of facial Action Units (AUs), which are codes that describe certain facial muscle movements (e.g. AU12 is lip corner puller, AU43 is eye blinking, etc.). Since we do not have ground-truth for the true locations of the gaze fixations in the PAL M-Learning Interaction dataset, we manually inspected the results. Our manual inspection showed that eye blinking results in spurious gaze estimations. As a post-processing step, we therefore removed the corresponding gaze estimations where an eye blink occurred. Given gaze fixations estimated in a clip, we extracted three types of eye movement features: (1) attention metrics; (2) heat map; and (3) gaze history images. We also compared eye movement features with the features extracted from AUs.

For computing the attention metrics, we divided the visual stimulus into seven regions of interest (ROIs) including the region enclosing the virtual agent, the region showing whose turn to ask a question, the regions enclosing the quiz question and the four options as shown in Fig. 3-(a). We computed two attention metrics [13], namely, gaze dwell fraction and gaze dwell time. Let T be the total time spent on the stimulus, given a total of N on-screen gaze estimations along vertical and horizontal directions with corresponding time instants, $\{x_i, y_i, t_i\}$, gaze dwell time is defined the amount of time a user spends viewing a ROI: $GDT(ROI) = \sum_{\{x_i,y_i\} \in ROI} (t_{i+1} - t_i)$. For computing gaze dwell fraction, we took into account the percentage of time a user gazes at a ROI: $GDF(ROI) = \frac{\sum_{\{x_i,y_i\} \in ROI} (t_{i+1} - t_i)}{T}$.

Heat map is one of the most widely used representations to analyse eye movements, which shows how looking is distributed over the stimulus. For each clip, we generated a 54×40 heat map as illustrated in Fig. 3-(b). We first vectorised the resulting heat map

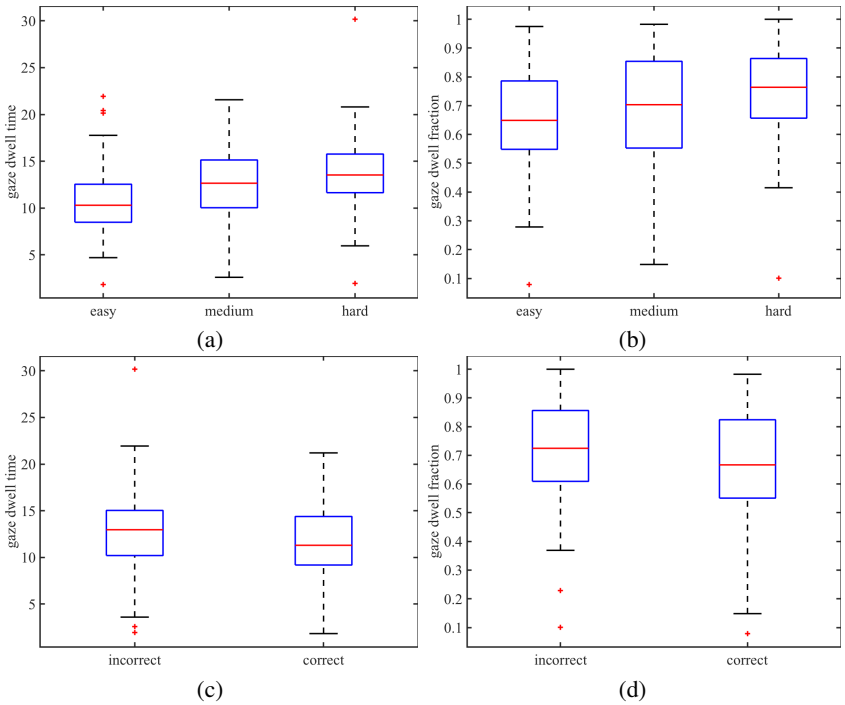


Fig. 4. (a-b) The effect of question's difficulty level on the gaze dwell time and gaze dwell fraction; (c-d) The effect of user's response (incorrect vs. correct) on the gaze dwell time and gaze dwell fraction.

and then applied Principal Component Analysis (PCA) to reduce its dimension from 2160 to 50.

Although heat map is a powerful representation to model gaze behaviours, it does not incorporate the temporal information. On the other hand, as reported in [21], people use gaze dynamics to judge whether a person knows the correct answer or not. Inspired by motion history images [1], we transformed gaze estimations into an image representation while encoding temporal information. We first divided the gaze estimations over time into temporal slices ($\{\Delta t_i\}_{i=1:S}$) of a duration of 1 s, and then for each temporal slice we built a heat map. From time-dependent heat maps, we computed a gaze history image (GHI) per clip as follows. If a user gazes at a certain location within a certain temporal slice Δt_i, we set $GHI_\tau(x, y, \Delta t_i) = \tau$. Otherwise, $GHI_\tau(x, y, \Delta t_i) = \max(0, GHI_\tau(x, y, \Delta t_{i-1}) - 1)$, where τ is set to total number of temporal slices in a clip, namely S. More explicitly, intensity at a certain location is a function of recency of gaze fixation, where higher intensity values are associated with a more recent gaze fixation. A GHI is illustrated in Fig. 3-(c), where more recent gaze fixations are highlighted in yellow, and dark blue regions correspond to areas that were never visited throughout the clip. As we did for the heat map, we vectorised the resulting gaze history image and applied PCA to reduce its dimension to 50.

Taken together, we extracted a total of 114 eye movement features, consisting of 7 gaze dwell time features, 7 gaze dwell fraction features, 50 heat map features and 50 gaze history image (GHI) features. In Fig. 4, we presented the impact of question difficulty level (i.e., easy, medium, hard) on the gaze dwell time and gaze dwell fraction (we only considered the 5 ROIs associated with the question box). One can clearly observe that (i) gaze dwell fraction increases as the question's difficulty level increases; and (ii) people knowing the correct answer tend to spend less time on the question box.

In addition to eye movement features, for comparison purposes, we extracted features from detected facial action units (AUs). We considered that an AU was successfully detected when the confidence score was higher than a threshold, and took into account a total of 7 AUs associated with upper face muscle movements only, including inner brow raiser (AU1), outer brow raiser (AU2), brow lowerer (AU4), upper lid raiser (AU5), cheek raiser (AU6), lid tightener (AU7), and blink (AU45). Then we counted the number of occurrences of each AU and converted these numbers into a histogram for each clip. We also took average of intensity values of each AU over the whole clip, resulting in a total of 14 features.

6 Knowledgeability Prediction

We provided the baseline results with Support Vector Machines (SVM) and Gaussian Processes (GP) as classification schemes. We trained a nonlinear SVM with a Radial Basis Function (RBF) kernel to discriminate correct and incorrect samples. We optimized the parameters in a subject-independent fashion. More explicitly, we evaluated the classification performance using a double cross validation approach. This is a common practice to ensure better generalizablity of the trained models to the unseen subjects. In the outer loop, each time we used all the data from one participant for testing, and all the data from the remaining 26 participants for training and validation. In the inner loop, we selected the best parameters (C and γ) on the training and validation sets using a leave-one subject out validation approach. For GP, we adopted single leave-one-subject-out cross validation approach, and used RBF kernel variety. We randomly subsampled from the class having larger number of samples in order to balance the data prior to training both of the classifiers, and reported classification results as average F-Score/accuracy over multiple runs.

In Table 2, we presented our classification results for each feature type. As mentioned before, this is the first attempt to predict human knowledgeability from eye movements without using any specialised device. We are aware of only one work that proposed an automatic method for predicting human knowledgeability from observable visual cues [6]. However, their work differs from our work along three aspects: (1) they focused on a different setting; (2) they used clips recorded from a static, third person vision perspective; and (3) they extracted visual features from facial action units, head pose and 3D gaze direction. Nevertheless, we compared our classification accuracy with the proposed method in [6]. Looking at Table 2, using heat map only in conjunction with SVM yielded the best classification accuracy of 59.1%, and outperformed the method based on visual features in [6] (56.1%) in a more challenging scenario.

Table 2. Knowledgeability classification in terms of F-Score ($\times 100$) and average accuracy (ACC%) on the PAL M-Learning Interaction dataset. The best results are highlighted in bold. (DwellTime: gaze dwell time; DwellFrac.: gaze dwell fraction; HeatMap: heat map; HistoryImage: gaze history image; ActionUnit: facial action units)

Feature Type	Support Vector Machines (SVM)			Gaussian Processes (GP)		
	F-Score		ACC	F-Score		ACC
	incorrect	correct	ave.	incorrect	correct	ave.
Human	73.2	45.2	59.2	73.2	45.2	59.2
DwellTime	58.9	**57.4**	58.2	58.0	**56.4**	**57.2**
DwellFrac.	57.9	50.5	54.5	57.2	51.4	54.5
HeatMap	**65.3**	50.0	**59.1**	**62.6**	50.0	**57.2**
HistoryImage	47.3	45.3	46.3	46.8	45.8	46.3
ActionUnit	52.7	45.8	49.5	53.1	50.4	51.8
VisualFeatures [6]	-	-	56.1	-	-	-

We observed three trends from Table 2. Firstly, SVM worked slightly better as compared to GP. Secondly, gaze dwell time and heat map features yielded better results with both classifiers as compared to gaze dwell fraction and gaze history image features. Finally, in our setting, action unit features performed just above chance, namely, yielding a classification accuracy of 51.8%.

7 Conclusion and Future Work

In this paper, we introduced a novel computational approach to the problem of predicting human knowledgeability from eye movements without using any additional sensor, in mobile learning settings. We designed a study using a mobile interactive platform, where we recorded users on video while they were playing quiz game with a virtual agent to answer a set of general knowledge questions. We then formalised the problem of human knowledgeability prediction as a binary problem, and used a multitude of eye movement features in conjunction with SVM and GP to predict whether a user gave the *correct* answer to a question or not. Our results showed that heat map in conjunction with SVM was able to discriminate users clicking the correct response and users clicking the incorrect response with an accuracy of 59.1%. We believe that combining eye movements together with action units will help to improve the accuracy. From our manual inspection of the recorded data, we also observed that people tend to divert their attention from the tablet when thinking or bored. Eye contact [25] might be an important cue to detect in these situations. Our research reported in this paper has demonstrated results that warrant further investigation of non-obstructive eye movement analysis to adapt system behaviours to users' individual profiles.

Acknowledgements

This work was funded by the Horizon 2020 Framework Programme of the European Union under grant agreement no. 643783 (project PAL).

References

1. Ahad, M.A.R., K. Tan, J., Kim, H., Ishikawa, S.: Motion history image: Its variants and applications **23**, 255–281 (03 2010)
2. Alyuz, N., Okur, E., Oktay, E., Genc, U., Aslan, S., Mete, S.E., Stanhill, D., Arnrich, B., Esme, A.A.: Towards an emotional engagement model: Can affective states of a learner be automatically detected in a 1: 1 learning scenario. In: Proceedings of the 6th Workshop on Personalization Approaches in Learning Environments (PALE 2016). 24th conference on User Modeling, Adaptation, and Personalization (UMAP 2016), CEUR workshop proceedings, this volume (2016)
3. Baltrušaitis, T., Robinson, P., Morency, L.P.: Openface: an open source facial behavior analysis toolkit. In: IEEE Winter Conference on Applications of Computer Vision (2016)
4. Baranes, A., Oudeyer, P.Y., Gottlieb, J.: Eye movements reveal epistemic curiosity in human observers. Vision Research **117**(Supplement C), 81 – 90 (2015)
5. Bednarik, R., Eivazi, S., Vrzakova, H.: A Computational Approach for Prediction of Problem-Solving Behavior Using Support Vector Machines and Eye-Tracking Data, pp. 111–134. Springer London, London (2013)
6. Bourai, A., Baltrušaitis, T., Morency, L.P.: Automatically predicting human knowledgeability through non-verbal cues. In: International Conference on Multimodal Interaction. pp. 60–67. ICMI 2017, ACM, New York, NY, USA (2017)
7. Broekens, D.J., W. A. Kosters, W., Vries, T.D.: Eye movements disclosure decisions in set. In: Benelux Conference on Artificial Intelligence. pp. 29–30 (2009)
8. Bulling, A., Roggen, D.: Recognition of visual memory recall processes using eye movement analysis. In: Proceedings of the 13th International Conference on Ubiquitous Computing. pp. 455–464. UbiComp '11, ACM, New York, NY, USA (2011)
9. Cole, M.J., Gwizdka, J., Liu, C., Belkin, N.J., Zhang, X.: Inferring user knowledge level from eye movement patterns. Information Processing & Management **49**(5), 1075 – 1091 (2013)
10. Huang, Q., Veeraraghavan, A., Sabharwal, A.: Tabletgaze: dataset and analysis for unconstrained appearance-based gaze estimation in mobile tablets. Machine Vision and Applications **28**(5), 445–461 (Aug 2017)
11. Knoblich, G., llinger, M., Spivey, M.: Tracking the eyes to obtain insight into insight problem solving (07 2005)
12. Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W., Torralba, A.: Eye tracking for everyone. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
13. Li, Y., Xu, P., Lagun, D., Navalpakkam, V.: Towards measuring and inferring user interest from gaze. In: Int. Conf. on World Wide Web Companion. pp. 525–533. WWW '17 Companion, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland (2017)
14. Open Trivia DB: Free to use, user-contributed trivia question database. opentdb.com, online; accessed 21 Feb 2018
15. Quoc Viet Hung, N., Tam, N.T., Tran, L.N., Aberer, K.: An evaluation of aggregation techniques in crowdsourcing. In: Lin, X., Manolopoulos, Y., Srivastava, D., Huang, G. (eds.) Web Information Systems Engineering – WISE 2013. pp. 1–15. Springer Berlin Heidelberg, Berlin, Heidelberg (2013)
16. Shrout, P., Fleiss, J.: Intraclass correlations: Uses in assessing rater reliability. Psychology Bull. (Jan 1979)
17. Surakka, V., Illi, M., Isokoski, P.: Voluntary eye movements in human-computer interaction. In: The Mind's Eye, pp. 473 – 491. North-Holland, Amsterdam (2003)

18. Tessendorf, B., Bulling, A., Roggen, D., Stiefmeier, T., Feilner, M., Derleth, P., Tröster, G.: Recognition of Hearing Needs from Body and Eye Movements to Improve Hearing Instruments, pp. 314–331. Springer Berlin Heidelberg, Berlin, Heidelberg (2011)
19. Underwood, G.: Cognitive Processes in Eye Guidance. Oxford University Press (2005)
20. Vendetti, M.S., Starr, A., Johnson, E.L., Modavi, K., Bunge, S.A.: Eye movements reveal optimal strategies for analogical reasoning. Frontiers in Psychology **8**, 932 (2017)
21. van Wermeskerken, M., Litchfield, D., van Gog, T.: Eye see what you are doing: Inferring task performance from eye movement data. In: European Conference on Eye Movements (2017)
22. Wood, E., Bulling, A.: Eyetab: Model-based gaze estimation on unmodified tablet computers. In: Proceedings of the Symposium on Eye Tracking Research and Applications. pp. 207–210. ETRA '14, ACM, New York, NY, USA (2014)
23. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: It's written all over your face: Full-face appearance-based gaze estimation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 2299–2308 (July 2017)
24. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence **PP**(99), 1–1 (2018)
25. Zhang, X., Sugano, Y., Bulling, A.: Everyday eye contact detection using unsupervised gaze target discovery. In: Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology. pp. 193–203. UIST '17, ACM, New York, NY, USA (2017)