VIP: A unifying framework for computational eye-gaze research

Keng-Teck Ma, Terence Sim, and Mohan Kankanhalli

National University of Singapore,
13 Computing Drive,
Singapore 117417,
{ktma,tsim,mohan}@comp.nus.edu.sg

Abstract. Eye-gaze is an emerging modality in many research areas and applications. We present our VIP framework, which captures the dependence of eye-gaze on Visual stimulus, Intent, and Person. The unifying framework characterizes current eye-gaze computational models. It allows computer scientists to formally define their research problems and to compare with other work. We review the state-of-art computational eye-gaze research and applications with reference to our framework. With the framework, we identify gaps in eye-gaze research and present our work on the new research problem of attribute classification. The accuracy of 0.92 is achieved for classification of Introvert/Extrovert.

Keywords: eye-gaze, framework, visual attention, classification

1 Introduction

"The eyes are the windows to the soul," goes an old English proverb. We agree. In fact, the window acts both ways: as a portal into the person's mind, and as a lens to perceive visual stimuli. In this regard, eye-gaze can provide invaluable clues both to the viewer, and to the object being viewed. This is the exciting premise, and promise, of research using eye-gaze data (see Section 2 for formal definition). Eye-gaze not only permits a fresh approach to existing problems (such as image segmentation), but also throws open a brave new world in which new applications may be created, and new inferences made.

We further advance eye-gaze research by proposing, in this paper, a unifying framework with which to reason about eye-gaze. We review existing models used in current eye-gaze research, and show that, while they are appropriate for their given applications, they are, alas, incomplete. We then propose our VIP eye-gaze framework, which captures the dependence of eye-gaze on Visual stimulus, Intent, and Person. By *visual stimulus* 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 mental states such as purpose of viewing the stimulus, the emotions elicited by the stimulus, etc. Finally, by *person* we mean the persistent attributes of the viewer of the visual stimulus, including identity, gender, age, and personality types. Because we have done careful survey of the field, we believe our VIP framework unifies and subsumes all existing eye-gaze models.

We illustrate the utility of our framework with the novel application of inferring the demographic and personality factors of the viewer. As far as we can tell, this inference of demographic profile and personality type for eye-gaze is pioneering; no one else has done this before. Indeed, the reader will quickly see, that given our VIP framework, many new research opportunities lie just ahead.

2 VIP Framework

In a controlled environment, eye-gaze information of a healthy subject is the automatic and mostly subconscious response of the the viewer's mental processes to the stimulus.

The *visual stimulus* can be an image, a video (with optional audio), a binocular image or an interactive stimulus such as video game. Over the past 25 years, there are extensive and active studies of the properties of the stimulus which affects the eye-gaze [5, 12]. In visual saliency literature, this is known as the bottom-up cues which include colors, brightness contrast and orientations.

The immediate mental processes and conditions have strong and obvious influences [37, 33, 6]. These processes include the top-down influences (knowledge, expectations, reward, and current goals) in the visual saliency literature. Emotions and fatigue are other examples of immediate mental factors. These factors will be coined as *intent*.

Recent psychological research shows that personal attributes of the viewer can affect the eye-gaze. These attributes are stable characteristics of the viewer which persist over months, years or even lifetime. Some persistent attributes are the viewer's identity, gender, age and personality types etc. Goldstein et al. [14] noted that "there are some significant differences in the observation behaviors between gender and age groups" when watching movies. Chua et al. [9] demonstrated that there are cultural differences in eye-movements. Personality has also been discovered as important in gaze modulation [31]. Shen and Itti's work on visual attention shows that the top-down influences are modulated by gender [34]. There are also identification systems which uses eye-gaze information as a biometric [3, 15, 30]. In layman's terms, the "who" and the "type" of the viewer are the persistent attributes. These attributes will be referred to as *person's* attributes.

We then formally define the eye-gaze data, E, as follows:

$$E = g(\{t_i, x_i, y_i, p_i, q_i, d_i, s_i, b_i, c_i\}_L, \{t_i, x_i, y_i, p_i, q_i, d_i, s_i, b_i, c_i\}_R) \qquad (1)$$

where g is a function of a *sequence* of eye-gaze data which

- i: the sequence number, i=1,2,...,n where n is the number of samples.
- t_i: time-stamp of the eye-gaze is related to the sampling rate. Usually the intervals are fixed.

- x_i: horizontal coordinates of the eye-gaze.
- y_i: vertical coordinates of the eye-gaze.
- p_i: horizontal location of the eye in the camera image (video-based eye-tracker only).
- q_i: vertical location of the eye in the camera image (video-based eye-tracker only).
- d_i: distance of the eye to eye-tracker.
- s_i: pupil size of the eye. (diameter or area)
- b_i: eye's opening magnitude. if $b_i = 0$, x_i,y_i and s_i are undefined since the eye is shut.
- c_i: tracking quality. (e.g. 0 = bad, 1 = excellent)
- L/R: left or right eye. Disparity can be used to compute depth or motion.

Together with other auxillary data, such as position of eye-tracker relative to screen/object of interests, 3D position of the eye-gaze can be computed. Examples of g are fixations, saccades and scanpaths vectors. If g is the sequences of fixations, then each fixation u_j is defined as:

$$u_j = \{\bar{x}, \bar{y}, \bar{s}, \bar{b}, t_{start}, t_{end}\} \tag{2}$$

where $\bar{x}$ is the mean value of all x_i in the fixation respectively, t_{start} is the start time of the fixation and t_{end} is the end time of the fixation. Thus g is defined as:

$$g = \{u_j\} \tag{3}$$

Since eye-gaze is influenced by the visual stimulus, intent and person; E can also be defined as a function of the 3 factors:

$$E = f(V, I, P) \tag{4}$$

where V is the visual stimulus' feature vector, I is the immediate mental states feature vector and P is the set of persistent personal attributes. Examples of V are the color and contrast feature vectors. Examples of I are tasks, skill levels or emotion states and emotion intensity. Examples of P are identity and gender.

In the ideal situation, g and f are equivalent. However, due to sensor's noise, computational limitations and incomplete model etc., they are not the exactly the same. In the computational models which we review, the objective of the system is to minimize some application-specific error measure between the ground-truth and the system's results. Hence, "$\approx$" is taken to mean the minimization of the error measure on the both sides of the equation in this paper.

We called this the VIP framework. With this framework as a reference, the features, computational model and assumptions of applications and research problems can be formally described and compared. New research directions are also easier to be discovered by identifying gaps of existing models. We will next survey the current models and applications and how they are completely defined by our framework.

Without loss of generality, consider the special case of which E depends only on V and I. Then either P is a constant or that E is independent of P. If P is a constant c, then we can rewrite $f(V, I, c)$ as $\underset{P=c}{f}(V, I)$. If P is not a constant, then f can be simplified to $f(V, I)$. For both conditions, we will refer to the simplified equations as the VI model. Also, f^{-1} means the inverse dependency of eye-gaze and the VIP factors. It does not necessarily imply that a corresponding f must exist.

2.1 V models

V models assume that without an explicit goal, attention is predominantly dependent on bottom-up cues [11, 17]. In other words, E is independent of P. And I is generally assumed to be the *same* for all viewers. Thus $\underset{I=c}{f}(V)$ defines these models.

It is commonly established that eye-gaze is a reliable proxy for visual attention and the V model is used by current saliency inference algorithms as the ground-truth model. In studies whereby ground truth saliency maps were generated from gaze data as the reference for comparison against computational models, a *single average* ground-truth saliency map was generated for each image [7, 17, 24, 22]. Hence, the reference model is $\underset{I=c}{f^{-1}}(\{E\})$ where $\{E\}$ is the set of the fixations for all human subjects and $\underset{I=c}{f^{-1}}$ is a function, e.g. Gaussian filter [16], which outputs the group-truth saliency map.

The image segmentation problem is another open research problem which has successfully exploited the eye-gaze data for better accuracy [28]. Based on the premise that the human eye invariably fixates within the interior of an object, the algorithm attempts to find the set of boundary contours surrounding the fixation. The segmentation problem can be effectively transformed to an energy minimization problem. By using multiple fixations, its performance is better than the single random fixation method proposed by Mishra et al. [23]. Mishra et al.'s method is in turn better than pure image-based segmentation algorithms [1, 2]. The assumption is that humans *generally* fixate on the most salient objects. The segmentation algorithm $h(f^{-1}(E), V)$ such that $f^{-1}(E)$ localizes the most salient object in the visual stimulus.

The real-time surveillance video summarization system proposed by Vural and Akgul can mix actions from different frames into the same video for more compact videos [35]. A real-time automated algorithm will detect video sections which actions which has occurred. Filtering is performed on the detected video section based on the fixations of the human operator. The $\underset{I=surveillance}{f^{-1}}(E)$ computes the ROI within the video frames V. For this application, I is implicitly assumed to be the mental state of a security personnel at work which is termed *surveillance*. The *task* of general surveillance, e.g. looking for suspicious actions and *domain knowledge* such as familiarity about the monitored environment are expected to be part of *surveillance* state of mind.

Generally, the V applications improve upon image-based algorithms by integrating E into the algorithms. There are many other such problems which benefited from the eye-gaze information [28, 23, 27].

2.2 I models

The *general* I models assume that some I can completely determined the specially selected E, i.e. $E \approx f(I)$. One such example is the activity recognition system by Bulling et al. [8]. They recorded saccades, fixations and blinks using a wearable electrooculography (EOG) device. It can classify 5 activity classes: copying a text, reading a printed paper, taking handwritten notes, watching a video, and browsing the Web. It opens up the wider applicability of eye-gaze data to other activities that are difficult, or even impossible, to detect using common sensing modalities. $f^{-1}(E)$ identifies the activities I from the eye-gaze features E. In this paper, both V (office environment) and P are assumed to be non-informative and are not included in their computational model.

Another example is the "Midas-touch" problem in gaze-based interactions systems. The problem is to infer I from E so that the systems can determine if an eye-gaze is observing or actioning (e.g. issuing a command). Bednarik et al. have used the features extracted from fixations, saccades and pupillary responses to determine the intentions of the user [4]. Their experimental results indicated that fixations and saccades features are more reliable than pupillary responses for predicting intentions.

2.3 P models

As eye movements are counterfeit resistant due to the complex neurological interactions and the extraocular muscle properties involved in their generation, they have been proposed as a viable biometric by various papers [15, 30]. In these papers, the stimulus and the tasks are the same during the training and testing phases. Hence, $E \approx \underset{V=c1,I=c2}{f}(P)$.

Kinnunen et al. [20] implement a stimulus V independent eye-gaze biometric. They identified the histogram of all angles the eye gaze travels during a short period, few seconds, as a potential predictor of a person's identity regardless of V and I. Hence, $f^{-1}(E)$ infers the person's identity P from the histogram E. Their methods are unlike those which use the same task and stimulus for identification.

Zhang et al. [38] uses the features extracted from videos of subjects talking to distinguish identical twins. Out of their 6 features, 3 of them are gaze data: gaze change, pupil movement and eye opening magnitude. In their system, the V varies from the bedroom, recording studios to convention halls filled with people. Therefore, the V is not of consequences to the accuracy. Their P ranges across different age, gender, ethnicity etc. Their system is able to distinguish between identical twins who have many common personal attributes.

To the best of our knowledge, other than biometric applications, there are no application which infers P from E. However, there are many advantages of using eye-gaze to infer other personal attributes as compared to conventional methods such as questionnaires and vision-based approach. Thus, we propose a novel attribute classification problem which is to accurately infer the personal attributes from the eye-gaze. It assumes that given the some stimulus ($c1$), viewers having common intents ($c2$) but differing personal attributes will have different eye-gaze patterns. Hence, $\underset{V=c1,I=c2}{f^{-1}}(E)$ will infer P. We achieve the accuracy of 0.92 with 52 subjects viewing 2 images for introvert/extrovert classification. Further details are presented in Section 3.

2.4 VI models

This model assumes that eye-gaze is dependent on both the visual stimulus features and the immediate mental states of the viewer, e.g. tasks or emotions.

The fixation prediction algorithms which combine both top-down influences and bottom-up cues are examples of applications which uses VI model. The objective of these algorithms is to find some $f(V, I)$ such that the error measure between $f(V, I)$ and E is small. The main directions of research are the methods of combining the features V and I and using new features [12].

In implicit tagging applications, affectiveness of the stimulus is automatically assessed from the viewer's various physiological signals, including pupillary dilation (PD) [25]. The PD is known to be influenced by emotions (I) and light intensity (V). Gao et al. attempts to use Adaptive Interference (AIC), with H* time-varying adaptive (HITV) algorithm to determine the emotions of the viewer [13]. Hence, $I \approx f^{-1}(E, V)$.

Yadati et al. has proposed a novel method for interactive personalized advertisement insertion for a single user [36]. The proposed system fuses in real time, the emotion type (from facial expressions), emotion intensity (from PD) and the affective values (from affective analysis of the video). The most effective advertisements are then inserted accordingly. It has better brand recall rate than the referenced affect-agnostic method. $f^{-1}(E, V)$ infers the I (emotional intensity) from both E (PD) and V (image-based affective analysis of the video segment).

Samsung Galaxy S IV is a smartphone with eye-tracking capability [32]. It can detect whether the user is looking at the screen and adjusts its response according to the displayed task. The "Smart Stay" feature will turn off the screen if the eyes are not detected; the "Smart Pause" feature will pause a playing video if the user looks away. Thus, $I \approx f^{-1}(E, V)$.

2.5 Other cases

The VP and IP models are not sufficiently explored by researchers. The VP model assumes that given some constant I, the eye-gaze are dependent on both the visual stimulus and the personal attributes. One potential research problem is attribute-specific fixation prediction.

The IP model assumes that E is either independent of V or that V is fixed. We do not know of any application or research problem with such assumptions. One potential research problem would be the co-inference of I and P from E that is $(I, P) \approx \underset{V=c}{f^{-1}}(E)$. For example, given some specially selected video and the eye-gaze features, the algorithm can infer gender *and* emotions.

From our extensive literature survey, there is no research problem which is formulated as the most general VIP model. This is clearly a big and interesting gap to be filled. One of the first step is to build a dataset which consists of all 3 factors. Much scientific insights can be gained from a comprehensive dataset which consists of all 3 factors. For example, new discoveries about the relationship and patterns can be found. Co-inference of 2 or even 3 factors may be possible. We have built a VIP dataset which is available at http://mmas.comp.nus.edu.sg/VIP.html.

Table 1 summarizes the features and applications for the various VIP models. From the table, it is clear that V models are well-researched and there are research gaps to be filled in the other models, especially the various combinations of P.

Table 1. Comparison of various VIP models. The underlined application: attribute classification is our contribution in this paper. We also make our VIP dataset which comprises of the 3 factors available at http://mmas.comp.nus.edu.sg/VIP.html. We also suggest open research problems for VP and IP model.

Model	Features	Examples of applications/problems	References
V	color, brightness, contrast, depth region of interests	fixation prediction, bottom-up saliency, image segmentation, image annotations	[5, 21, 7, 17, 24, 22] [28, 23, 27]
I	tasks, fatigue, emotions	activity classification, fatigue detection, emotions classification	[37, 8]
P	identity, demographic, personality	biometric, attribute classification	[3, 15, 30, 20] Section 3
VI	features of V and I	saliency models, video summarization, interactive advertisement	[12, 19, 36]
VP	features of V and P	attribute-specific fixation prediction	Open area
IP	features of I and P	$(I, P) = f^{-1}(E)$	Open area
VIP	features of V, I and P	VIP dataset	Open area

3 Attribute classification

From the VIP framework, we have identified the new problem of personal attribute P classification from eye-gaze information E. We define V to be constant for the training and the inferencing. I is assumed to be free-viewing. Thus, the problem is defined as:

$$P = \underset{V=c,I=free-viewing}{f^{-1}}(E) \tag{5}$$

where P is the persistent personal attribute, e.g. gender. $f^{-1}_{V=c,I=free-viewing}$ is the classifier which was trained on eye-gaze information of other subjects when free-viewing the same stimulus. The information is labeled with their corresponding P. E is the eye-gaze information of the test subject. This problem is a P model since the V and I are constants.

Many of the personal attributes, such as gender, age, culture and personality types are routinely collected by many organizations. These attributes are collectively known as demographic/personality profile. The profiling is used for marketing, personnel screening etc. The advantages of eye-gaze over other modalities are low latency, no purposeful thoughts required and it is non-obtrusive.

Personal attribute classification is analogous to taking a survey. The eye-gaze information in response to an image is similar to taking a survey at a sub-conscious level. Instead of questions, visual stimuli are presented. Similar to the question in a survey, only eye-gaze data of purpose-selected stimulus can accurately determine the value of the intended attribute.

3.1 Experimental Setup

The images were selected from the NUSEF [28] dataset, which contains both neutral and affective images. Out of 758 images, 150 were randomly selected.

72 subjects were recruited from a mixture of undergraduate, postgraduate and working adult population. The male and female subjects are recruited separately to ensure an even distribution. They were tasked to view the 150 images in a free-viewing settings (i.e. without assigned task). Each image was displayed for 5 seconds, followed by 2 seconds viewing of a gray screen. The images were displayed in random order. Their eye-gaze data was recorded with a binocular infra-red based remote eye-tracking device SMI RED 250. The recording was done at 120Hz. The subjects were seated at 50 centimeters distance from a 22 inch LCD monitor with 1680x1050 resolution. This setup is similar to other ones used in eye-gaze research [28].

Before start of the viewing experiment, the subjects also provided their demographic data: gender, age-group, ethnicity, religion, field of study/work, highest education qualifications, income group, expenditure group and nationality. 3 personality type questions are posed based on the Jung's Psychological types [18].

The recorded eye-gaze data were preprocessed with the SMI SDK to extract the fixations from the preferred eye as chosen by the subjects. The recorded data of the 52 subjects, 27 females and 25 males, who have fixations for more than 100 images were used for the attribute classification problems. The number of subjects are comparable to similar studies in eye-gaze experiments [31, 10].

3.2 Features Selection

As this is the first work on using eye-movement data to classify demographic and personality attributes, there is no prior research to directly leverage on. From our preliminary inspections of the fixation data, we select the potential 20 features as follows:

- mean value of the horizontal coordinates, x, of the fixations: $\bar{x}$
- mean value of the vertical coordinates, y of the fixations: $\bar{y}$
- mean value of the fixations' duration: $\bar{d}$
- triangle matrix of covariance of x and y: σ_x, σ_y and σ_{xy}
- standard deviation of duration: σ_d
- standard deviation of pupil size divided by mean value of pupil size: $\sigma_p/\bar{p}$
- 1^{st} fixation: x_1, y_1, d_1
- 2^{nd} fixation: x_2, y_2, d_2
- fixation with the longest duration: x_L, y_L, d_L
- total fixation duration: D
- number of fixations: N

To select the relevant features for classification, a correlation analysis method was applied. The analysis is performed for each image separately. As an example, for the image "dog.jpg', the analysis is applied to $\bar{x}$ of all subjects and the corresponding attribute of the subjects (female=1, male=0). The zeroth lag of the normalized covariance function is used to compute the correlation coefficients and the hypothesis of no correlation. Each p-value is the probability of getting a correlation as large as the observed value by random chance, when the true correlation is zero. The correlation is defined as *significant* if $p - value < 0.05$. This means that the probability of observing a correlation due to statistical fluke is only 5%.

We want to select the features which are highly correlated with the attributes' values and have low $p - value$ for many images. Since each pair (feature,image) has a 0.05 probability of being *significantly* correlated due to random coincidences, the number of expected correlated images for a feature is $0.05 * 150 = 7.5$ for the set of 150 images. Therefore, only features which *significantly* correlates with the attribute for more than 7.5 images are selected. The results are summarized in Table 2.

Hence, the features E selected are:

- *Male/Female*: σ_x, σ_y, $\sigma_p/\bar{p}$
- *Religious/None*: $\bar{x}$, $\sigma_p/\bar{p}$, x_1, x_2, d_2, x_L, D, N
- *Extrovert/Introvert*: σ_{xy}
- *Sensing/Intuition*: $\bar{d}$, σ_d, y_1, y_2, d_2, d_L, D
- *Thinking/Feeling*: $\sigma_p/\bar{p}$, y_2, d_2, x_L, y_L

The correlation analysis results shows that *Male/Female* have different variations of fixations and that their pupillary dilations are different. For religiosity, the 2 groups fixated on different parts of the images ($\bar{x}$, x_1, x_2). The fixation durations also differs. For *Extrovert/Introvert* groups, only the σ_{xy} is found to be significant. For *Sensing/Intuition*, the various fixation durations features, $\bar{d}$, σ_d, d_2, d_L, D, correlates positively with the *Sensing* group. This corresponds well with the characteristic of *Sensing* type who will spend more time to examine a stimuli before making a judgment. In *Thinking/Feeling* groups, the $\sigma_p/\bar{p}$ feature is a good indicator for emotions and it correlates positively with the *Feeling* group. In summary, the correlation analysis are reasonable and consistent with prior knowledge or research results [34].

Table 2. Correlation Analysis. The values in the table shows the number of images which $p-value < 0.05$ for the feature. The features which have less than 7.5 (0.05∗150) images are considered to be statistical coincidences, and are not selected. The features which are selected as underlined.

	$\bar{x}$	$\bar{y}$	$\bar{d}$	σ_x	σ_y	σ_{xy}	σ_d	$\sigma_p/\bar{p}$	x_1	y_1	d_1	x_2	y_2	d_2	x_L	y_L	d_L	D	N
Male/Female	1	1	0	$\underline{18}$	$\underline{8}$	6	0	$\underline{14}$	2	2	0	5	4	4	2	3	0	3	3
Religious/None	$\underline{17}$	2	5	3	6	3	3	$\underline{8}$	$\underline{17}$	1	3	$\underline{9}$	5	$\underline{8}$	$\underline{12}$	4	5	$\underline{68}$	$\underline{46}$
Extrovert/Introvert	2	0	0	3	6	$\underline{8}$	1	3	1	3	3	1	5	2	3	1	1	0	2
Sensing/Intuition	0	7	$\underline{16}$	3	2	4	$\underline{10}$	7	0	$\underline{9}$	7	3	$\underline{8}$	$\underline{13}$	4	3	$\underline{12}$	$\underline{12}$	4
Thinking/Feeling	6	6	3	0	2	0	4	$\underline{20}$	3	7	5	7	$\underline{11}$	9	$\underline{12}$	$\underline{14}$	5	6	7

3.3 Classifier and Training

We used the standard linear SVM classifier in the Matlab Biometric Toolbox, with the default parameters and auto-scaling.

For cross-validation, we applied the leave-one-out method. This method is most suitable as there are insufficient number of subjects per image for k-fold cross-validation, or train-validate-test division.

3.4 Empirical Results and Analysis

Table 3 summarizes the experimental results. Except for gender, the mean accuracies are lower than the prior distribution. This validates our claim that only certain images are useful for certain attribute classification. This also indicates that differences of gaze information of gender are greater and can be more accurately differentiated. The maximum accuracies are higher than the prior distributions for all factors, indicating that it is possible to classify the factors for using those images respectively. Male/Female, Religious/None and Extrovert/Introvert classifiers have many images which have higher accuracy than prior probability. Thus it is easier to select an appropriate image to suit the application requirements for these attributes. The Religious/None attribute has the highest maximum accuracy (0.78) while the Extrovert/Introvert the lowest (0.66). This suggests that religiosity has more influence on eye-gaze information compared to Extrovert/Introvert attribute.

3.5 Classification using eye-gaze from multiple images

We further conducted a set of experiments in which attributes are classified by gaze information of *multiple* images. There are many methods of combining the classifiers from single image classification. We experimented on the voting and tree ensemble methods.

The voting ensembles classifier is implemented as follows. For each subject, the classification results from the single image classifiers vote for the final class. For example, if a subject viewed 5 images and the respective classifiers' results are male, male, female, female, female; then the final classification result is female

Table 3. Accuracy of the classifiers. Prior probability refers to the prior proportion of the majority group. In our dataset, there are 27 females and 25 males subjects, thus prior probability for gender is $27/52 = 0.52$. Images refers to the number of images which classifiers' accuracies are higher than prior probability.

	prior	mean	max	images
Male/Female	0.52	0.54	0.75	94
Religious/None	0.63	0.58	0.78	46
Extrovert/Introvert	0.52	0.51	0.66	80
Sensing/Intuition	0.62	0.52	0.76	26
Thinking/Feeling	0.63	0.54	0.73	24

(3 votes vs 2 votes for male). For this method, the selection of the classifiers is critical to the accuracy rate. The selected single-image classifiers should be also independent for high accuracy. There are 3 selection methods. Using the single best classifier: *single*, using all classifiers: *all* and using the top k classifiers, k is the optimal number of classifiers: *greedy*.

We also use construct an decision tree method where each internal node is a single-image classifier.

The experimental results are shown in Table 4. Clearly, using multiple images outperforms even the best single image classifier. The *all* ensembles have the worst accuracies. This is consistent with our observations that only some images are suitable for attribute classification. The *tree* ensembles are generally good and only a few images are required. Thus it is suitable for applications which the users may not be willing to view too many images.

Table 4. Mean accuracy of the multiple image classifiers. For *greedy* and *tree*, the number in the parentheses indicate the number of classifiers selected. The best accuracies for each factor are underlined.

	single	all	greedy	tree
Male/Female	0.75	0.58	0.87 (3)	0.85 (3)
Religious/None	0.78	0.65	0.88 (8)	0.84 (2)
Extrovert/Introvert	0.66	0.53	0.80 (12)	0.92 (2)
Sensing/Intuition	0.76	0.52	0.80 (3)	0.92 (5)
Thinking/Feeling	0.73	0.54	0.90 (13)	0.90 (4)

4 Conclusion

In conclusion, we proposed a novel and unifying VIP framework which formally defines the eye-gaze computational models. This framework will facilitate the advances of computational eye-gaze research as new problems can be more easily identified. Table 5 shows some examples of eye-gaze applications and their VIP formulations. Secondly, we identified the new research problem: attribute classification. Thirdly, we have built a complete VIP dataset and make it publicly available.

Table 5. Some examples of applications and their corresponding VIP models.

Application	Formulation	f/f^{-1}	E	V	I	P
Saliency [7]	$E \approx \underset{I=c}{f}(V)$	Information Maximization	Gaussian Filter of Fixations	Colors	Free-view	–
Saliency [24]	$E \approx \underset{I=c}{f}(V)$	Conspicuity, Normalization, Summation	Gaussian Filter of Fixations	Colors, Orientation	Free-view	–
Image Segmentation [28]	$V \approx \underset{I=c}{f^{-1}}(E)$	Energy Minimization	Fixations	Most Salient Object	Free-view	–
Video Summarization [35]	$V \approx \underset{I=c}{f^{-1}}(E)$	Energy Minimization	Gaze	Motions	Surveillance	–
Activity Recognition [8]	$I \approx f^{-1}(E)$	mRMR [26], SVM	Saccades, Fixations, Blinks	–	Activity	–
Midas Touch [4]	$I \approx f^{-1}(E)$	Normalization, SVM	Fixation	–	Act/ Observe	–
Biometric [20]	$P \approx f^{-1}(E)$	UBM [29], GMM	Gaze	–	–	Identity
Twins Identification [38]	$P \approx \underset{I=c}{f^{-1}}(E)$	Alignment, GMM, SVM	Gaze, Pupil Movement, Opening Magnitude	–	Talking	Identity
Implicit Tagging [13]	$I \approx f^{-1}(E, V)$	AIC, HITV	Pupillary Dilations	Intensity	Emotions	–
Smart Pause [32]	$I \approx f^{-1}(E, V)$	Proprietary	Gaze	Video/ Other	Pause/ Play/Other	–
Interactive Ads [36]	$I \approx f^{-1}(E, V)$	Fusion	Pupillary Dilations	Affect	Emotions	–
Attribute Classification	$P \approx \underset{V=c1,I=c2}{f^{-1}}(E)$	Correlation, SVM	Fixations	Specific Images	Free-View	Demography, Personality

Acknowledgments. This work is funded by the A*STAR PSF Grant No. 102- 101-0029 on Characterizing and Exploiting Human Visual Attention for Automated Image Understanding and Description.

References

1. Arbeláez, P., Cohen, L.: Constrained image segmentation from hierarchical boundaries. In: CVPR 2008. pp. 1–8. IEEE (2008)
2. Bagon, S., Boiman, O., Irani, M.: What is a good image segment? a unified approach to segment extraction. In: ECCV 2008, pp. 30–44. Springer (2008)
3. Bednarik, R., Kinnunen, T., Mihaila, A., Fränti, P.: Eye-movements as a biometric. Image analysis pp. 16–26 (2005)
4. Bednarik, R., Vrzakova, H., Hradis, M.: What do you want to do next: a novel approach for intent prediction in gaze-based interaction. In: Proceedings of the Symposium on Eye Tracking Research and Applications. pp. 83–90. ACM (2012)
5. Borji, A., Itti, L.: State-of-the-art in visual attention modeling. Pattern Analysis and Machine Intelligence, IEEE Transactions on 35(1), 185–207 (Jan 2013)
6. Bradley, M.M., Miccoli, L., Escrig, M.A., Lang, P.J.: The pupil as a measure of emotional arousal and autonomic activation. Psychophysiology 45(4), 602–607 (2008)
7. Bruce, N., Tsotsos, J.: Saliency based on information maximization. Advances in neural information processing systems 18, 155 (2006)
8. Bulling, A., Ward, J., Gellersen, H., Troster, G.: Eye movement analysis for activity recognition using electrooculography. Pattern Analysis and Machine Intelligence 33(4), 741–753 (2011)
9. Chua, H., Boland, J., Nisbett, R.: 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. Dorr, M., Martinetz, T., Gegenfurtner, K., Barth, E.: Variability of eye movements when viewing dynamic natural scenes. Journal of vision 10(10) (2010)
11. Elazary, L., Itti, L.: Interesting objects are visually salient. Journal of Vision 8(3) (2008)
12. Frintrop, S., Rome, E., Christensen, H.I.: Computational visual attention systems and their cognitive foundations: A survey. ACM Transactions on Applied Perception (TAP) 7(1), 6 (2010)
13. Gao, Y., Barreto, A., Adjouadi, M.: Monitoring and processing of the pupil diameter signal for affective assessment of a computer user. In: Human-Computer Interaction. New Trends, pp. 49–58. Springer (2009)
14. Goldstein, R., Woods, R., Peli, E.: Where people look when watching movies: Do all viewers look at the same place? Computers in biology and medicine 37(7), 957–964 (2007)
15. Holland, C., Komogortsev, O.V.: Biometric identification via eye movement scanpaths in reading. In: Biometrics (IJCB), 2011 International Joint Conference on. pp. 1–8. IEEE (2011)
16. Judd, T., Durand, F., Torralba, A.: A benchmark of computational models of saliency to predict human fixations. Tech. rep., MIT (Jan 2012)
17. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: IEEE International Conference on Computer Vision (ICCV) (2009)
18. Jung, C.G., Baynes, H., Hull, R.: Psychological types. Routledge London, UK (1991)
19. Katti, H., Yadati, K., Kankanhalli, M., Chua, T.S.: Affective video summarization and story board generation using pupillary dilation and eye gaze. In: Multimedia (ISM), 2011 IEEE International Symposium on. pp. 319–326. IEEE (2011)
20. Kinnunen, T., Sedlak, F., Bednarik, R.: Towards task-independent person authentication using eye movement signals. In: Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications. pp. 187–190. ACM (2010)

21. Lang, C., Nguyen, T.V., Katti, H., Yadati, K., Kankanhalli, M., Yan, S.: Depth matters: influence of depth cues on visual saliency. In: Computer Vision–ECCV 2012, pp. 101–115. Springer (2012)
22. Le Meur, O., Le Callet, P., Barba, D., Thoreau, D.: A coherent computational approach to model bottom-up visual attention. Pattern Analysis and Machine Intelligence 28(5), 802–817 (2006)
23. Mishra, A., Aloimonos, Y., Cheong, F.L.: Active segmentation with fixation. In: Computer Vision, 2009 IEEE 12th International Conference on. pp. 468–475. IEEE (2009)
24. Ouerhani, N., Von Wartburg, R., Hugli, H., Muri, R.: Empirical validation of the saliency-based model of visual attention. Electronic letters on computer vision and image analysis 3(1), 13–24 (2004)
25. Pantic, M., Vinciarelli, A.: Implicit human-centered tagging [social sciences]. Signal Processing Magazine, IEEE 26(6), 173–180 (2009)
26. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. Pattern Analysis and Machine Intelligence, IEEE Transactions on 27(8), 1226–1238 (2005)
27. Ramanathan, S., Katti, H., Huang, R., Chua, T.S., Kankanhalli, M.: Automated localization of affective objects and actions in images via caption text-cum-eye gaze analysis. In: Proceedings of the 17th ACM international conference on Multimedia. pp. 729–732. ACM (2009)
28. Ramanathan, S., Katti, H., Sebe, N., Kankanhalli, M., Chua, T.S.: An eye fixation database for saliency detection in images. In: ECCV 2010. Crete, Greece (2010)
29. Reynolds, D.A., Quatieri, T.F., Dunn, R.B.: Speaker verification using adapted gaussian mixture models. Digital signal processing 10(1), 19–41 (2000)
30. Rigas, I., Economou, G., Fotopoulos, S.: Human eye movements as a trait for biometrical identification. In: Biometrics: Theory, Applications and Systems (BTAS), 2012 IEEE Fifth International Conference on. pp. 217–222. IEEE (2012)
31. Risko, E.F., Anderson, N.C., Lanthier, S., Kingstone, A.: Curious eyes: Individual differences in personality predict eye movement behavior in scene-viewing. Cognition (2011)
32. Samsung Galaxy S4 - Life Task. `http://www.samsung.com/global/microsite/galaxys4/lifetask.html#page=pausescroll`, accessed: 02/04/2013
33. Schleicher, R., Galley, N., Briest, S., Galley, L.: Blinks and saccades as indicators of fatigue in sleepiness warnings: looking tired? Ergonomics 51(7), 982–1010 (2008)
34. Shen, J., Itti, L.: Top-down influences on visual attention during listening are modulated by observer sex. Vision research 65, 62–76 (2012)
35. Vural, U., Akgul, Y.S.: Eye-gaze based real-time surveillance video synopsis. Pattern Recognition Letters 30(12), 1151–1159 (2009)
36. Yadati, K., Katti, H., Kankanhalli, M.: Interactive video advertising: A multimodal affective approach. In: Advances in Multimedia Modeling, pp. 106–117. Springer (2013)
37. Yarbus, A., Haigh, B., Rigss, L.: Eye movements and vision, vol. 2. Plenum press New York (1967)
38. Zhang, L., Nejati, H., Foo, L., Ma, K.T., Guo, D., Sim, T.: A talking profile to distinguish identical twins. In: Proceedings of the 10th international conference on Automatic Face and Gesture Recognition. IEEE (2013)