

# Eye-2-I: Eye-tracking for just-in-time implicit user profiling

Keng-Teck Ma  
Agency for Science,  
Technology And Research  
1 Fusionopolis Way  
Singapore 138632  
makt@i2r.a-star.edu.sg

Terence Sim  
National University of  
Singapore  
13 Computing Drive  
Singapore 117417  
tsim@comp.nus.edu.sg

Qianli Xu  
Agency for Science,  
Technology And Research  
1 Fusionopolis Way  
Singapore 138632  
qxu@i2r.a-star.edu.sg

Mohan Kankanhalli  
National University of  
Singapore  
13 Computing Drive  
Singapore 117417  
mohan@comp.nus.edu.sg

Liyuan Li  
Agency for Science,  
Technology And Research  
1 Fusionopolis Way  
Singapore 138632  
lyli@i2r.a-star.edu.sg

Rosary Lim  
Agency for Science,  
Technology And Research  
1 Fusionopolis Way  
Singapore 138632  
rosary-lim@i2r.a-  
star.edu.sg

## ABSTRACT

For many applications, such as targeted advertising and content recommendation, knowing users' traits and interests is a prerequisite. User profiling is a helpful approach for this purpose. However, current methods, i.e. self-reporting, web-activity monitoring and social media mining are either intrusive or require data over long periods of time. Recently, there is growing evidence in cognitive science that a variety of users' profile is significantly correlated with eye-tracking data. We propose a novel just-in-time implicit profiling method, Eye-2-I, which learns the user's interests, demographic and personality traits from the eye-tracking data while the user is watching videos. Although seemingly conspicuous by closely monitoring the user's eye behaviors, our method is unobtrusive and privacy-preserving owing to its unique characteristics, including (1) fast speed - the profile is available by the first video shot, typically few seconds, and (2) self-contained - not relying on historical data or functional modules. As a proof-of-concept, our method is evaluated in a user study with 51 subjects. It achieved a mean accuracy of 0.89 on 37 attributes of user profile with 9 minutes of eye-tracking data.

## Categories and Subject Descriptors

H3.4 [Systems and Software]: User profiles and alert services

## General Terms

Human factors; Classification;



Figure 1: A screen capture from our demo video. This top shows the eye fixation on the video. The bottom shows the output of Eye-2-I. The system is able to infer the demographic, personality and interests from the user's eye-tracking data.

## Keywords

eye-gaze, profiling, classification

## 1. INTRODUCTION

Providing personalized services, such as targeted advertising, content recommendation and multimedia retrieval [29], has been important to users (survey by Adobe [1]); and by natural extension, service providers. User profiling has been proposed to tackle this issue, whereby personal information (e.g., interests, traits, and demographic data) are inferred ei-

---

ther directly from user feedback or inferred indirectly from past behavior record, such as web-activity or social media history. However, such practices are severely hindered by the availability of historical data, data impurity, and privacy and security concerns. It remains to be addressed how to make timely inferences of user profiles based on a data collection process that is unobtrusive to the users.

*It is our vision that answers to this question lies in deeper understanding of users' natural behaviors in the respective interaction context in a just-in-time manner.*

It is well-established in the cognitive science and psychology communities that our traits and interests significantly influence our subconscious responses. Inspired by implicit tagging where the meta-data about a multimedia content is derived from the observer's natural response [31], we propose to infer the user's traits and interests from the eye-tracking data.

Eye-tracking data, including fixations, blinks and dilations, captures an automatic and subconscious response, which is influenced by a person's interests [5], traits [9, 13, 34], and attention [3, 11]. In essence, eye-tracking data is heavily influenced by a person's profile. As such, using machine learning techniques, such as supervised learning, these data can be used to infer one's profile.

We are aware that closely monitoring the user's eye-gaze is conspicuous and may lead to privacy and security concerns, *per se*. We boldly utilize this unconventional media in hope of pushing the boundary of interaction design with a better understanding of the latent user needs. Meanwhile, we expect that those concerns would be effectively alleviated if the system can perform accurate profiling within reasonably short period of time (*i.e.*, just-in-time), thus mitigating the need for storing any personal information. In other words, the profiling is conducted on-the-fly, and the lifetime of personal data is strictly confined to a service session, e.g. the duration of a flight. As compared to conventional methods (e.g., self-reporting, web-activity monitoring and social media mining), the method is unobtrusive and privacy-preserving because it does not keep historical, personal information. In addition, services built with this technology should be deployed with explicit consent of users regarding the usage of their eye-tracking data.

To the best of our knowledge, we are the first to propose a user profiling system, Eye-2-I, which uses Eye-tracking data for just-In-time and Implicit profiling to infer a comprehensive set of users' attributes while they are watching a video. The profile is available by the first shot, typically few seconds. With empirical evidence, we demonstrate the capability of using eye-tracking data for inferring the complete users' profile of 8 demographic traits, 3 personality types, 26 topics of interest and emotions. In sum, our method offers three unique features: timeliness, implicitness and comprehensiveness. In the current framework, eye-tracking data is captured using specialized devices (SMI RED 250) to ensure data fidelity. Alternatively, one may use a standard video camera [16] which are equipped on devices such as laptops, tablets, smart-phones, gaming consoles and smart televisions [4]. As accurate eye-tracking can be achieved at more affordable cost [28], we expect eye-tracking technology to be more wide-spread in the near future.

## 2. BACKGROUND

Self-reporting is a simple and direct method for profiling.

It has response time of several minutes and is obtrusive. It is our vision that profiling can be incorporated in natural interactions with a system, e.g. video watching.

Alternatively, profiling can be done using historical web activity data, including views, link-clicks and searches. For example, as users browse Google's partner websites, it stores a HTTP cookie in a user's browser to understand the types of pages that user is visiting, usually called user-tracking. This information is used to show ads that might appeal to the users based on their inferred interest and demographic categories [15].

The social media also provides a rich source of data for user profiling. Kosinski et al. used the history of "Like" in Facebook to infer private traits and attributes [19]. Posting to Twitter can also reveal much about the user's traits such as ethnicity and political affiliation as shown by Pennacchiotti and Popescu [24]. Personality can also be revealed from Twitter's history [26]. Cristani et al. showed that personality traits can also be inferred from one's "favourite" Flickr images [10].

While our proposed method also monitors users' behaviors to infer their profile, we track a different types of behavior, namely eye-tracking data. With eye-tracking, the response time is in seconds and minutes, instead of hours or days as with tracking users' history of web or social media activity. Thus profiles are made available sooner and will be more updated and relevant.

Behaviors can be conscious and purposeful, such as clicking on hyperlink, posting a tweet, tagging an image as a "favourite"; or subconscious responses, such as pupil dilations, blinks and fixations. Conscious behaviors are more robust against irrelevant factors, e.g. environmental noise and lighting changes. But subconscious responses are more resistant to manipulations and deception [25].

Depending on the scope of the behavior, a single user can be identified from their browsers (e.g. with web cookies), user accounts or service sessions. Eye-2-I track eye movement behaviors and store the profile within a service session. The duration of the session is application dependent. By default, the profile is discarded after each session for privacy protection. If privacy is not a concern, for example, in a fully protected or trusted environment, the profile can persist across multiple sessions using any existing methods, e.g. user account or web cookie.

From Table 1, it is clear that Eye-2-I is unique among the various profiling methods. Its unique properties open an entirely new approach to user profiling. This is further elaborated in our example application of personalized in-flight entertainment system in Section 3.

Facial features provides an alternative means of just-in-time profiling implicitly. Personal traits such as gender, age and ethnicity can be inferred from facial features [8]. However, our method can also be used to predict other demographic factors which may not manifest in appearance-based methods, e.g. religiosity. Another clear advantage of using eye-gaze is that transient mental states such as topics of interest can be revealed through interactions between the eye-gaze and regions of interest in the video content. Table 2 shows the comparisons between the two modalities.

Our prior work infers demographic and personality traits from eye-tracking data while users are viewing images [20]. Alt et al. proposed how gaze data on web-pages can be used to infer attention and to exploit this for adaptive content,

	Web	Social Media	Eye-2-I
Response	hours	days	minutes
Behavior	conscious	conscious	subconscious
Scope	browser	account	session

**Table 1: Comparison of the behavior profiling methods. Response refers to the amount of time required to acquire sufficient data for comprehensive profiling. Behavior can be either conscious, e.g. clicking on hyperlink; or subconscious, e.g. pupil dilations, blinks. Scope refers to the scope of the behaviors used to track the users.**

	Eye-gaze	Face	Eye-2-I
Gender	[13, 20]	[8]	Y
Age	[13]	[8]	Y
Ethnicity	[9]	[8]	Y
Personality	[34, 20]	[21]	Y
Religiosity	[20]		Y
Interests	[5]		Y
Field of work/study			Y
Education			Y
Socioeconomic			Y

**Table 2: Comparison of the attributes which are correlated and/or inferred with eye-gaze, face and our proposed system: Eye-2-I. No prior work provides comprehensive user profile from either face or eye-tracking data.**

i.e. advertising [3]. Eye-2-I differs from these work in that we are using eye-tracking data from video-viewing and our output is a comprehensive profile, including topics of interests. This is the first work we know of which infers general topics of interests which may not be present in the visual content. This is different from prior work which infer direct interests in the visual content, e.g. an advertisement banner. Eye-tracking data from video-viewing has temporal ordering across different shots and results in higher accuracies than independent single shot classifiers, as shown in our experimental results.

### 3. EXAMPLE APPLICATION

While the idea of using human eye-gaze behavior to predict user profile is not restricted to specific application scenarios, we have been motivated by a promising use case of personalized in-flight entertainment system.

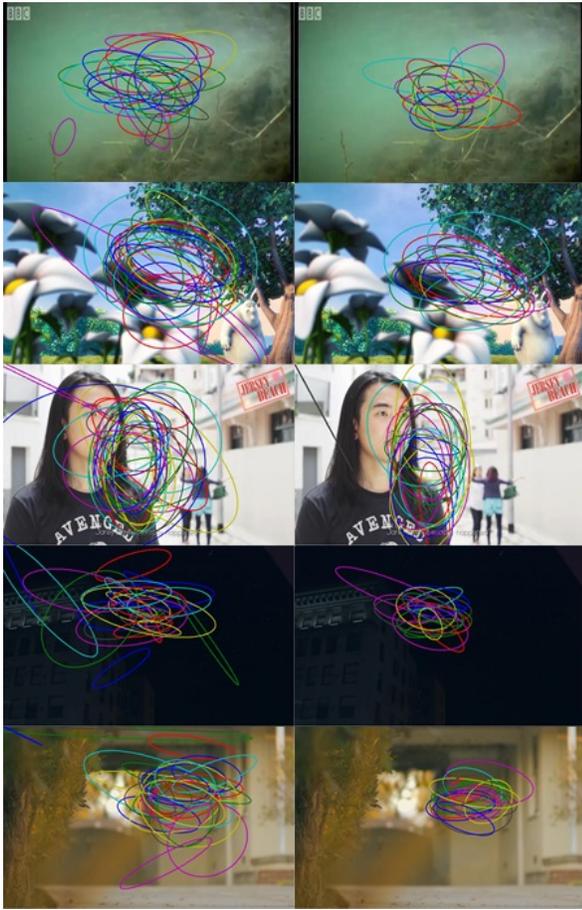
Today, in-flight entertainment system is not personalized for a variety of reasons. Firstly, it is not effective to request users to enter their detailed profiles for the service durations which are on average 2.16 hours for commercial flights [2]. Secondly, as these systems are not the users’ own devices, user-tracking methods such as web cookies is not possible. Thirdly, requiring the users to sign in to their existing social media accounts so as to retrieve their profiles is challenging. Some users do not have relevant accounts; for example they may be too young, are not technologically savvy, or are adverse to social media etc.

Watching videos is a popular activity during a flight and the physical conditions of the personal televisions setup in many commercial planes are favorable to this application. The relatively controlled and consistent settings (i.e. identical screen size, restricted viewing angle and eye-screen distance, partially controllable lighting conditions, etc.) makes it technically possible to collect eye-tracking data with reasonable accuracy. With additional setup of eye-tracking devices, a system collects user eye-tracking data and conducts user profiling just-in-time. Since the eye gaze is implicit and subconscious, there is little to no effort required from the users. The users’ profile information can be used for multiple purposes, such as targeted advertising, content recommendation and personalized multimedia retrieval. Finally, since profiling is performed just-in-time, privacy concerns can be mitigated by removing their profiles from the system once their plane lands. Like other profiling methods, the system can also show the passengers the terms of usage of the eye-tracking data. They may choose to participate or not based on personal preference. Properly applied, the proposed system can greatly enhance the in-flight service level.

### 4. COGNITIVE RESEARCH

Both intuitively and experimentally. eye tracking data, such as fixation durations are also correlated with interests and attention. Rayner’s experiments found that subjects spent more time looking at the type of ad they were instructed to pay attention to [27]. Similarly, Alt et al is able to infer interests from eye-tracking data [3].

In cognitive research, studies also show that different groups of people have different eye movement patterns. Among the most well studied traits are gender, age, culture, intelligence and personality.



**Figure 2: Some examples of fixations differences for gender (left column: female; right column: male). Center of ellipse is the mean position of the fixations of the shot, the shape and size is the covariance. For many shots, female subjects have greater variance in fixations’ positions.**

Goldstein et al. examined the viewing patterns when watching a movie and observed that male and older subjects were more likely to look at the same place than female and younger subjects [13]. In other words, male and older subjects have less variance in their eye-movements. Similarly, Shen et al.’s work on visual attention while watching a conversation shows that the top-down influences are modulated by gender [30]. In their experiments, men gazed more often at the mouth and women at the eyes of the speaker. Women more often exhibited “distracted” saccades directed away from the speaker and towards a background scene element. Again, male subjects have less variance in fixations positions. These findings on gender differences are also found in our dataset, as shown in Figure 2.

Chua et al. measured the eye gaze differences between American and Chinese participants in scene perception [9]. Chinese participants purportedly attend to the background information more than did American participants. Asians attended to a larger spatial region than do Americans [6].

Wu et al. discovered that the personality relates to fixations towards eye region [34].

arts & humanities	automotive	business
finance & insurance	entertainment	Internet
computer & electronics	real estate	local
reference & education	recreation	science
news & current events	telecomms	sports
beauty & personal care	animals	games
food & drink	industries	shopping
photos & videos	lifestyle	travel
home & gardening	social network	society

**Table 3: Topics of interest from Google Ads. The topics will be referenced to by their first words in this paper.**

Vigneau et al.’s study used regression analyses on eye movements as significant predictors of Raven Advanced Progressive Matrices test performance [32]. Raven is a standardized intelligence test. Intelligence is known to be significantly correlated with education level and socioeconomic status.

## 5. DATA COLLECTION

The evaluation dataset is the first multi-modal dataset (facial expressions, eye-gazes and text) coupled with anonymous demographic profiles, personality traits and topics of interest of 51 participants. It is available for non-commercial and not-for-profit purposes.

### 5.1 Participants

Fifty-one participants were recruited for the 1-hour paid experiment from an undergraduate, postgraduate and working adults population. They have perfect or corrected-to-perfect eye-sight and have good understanding of English language.

### 5.2 Procedure

The subjects were asked to view all four videos (with audio) in a free-viewing settings (i.e. without assigned task). Specifically, they were instructed to view the videos as they would watch in their leisure time on their computer or television. Our experiment was approved by the Institutional Review Board for ethical research.

Their eye-gaze data was recorded with a binocular infrared based remote eye-tracking device SMI RED 250. The recording was done at 60Hz. The subjects were seated at 50 centimeters away from a 22 inch LCD monitor with 1680x1050 resolution.

A web-camera is also set up to analyze their facial expressions. The eMotion emotion analyzer tracks the face and returns a streaming probability for neutral, happy, surprise, anger, sad, fear, and disgust [12].

We considered carefully of the trade-off between having more accurate and clean eye-tracking data using physical restrains; and having participants in a more realistic setup with freedom of eye, head and body movements. As our objective is to profile the subjects implicitly and unobtrusively, the subjects were not restrained by any physical contraption, e.g. chin rest or head rest. This setup is different from most other fixation datasets [33].

To obtain good quality eye-tracking data, the subjects were given instructions to keep their eyes on the screen and to remain in a relaxed and natural posture, with minimal movements. We noted that some subjects did not follow

Video	Ratings	Valence	Arousal	PV
<i>Documentary</i>	3.96(0.96)	6.24(1.26)	5.35(1.93)	0
<i>Animation</i>	4.04(0.89)	6.65(1.73)	6.20(1.50)	2
<i>Satire</i>	3.73(1.10)	7.02(1.39)	6.00(1.87)	2
<i>Romance</i>	4.08(0.74)	3.84(1.62)	5.73(1.42)	3

**Table 5: Statistics of participants’ feedbacks. In the first 3 columns, the first number is the mean and the number in parentheses is the standard deviation. The last column, PV, indicates the number of participants who had already viewed the videos prior to the user study.**

the instructions. These subjects were too engaged with the content that they moved unconsciously. For example, a few subjects were laughing heartily with significant head and body movements while watching the *animation* and *satire* videos. Nevertheless, the data collected are of high quality, due to users’ high engagement with the content; multiple calibrations per subjects; and tight control of the calibration process.

### 5.3 Videos

Some videos are more likely than others to elicit eye-gaze behaviors which are suitable profiling of the different attributes. We have carefully selected 4 videos with different genres, number of acts, languages, cast make-up and affect. The characteristics of the videos are summarized in Table 4. The duration of each video was about 10 minutes. All videos were presented to every participant in random order.

### 5.4 User’s feedback

The participants were tasked to answer questions after watching the video: rating (1-5, dislike to like), emotional valence (1-9, sad to cheerful), emotional arousal (1-9, calm to excited). They selected topics which were related to videos from a list (Table 3). The participants were also asked if they had viewed the videos before. Table 5 shows the mean and standard deviations of the feedback for the videos. Only very few participants have viewed the videos before the experiment. The participants also answered questions on their demography and personality (Table 7).

## 6. METHODOLOGY

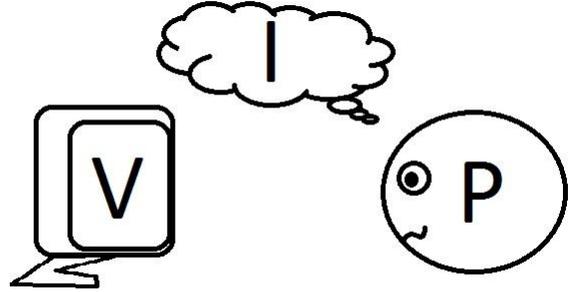
There is abundance of study linking eye-movements with various attributes of user profile (Section 4), however none has attempted to automatically predict the comprehensive user profile from eye-tracking data. Our approach is the first to establish the feasibility of this.

As a proof-of-concept, we made a deliberate choice to use standard supervised machine learning technique, support vector machine (SVM) and simple statistical features to infer the profile from eye-tracking data. This straightforward approach More advanced techniques and features are suggested and discussed in Section 8. Our technical contribution is using the incremental classifiers to improve on the accuracy as compared to single image classifier. The contribution improves accuracy significantly (See Figure 5).

Firstly, we identified the profile’s attributes which are of interests to multimedia applications. Next, we extracted statistical features from the eye-tracking data. Then with these features, we trained the SVM with labeled data for

each shot. Finally, the classification results of each shot is concatenated and used as input feature to the incremental classifier.

### 6.1 VIP model



**Figure 3: The VIP factors which will affect eye-gaze. V: visual stimuli, I: intent and P: person. All 3 factors will affect the eye-gaze of the viewer. However, in current research models, only one or two of the factors are considered.**

In our prior work, we proposed the VIP framework (Figure 3) which characterizes computational eye-gaze research [20]. It states that eye-gaze is a function of Visual stimulus, Intents and Personal traits. By *visual stimuli* we include any visual modality, such as traditional images and videos, and also novel mediums like 3D images and games. By *intent* we refer to the immediate state of the mind such as purpose of viewing the stimuli, the emotions elicited by the stimuli, etc. Finally, by *person* we mean the persistent traits of the viewer of the visual stimuli, including identity, gender, age, and personality types.

We formulate our application as:

$$\{I, P\} = f^{-1}(E_V) \quad (1)$$

That is for a video shot ( $V$ ), our Eye-2-I algorithms:  $f^{-1}$  infers interests ( $I$ ) and personal traits ( $P$ ) from the eye-gaze data of each shot ( $E_V$ ).

This formulation succinctly summarizes our novel contributions of inferring **both** interests and personal traits from eye-tracking data.

### 6.2 Traits and interests profiling

We identify the following personal traits for profiling: gender, age-group, ethnicity, religion, field of study/work, highest education qualifications and income groups (personal and household). Many of these traits are used in market segmentation and targeted advertising [14].

Jacob et al. found that advertisements were evaluated more positively the more they cohered with participants’ personality types [17]. Personality are also useful with other applications such as movie recommendation [22]. Eye-2-I infers the Carl Jung’s personality types: Extrovert/Introvert, Sensing/Intuition and Thinking/Feeling for the eye-gaze [18].

For the inference of interests, we have selected the same set of categories as Google Ads system shown in Table 3.

### 6.3 Features extraction

Video	Genres	Acts	Languages	Cast	Affect
1	documentary, animal	1	British English	1 man	Neutral, Calm
2	animation, animal, comedy	3	no speech	4 animals	Cheerful, Excited
3	local, satire, television	multiple	multilingual	multiple persons	Cheerful, Excited
4	romance	1	American English	1 man, 1 woman	Sad, Neutral

Table 4: Summary of the characteristics of the videos.

$\bar{x}, \bar{y}$	mean value of the coordinates, $x, y$ , of the fixations
$d$	mean value of the fixations’ duration
$\sigma_x, \sigma_y, \sigma_{xy}$	triangle matrix of covariance of $x$ and $y$
$\sigma_d$	standard deviation of duration
$\hat{p} = \sigma_p / \bar{p}$	normalized pupil dilation
$x_1, y_1, d_1$	1 <sup>st</sup> fixation
$x_2, y_2, d_2$	2 <sup>nd</sup> fixation
$x_L, y_L, d_L$	fixation with the longest duration
$D$	total fixation duration
$N$	number of fixations

Table 6: Statistical features used.

The statistical features are extracted from the eye-tracking data of each shot for classification. Nineteen features are identified as in [20] for inferring personal traits for images viewing activity. These features are found to be different among people with different traits from prior research [7, 9, 13]. The features are shown in Table 6.

We considered extracting only one feature vector from eye-tracking data of each video. However, the eye fixations over the entire video is too diverse and will not be useful for our purpose. On the other extreme, eye fixations on a single frame is insufficient for classification. Therefore, we adopted “shot” as the basic unit for feature extraction and annotation, where a shot means a video clip that is continuously shown without significant change of shooting orientation. Shot segmentation allows eye fixation data on each set of content-coherent and semantic-similar frames to be classified independently.

## 6.4 Incremental classification

The main challenge for user profiling from eye-gaze is the strong dependency of the visual and semantic content. Only some visual content are suitable for inferring certain attributes, e.g. *gender* [20]. We overcome this by using an ordered ensemble of classifiers as explained below.

We perform supervised learning to classify the extracted features to the respective attributes (demographic, personality, interests) for every shot. For any single shot, the classification accuracy are low for some attributes and better for others (results for single-shot classifiers are in the supplementary material). Instead of returning the mixed results, we can exploit the temporal ordering of the shots to incrementally improve the results by combining the classification results of the same attribute from previous shots from the same video. To this end, we implemented a supervised meta-classifier which treats the ordered set of shot classification results as the input features. The size of feature vector is equal to the current shot index.

The meta-classifiers learn the relative “weights” of each individual shots with respect to the attribute being classi-

Traits	Majority( <i>M</i> Ratio)	Minority
<i>gender</i>	female(0.59)	male(0.41)
<i>agegroup</i>	$\leq 24$ (0.76)	$\geq 25$ (0.24)
<i>ethnicity</i>	chinese(0.69)	others(0.31)
<i>religiosity</i>	religious(0.67)	none(0.33)
<i>specialty</i>	sci&eng(0.65)	others(0.35)
<i>education</i>	tertiary(0.69)	post-grad(0.31)
<i>income</i>	0-999(0.71)	$\geq 1000$ (0.29)
<i>household</i>	1-4999(0.75)	$\geq 5000$ (0.25)
<i>ei</i>	Introvert(0.53)	Extrovert(0.47)
<i>sn</i>	Sensing(0.57)	Intuition(0.43)
<i>tf</i>	Feeling(0.63)	Thinking(0.37)

Table 7: Grouping of traits for the dataset. The numbers in parentheses show the distribution of the traits.

fied. As more shots are shown to the users, the incremental classifiers have more information to infer the attribute correctly. Hence, this method improves classification accuracy when the video contains sufficient shots with relevant visual content.

## 7. EMPIRICAL EXPERIMENTS

The objective of the experiments is to validate our claim that user profile can be accurately inferred from eye-tracking data. We classify each attribute (trait or topic of interest) into 2 possible classes. For topics of interest, the 2 classes are “interested” and “not-interested”. For traits with multiple possible values, we consolidated them into 2 groups for a more even distribution. Table 7 shows the groupings of traits and the distributions of the population. In the table, *M*Ratio is defined as the fraction of the majority class in the population (e.g. *female* = 0.59).

For each shot in each video, the statistical vectors are extracted from the fixations of each person. A linear support vector machine (SVM) classifier was trained per shot per attribute. We used the standard linear SVM classifier in the Matlab Biometric Toolbox, with the default parameters and auto-scaling.

Using incremental classification method, the ordered classification results from the previous and current shots formed the input feature vector for the meta-classifier, also a SVM (same implementation and parameters as the per-shot classifiers). Leave-one-out cross validation was used to evaluate the meta-classifiers, i.e. a single subject is left out of the training set in each round. Figure 4 shows the example of classifying *gender* trait for *satire* video.

For readers who are interested in the other classification metrics, we have included the full experimental results for

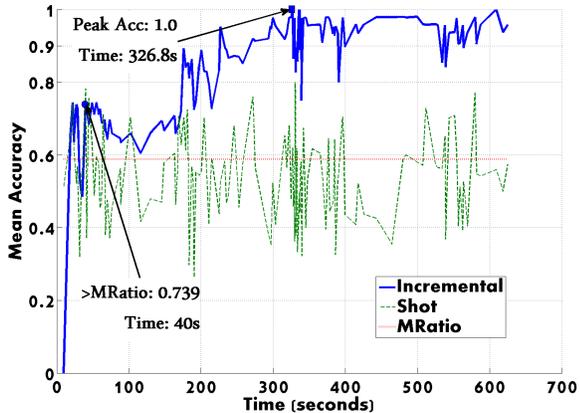


Figure 4: Mean Accuracy vs Time plot for *gender* trait classification with *satire* video. Except for a few shots at the beginning of the video, *Shot* classifiers’ accuracies are lower than *Incremental*. *Incremental* classifier’s accuracies improve over time. After 40 seconds, its accuracy is consistently higher than *MRatio* as defined in Table 7. It peaked at perfect accuracy after 326.8 seconds, after which accuracy of  $> 0.9$  was sustained with a few exceptions.

both *Shot* and *Incremental* classifiers and their classification metrics, that is *sensitivity*, *specificity*, *precision*, *recall* and *F1* score in the supplementary material (<http://1drv.ms/1vUPMEu>). The correlation analysis (*P* and *R*) between the 19 features and 37 attributes for each shot (537) are also included. Our preliminary analysis shows significant correlation of the certain set of attributes and features for many shots, for example  $\sigma_x$  and *gender*. This result is consistent with prior cognitive science study [30, 13]. While more detailed analysis is beyond the scope of this paper, interested readers are strongly encouraged to analyze our supplementary materials.

## 7.1 Data preparation

We refer to personal traits (e.g. gender, age, personality types) and topics of interest (e.g. animals, computers) collectively as *attributes*.

For the presentation of the experimental results, the attributes are abbreviated as: field of study/work  $\Rightarrow$  *specialty*, highest education qualifications  $\Rightarrow$  *education*, personal income  $\Rightarrow$  *personal*, household income  $\Rightarrow$  *household*; extrovert/introvert  $\Rightarrow$  *ei*; sensing/intuition  $\Rightarrow$  *sn* and thinking/feeling  $\Rightarrow$  *tf*.

The recorded eye-gaze data were preprocessed by the vendor’s software to extract the fixations. Fixations from the preferred eye as indicated by the subjects were used. Missing eye-tracking data was ignored for the computation of the statistics.

For the inference of topics of interest, only 1 participant indicated interests in *real* estate. Hence, this topic is removed for consideration, leaving 26 topics of interest.

Each video is manually segmented into shots. The number of shots are: 107, 153, 135 and 140 respectively.

## 7.2 Experimental Results

First, we show the overall accuracy for our classifiers. As

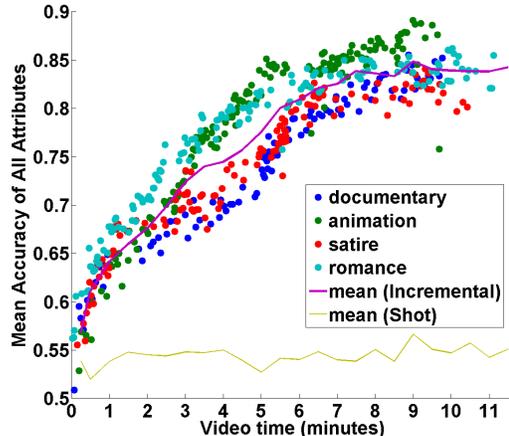


Figure 5: The lines plots the mean accuracy for all videos for *Incremental* and *Shot* respectively. *Incremental* classifier is more accurate than *Shot*. Mean *Incremental* accuracy for all attributes, all videos, at 15, 30, 60, 120, 240, 480, 960 seconds are 0.57, 0.61, 0.64, 0.68, 0.74 and 0.84 respectively. Mean *Incremental* accuracy of all attributes vs video time plot for each video are plotted as dots.

there is no comparable prior work, chance (0.5) is the only possible baseline comparison. Figure 5 shows that the mean accuracy for all of our classifiers are greater than chance. With more data, the mean *Incremental* accuracy steadily increases and peaks at 450 seconds (7:30 minutes) at 0.84. On average, *animation* is most accurate; it also reached 0.89 accuracy with 539.5 seconds (9 minutes) of data.

Next we investigate the *Incremental* accuracy for individual attributes. We chose *MRatio* for baseline comparison, where *MRatio* is defined as the fraction of the majority class in the population, as shown in Table 7. Assuming that the distributions of population who will watch the video is the only information known in advance, *MRatio* is the best accuracy from any deterministic classifier. Another possible baseline is chance (0.5) but *MRatio* is always equal or higher than that.

From Figure 5, it is clear that the *Incremental* accuracy is positively correlated with the amount of data. So we consider the scenario with the maximum amount of data, which is at the end of each video. The end accuracy of each video is thusly computed.

In Figures 6 and 7, for each attribute, we compare *MRatio* against end accuracy of each video. For every attribute, there is at least one video which end accuracy is higher than *MRatio*. Some of the attributes such as *industries* and *telecomms*, have *MRatio* which are higher than 0.9. Despite that, the incremental classification method is still better than *MRatio* for at least one video.

Furthermore, we observe that the best accuracies are in the similar range with the widely reported work by Kosinski et al. [19]. Notwithstanding the differences between types of behavior tracked (Facebook’s “Like” vs eye tracking data) and the amount of training data, accuracies of higher than 0.9 were obtained for *gender* and *ethnicity* for both work.

In addition, these results enable us to choose the best video for a specific attribute. For example, at the end of

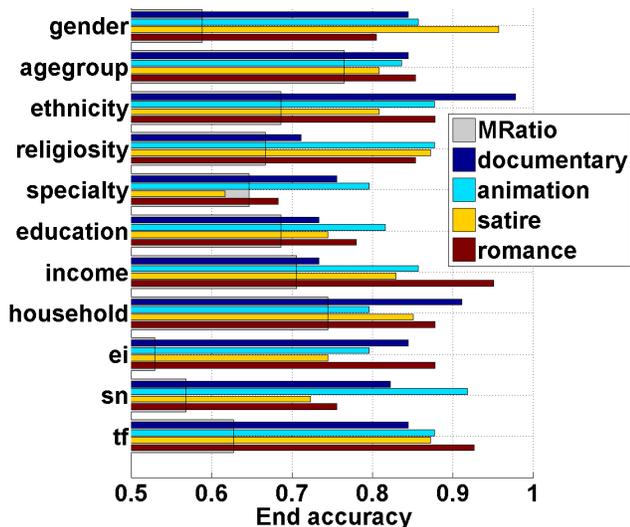


Figure 6: Traits vs end accuracy of each video and *MRatio*.

the *romance* video, *lifestyle* and *games* are perfectly predicted. A practical example is automotive advertising. The *romance* video will be most useful as it has highest accuracy for *income* and *automotive* topic of interests. Advertisers can then target the users who have higher income and interests in their product category. Since *romance* also has highest accuracy for the Thinking/Feeling personality type, the advertiser can display the more personalized advertisements (e.g. appeal to logic or emotion) based on the predicted personality type [17].

## 8. DISCUSSIONS

Our experimental results, while promising, have several limitations and unanswered questions. First, only binary classification is performed. While this is a good fit for some attributes, there are many attributes for which multi-class is more suitable.

Second, we have not made in-depth investigation on the generalizability of the method. Our sample size of 51 participants and 4 videos is a bit small to draw a definite conclusion that profile can be inferred in the general population with our method. With additional resources, this problem can be further addressed by recruiting more participants from the general population, especially seniors and children, and to include more diverse videos. With a larger population, we can also perform multi-class classifications which are more challenging and useful. For comparison, our eye-tracking dataset is the second highest in the number of subjects and the number of video shots among the publicly available ones [33]. The significant amount of resources and expertise which are needed to collect high quality eye-tracking data is a problem which should be overcome for this approach to fully take-off.

Third, one limitation in our current setup for Eye-2-I is the requirement for sufficient labeled eye-tracking data for each video. In some applications, such as our in-flight entertainment system example, this is not a major problem as the library of videos are limited. For other applications with large collection, such as You-Tube, this can be overcome us-

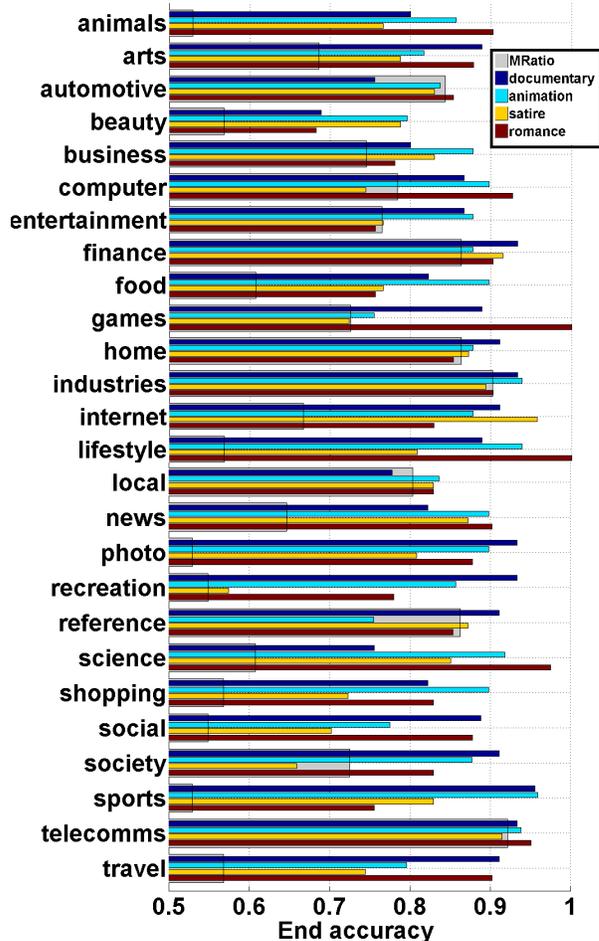


Figure 7: Topics of interests vs end accuracy of each video and *MRatio*.

ing crowd-sourcing. Each video is initialized with a profile of the expected population distributions, e.g. 0.5 male, 0.5 female. Each video’s historical distributions of viewers, if available, can also be used for initialization, e.g. nursery rhymes videos are initialized with higher ratio of young children. After watching a video, a user will be prompted to update his/her inferred profile. A suitable classification algorithm will use the newly labeled eye-tracking data for online learning, after that the labeled data can be safely discarded. Online machine learning is a model of induction that learns one instance at a time. The goal in online learning is to predict labels for instances. The key defining characteristic of on-line learning is that soon after the prediction is made, the true label of the instance is discovered. This information can then be used to refine the prediction hypothesis used by the algorithm. The goal of the algorithm is to make predictions that are close to the true labels. As more labeled data becomes available, the system’s accuracy will improve. We are also exploring methods in cross-media understanding to overcome this limitation [36].

An important scientific question to ask is what are the visual or semantic features which can determine if a visual stimulus is more suitable for classification of an attribute, e.g. *gender*. The answer to this question demands contributions from multiple disciplines such as behavioral psychology, computer science and even neuro-psychology. Our experimental results provide some hints. We observe that videos which involve stronger emotions, e.g. *romance* and *animation* are better than the *documentary* video for profiling. However, psychophysics experiments which isolate these factors for robust analysis are beyond the scope of this paper.

For deployment in an unconstrained environment, there are three factors which can be explored in future work. Firstly, exogenous factors, such as lighting and environmental sounds will affect eye-movements. How can these be managed? Secondly, the non-linear dependencies between attributes, e.g. young male and old female may have high similarities for some eye-gaze features. Is the linear SVM good enough to disentangle these dependencies, or more sophisticated methods such as deep-learning be needed? Third is the effects on repeated viewing of the same or similar video. Is the profiling method stable across multiple viewings?

Clearly there is much room for improvements and gaining new insights about user profiling with eye-tracking data. While our approach showcases the possibility of such an endeavor, we are limited by our resources, knowledge and imaginations. Hence, we humbly and earnestly invite other researchers to explore new possibilities of this unconventional method. To this end, we made our dataset, which took us considerable resources to collect, publicly and freely available.

Despite these limitations and unanswered questions, Eye-2-I may have good potential for radically new designs. There should be some unrealized applications which require a detailed and accurate user profile within minutes, which is not supported by other methods. Self-reporting are intrusive and error-prone. Web-tracking and social media mining need hours and days respectively. Appearance methods such as faces, while fast, are limited to attributes such as gender, age etc. In our experiments, Eye-2-I is able to provide a detailed profile of demographic, personality and

topics of interests, from 539.5 seconds of eye-tracking data while viewing the *animation* video, the mean accuracy of 0.89 can be achieved. Furthermore, while video watching was the chosen context in our user study, we theorize that our method could work with other visual interactions which have temporal ordering, e.g. gaming.

## 9. FUTURE WORK

As described in Section 5, faces also provide implicit and just-in-time information about the users. Together with pupil dilations [7] and video content analysis [35], the affective state of the users can be estimated and a richer set of profiles can be made available. Faces can also be used to enhance Eye-2-I profile on appearance evident attributes, e.g. *gender* and *age*.

We are also investigating more advanced features to improve on the classification results. One possible method is a region-based feature. Barber and Legge reported that people with different interests will fixate in different region of interests in a scene [5]. Such feature is more finely grained to differentiate the amount of attention given by a user in the different ROI of a given scene.

To extend our work such that profiling can be performed without any prior training data from a given data, we will explore the various techniques in transfer learning [23]. One potential way forward is to identify both low-level and semantic features which cause the differences of the eye-gaze patterns.

## 10. CONCLUSION

We proposed and validated the first just-in-time and implicit user profiling method using eye-tracking data. While our experimental setup have several limitations as discussed, we believe the the unique combination of features for our method have potential to support ground-breaking applications. Based on the promising results regarding both the prediction accuracy and response time, we believe just-in-time implicit user profiling is readily achievable in the context of video watching. Given that so much can be known just from one’s eye-gaze, the truth lying in the proverb - *“The eyes are the window of the soul”* - appositely motivate us to explore new territories of human understanding.

## 11. REFERENCES

- [1] Adobe Systems and Edelman Berland. Click here: The state of online advertising. [http://www.adobe.com/aboutadobe/pressroom/pdfs/Adobe\\_State\\_of\\_Online\\_Advertising\\_Study.pdf](http://www.adobe.com/aboutadobe/pressroom/pdfs/Adobe_State_of_Online_Advertising_Study.pdf), 2012.
- [2] B. C. Airplanes. Statistical summary of commercial jet airplane accidents: Worldwide operations 1959-2013. *Aviation Safety, Boeing Commercial Airlines, Seattle, Washington*, 2014.
- [3] F. Alt, A. S. Shirazi, A. Schmidt, and J. Mennenöh. Increasing the user’s attention on the web: using implicit interaction based on gaze behavior to tailor content. In *Proceedings of the 7th Nordic Conference on Human-Computer Interaction: Making Sense Through Design*, pages 544–553. ACM, 2012.
- [4] Arminta Syed. Mirametrix’ eye tracking technology & analytics empowers SMART TVs’ advertisers and content publishers.

- <http://www.mirametrix.com/SmartTVsmarter/>, August 2014. Accessed: 12/08/2014.
- [5] P. J. Barber and D. Legge. *Psychological types*, chapter 4: Information Acquisition. Methuen, London, UK, 1976.
- [6] S. P. . N. R. E. Bodurog glu, A. Cultural differences in visuospatial working memory and attention. *Midwestern Conference on Culture, Language, and Cognition*, 2005.
- [7] M. M. Bradley, L. Miccoli, M. A. Escrig, and P. J. Lang. The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology*, 45(4):602–607, 2008.
- [8] S. Buchala, N. Davey, T. M. Gale, and R. J. Frank. Principal component analysis of gender, ethnicity, age, and identity of face images. *Proc. IEEE ICMI*, 2005.
- [9] H. Chua, J. Boland, and R. Nisbett. Cultural variation in eye movements during scene perception. *Proceedings of the National Academy of Sciences of the United States of America*, 102(35):12629–12633, 2005.
- [10] M. Cristani, A. Vinciarelli, C. Segalin, and A. Perina. Unveiling the multimedia unconscious: Implicit cognitive processes and multimedia content analysis. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 213–222. ACM, 2013.
- [11] C. M. Cristina Conati. Eye-tracking for user modeling in exploratory learning environments: an empirical evaluation. *Knowledge-Based Systems*, 20(6):557–574, 2007.
- [12] T. Gevers. eMotion emotion analyzer. <http://http://visual-recognition.nl/>.
- [13] R. Goldstein, R. Woods, and E. Peli. Where people look when watching movies: Do all viewers look at the same place? *Computers in biology and medicine*, 37(7):957–964, 2007.
- [14] Google. How ads are targeted to your site. <https://support.google.com/adsense/answer/9713?hl=en>. Accessed: 14/03/2014.
- [15] Google. How Google infers interest and demographic categories. [https://support.google.com/adsense/answer/140378?hl=en&ref\\_topic=23402](https://support.google.com/adsense/answer/140378?hl=en&ref_topic=23402). Accessed: 14/03/2014.
- [16] D. W. Hansen and Q. Ji. In the eye of the beholder: A survey of models for eyes and gaze. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(3):478–500, 2010.
- [17] J. B. Hirsh, S. K. Kang, and G. V. Bodenhausen. Personalized persuasion tailoring persuasive appeals to recipients’ personality traits. *Psychological science*, 23(6):578–581, 2012.
- [18] C. G. Jung, H. Baynes, and R. Hull. *Psychological types*. Routledge London, UK, 1991.
- [19] M. Kosinski, D. Stillwell, and T. Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 2013.
- [20] K.-T. Ma, T. Sim, and M. Kankanhalli. VIP: A unifying framework for computational eye-gaze research. In *4th International Workshop on Human Behavior Understanding*. Springer, 2013.
- [21] L. Martens. *Automatic Person and Personality Recognition from Facial Expressions*. PhD thesis, Tilburg University, 2012.
- [22] C. Ono, M. Kurokawa, Y. Motomura, and H. Asoh. A context-aware preference model using a bayesian network for recommendation and promotion. In *User Modeling 2007*, pages 247–257. Springer, 2007.
- [23] S. J. Pan and Q. Yang. A survey on transfer learning. *Knowledge and Data Engineering, IEEE Transactions on*, 22(10):1345–1359, 2010.
- [24] M. Pennacchiotti and A.-M. Popescu. A machine learning approach to twitter user classification. *ICWSM*, 11:281–288, 2011.
- [25] J. W. Pennebaker and C. H. Chew. Behavioral inhibition and electrodermal activity during deception. *Journal of personality and social psychology*, 49(5):1427, 1985.
- [26] L. Qiu, H. Lin, J. Ramsay, and F. Yang. You are what you tweet: Personality expression and perception on twitter. *Journal of Research in Personality*, 46(6):710–718, 2012.
- [27] K. Rayner, C. M. Rotello, A. J. Stewart, J. Keir, and S. A. Duffy. Integrating text and pictorial information: eye movements when looking at print advertisements. *Journal of Experimental Psychology: Applied*, 7(3):219, 2001.
- [28] L. Rosenberg. Gaze-responsive video advertisement display, 2006. US Patent App. 11/465,777.
- [29] N. Sebe and Q. Tian. Personalized multimedia retrieval: the new trend? In *Proceedings of the international workshop on Workshop on multimedia information retrieval*, pages 299–306. ACM, 2007.
- [30] J. Shen and L. Itti. Top-down influences on visual attention during listening are modulated by observer sex. *Vision research*, 65:62–76, 2012.
- [31] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3:42–55, April 2012. Issue 1.
- [32] F. Vigneau, A. F. Caissie, and D. A. Bors. Eye-movement analysis demonstrates strategic influences on intelligence. *Intelligence*, 34(3):261–272, 2006.
- [33] S. Winkler and R. Subramanian. Overview of eye tracking datasets. In *Workshop on Quality of Multimedia Experience*, 2013.
- [34] D. W.-L. Wu, W. F. Bischof, N. C. Anderson, T. Jakobsen, and A. Kingstone. The influence of personality on social attention. *Personality and Individual Differences*, 2013.
- [35] K. Yadati, H. Katti, and M. Kankanhalli. CAVVA: Computational affective video-in-video advertising. *IEEE Transactions on Multimedia*, 16(1), 2014.
- [36] Y. Yang, D. Xu, F. Nie, J. Luo, and Y. Zhuang. Ranking with local regression and global alignment for cross media retrieval. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 175–184. ACM, 2009.