Detecting High-functioning Autism in Adults Using Eye Tracking and Machine Learning

Victoria Yaneva *, Le An Ha *, Sukru Eraslan †, Yeliz Yesilada †, and Ruslan Mitkov *

***Abstract*—The purpose of this study is to test whether visual processing differences between adults with and without high-functioning autism captured through eye tracking can be used to detect autism. We record the eye movements of adult participants with and without autism while they look for information within web pages. We then use the recorded eye-tracking data to train machine learning classifiers to detect the condition. The data was collected as part of two separate studies involving a total of 71 unique participants (31 with autism and 40 control), which enabled the evaluation of the approach on two separate groups of participants, using different stimuli and tasks. We explore the effects of a number of gaze-based and other variables, showing that autism can be detected automatically with around 74% accuracy. These results confirm that eye-tracking data can be used for the automatic detection of high-functioning autism in adults and that visual processing differences between the two groups exist when processing web pages.**

***Index Terms*—Autism, Eye Tracking, Web, Screening, Diagnostic Classification, Detection.**

Open Data

Eye-tracking data, code, and materials are available in our external repository at https://tinyurl.com/detectingautism.

I. Introduction

Many disorders and diseases that do not have a clinical biomarker are at risk of being either misdiagnosed or diagnosed during their later stages. One such neurodevelopmental disorder is Autism Spectrum Disorder (ASD), which affects communication and social interaction [1]. As autism is a highly heterogeneous condition, the term "spectrum" is used to signify the different types and levels of support that different individuals might need, where "high-functioning autism" signifies a high level of independence and ability. While individuals with high-functioning autism have a normal IQ range, they may process information differently, especially in situations requiring social interaction, the understanding of semantics and pragmatics, or the transfer of knowledge from one domain to another. Many people on the spectrum may also have atypical sensory processing and attention shifting patterns (Section I-A,) as well as a preference for specific routines [2].

Currently, the ASD diagnostic procedure is a highly subjective assessment process. It is restricted to behavioural, historical, and parent-report information [3], [4], which is then interpreted by a qualified clinician. The case-by-case basis of the decision is necessary as it allows to treat each patient according to their circumstances but, at the same time, it leads to a lack of consistency and reliability [5], [6].

Obtaining an early diagnosis is more likely achieved when ASD symptoms are severe [5], and, conversely, people with high-functioning autism seeking a diagnosis in their adulthood are especially difficult to diagnose [7]. Some of the reasons are that the symptoms of high-functioning autism are not as obvious; developing coping strategies throughout life (e.g., learning to avoid triggers) masks the presentation of relevant symptoms; and that, unlike for children, critical incidents with adults are not monitored by school staff or parents. It would therefore be beneficial to develop a screening method for identifying high-functioning autism that does not rely on parental and school reports and that is sensitive enough to capture the fine-grained differences between adults who are on the spectrum and those who are not.

In this paper, we test the hypothesis that *visual processing differences between adults with and without high-functioning autism captured through eye tracking can be used to detect autism automatically*. This approach is based on the idea that the eye-tracking data captures differences in the cognitive profiles of the two groups when executing information-searching tasks, and then these differences, as learned by a machine-learning classifier, can be used as a marker of the condition.

A. Autism Detection

The most rigorously validated autism-detection models for adults which use behavioural data are based on resting-state fMRI, owing to the availability of data sets collected in different centres and used as unseen data for evaluation. The accuracy of these classifiers varies between 79% [8] and 86% [3] for leave-one-out cross-validation (LOOCV) and between 71% [8] and 83% [9] when tested on unseen data. Another study using only LOOCV reports 76.7% [10]. The best result of 86% (subsequently 80% when evaluated on unseen data) is based on training data from 12 participants with ASD and 12 Control participants and achieves 100% sensitivity (recall) and 66.7% specificity (precision). While these studies provide a promising direction in autism detection, collecting the fMRI data is a very expensive and obtrusive procedure and is not suitable for pregnant women, nor for people with sensory issues, metal implants, claustrophobia, head trauma, etc., which limits the applicability of the approach. Nevertheless, to the best of our knowledge, these results represent the state of the art in automatic autism detection with behavioural data.

*The Research institute for Information and Language Processing, University of Wolverhampton, (v.yaneva | ha.l.a | r.mitkov@wlv.ac.uk)

†Computer Engineering Program, Middle East Technical University Northern Cyprus Campus, (seraslan | yyeliz@metu.edu.tr)

Studies using EEG and speech data report results from 10-fold cross validation instead of LOOCV, where the accuracy is 94% for EEG data [11] and 93% for speech data [12]. These are potentially overoptimistic as different data segments from the same participant are assigned to the training and testing sets, thus increasing the similarity between the two. In other words, this evaluation set-up does not correspond to real-world applications where the system has to categorize a user, portions of whose data are not included in the training set.

The differences in visual attention between people with autism and neurotypical people are well documented in the literature (e.g., [13]–[18]). Atypical visual-attention patterns reflect higher-order differences in information processing, as the focus of attention directs the input of information from the environment. Visual attention is related to concentration, interest, perception, learning, the ability to form joint attention, cognitive effort and other indicators, the combination of which can be used to detect autism. For example, many people with ASD tend to avoid the eye region when looking at faces [13], [16] and this phenomenon has been extensively investigated in relation to social interaction difficulties, which are one of the diagnostic criteria for ASD. Furthermore, eye-tracking data from visual attention tasks has been shown to correlate well with brain activity differences. Evidence from a large sample of 294 ASD subjects suggests that the differences in the visual attention to the eye region could be due to smaller volume of the right anterior cerebellum in subjects with ASD and that *"eye tracking may be a promising neuro-anatomically based stratifying biomarker of ASD"* [13].

Eye tracking has mainly been explored as a biomarker of autism in infants, toddlers, and young children [14]–[17]. A large contribution of these studies relates to the challenges of accurately recording gaze in such young subjects and they establish eye tracking as a promising direction for autism detection. With regards to young children (between the ages of four and six), [16] reports a study involving ASD and typically-developing children watching a silent video of a woman mouthing the English alphabet. The reported accuracy is 85.1%. This was achieved using five-fold cross-validation with no note on how the data was split (i.e. whether segments of the data of the same participant were used for training and testing, as in the EEG and speech studies above). In another study, [17] perform classification of children between the ages of four and 11 looking at pictures of faces. Using LOOCV, the reported accuracy is 88.51%. It was unclear, however, how severe the autism of the children from the ASD group was. The study presented here differs from previous research in that it involves adults as opposed to children, and it focuses on high-functioning autism, which can be challenging to detect.

B. Overview of the Proposed Approach

Unlike previous research, the focus of this paper is on testing whether gaze data can be used for fine-grained distinction between adults with and without high-functioning autism and whether data from familiar activities such as web-page processing can be useful for this purpose. The motivation behind this approach is first and foremost related to the fact that eye-movement data does not rely on subjective interpretation of whether or not a given behaviour occurred, and that: i) looking for information on web pages is a highly familiar and naturalistic task; ii) gaze can be reliably captured through eye tracking; and iii) eye-tracking data is relatively less costly and easier to record, as devices and software that use gaze for navigation or gaming have been available on the market for years (e.g., Samsung Galaxy S4, Tobii Eye Trackers for PC Gaming, etc.). It is important to note that the proposed screening method is not intended to substitute clinical diagnostic procedures where these are available. Rather, its purpose is to be used to identify autistic traits at a wider level and to provide the means to refer for further assessment those people who might be at risk.

The experiments presented in this paper test our main hypothesis using data derived from different participant groups, stimuli and tasks in two independent data collection rounds involving a total of 71 unique adult participants (31 with high-functioning ASD and 40 Control). In both rounds of data collection, the two groups of participants completed information-processing tasks while looking at web pages and having their eye movements recorded by an eye tracker. Using data from the first round only, we trained an initial autism-detection tool which achieved an accuracy of 75% and was presented in [19]. Encouraged by this result, we proceeded to collect more data and investigate new research questions using both data sets. As a result, the present manuscript has the following distinct objectives, which have not been previously explored:

1) To test whether classifiers based on eye-tracking data from visual processing tasks can be used to detect high-functioning autism in adults with *consistent* accuracy when trained on new data using a different stimulus set, as well as different participants and tasks. The consistency of the prediction across various conditions is informative of the reliability of the approach.
2) To test whether people with autism exhibit different visual processing patterns with and without specific information-location tasks across different time conditions. The implications of this question are related to distinguishing whether the atypical attention patterns are directly influenced by having a specific information-location task or exist independently as a natural approach to visual processing in the absence of an explicit instruction. In addition to task effects, here we are also concerned with the effects of task duration.
3) To investigate the effects of a number of factors over the classification accuracy, which were not previously investigated. These include a number of gaze-based and page-related features, a larger number of tasks, particularly in terms of eliciting attention-shifting differences, as well as different approaches and granularity levels for defining the Areas of Interest (AOIs) on the web pages.
4) To aid independent research by making new data available.

In the next sections, we refer to experiments conducted with the two rounds of data collection as Study 1 and Study 2, respectively. For the purpose of direct comparison and joint analysis, we also present key results from the study reported in [19], which are clearly marked in the relevant tables.

TABLE I
INFORMATION ABOUT THE PARTICIPANTS

	Study 1		Study 2	
	ASD	**Control**	**ASD**	**Control**
Participants	18	18	19	25
Excluded	3	3	0	6
Included	15	15	19	19
Gender (F, M)	(6, 9)	(7, 8)	(8, 11)	(13, 6)
Age (m, SD)	37 (9.14)	33.6 (8.6)	41 (14)	32.2 (9.9)
School (years) (m, SD)	16 (3.3)	18.4 (2.5)	15.8 (2.4)	17.4 (2.9)

II. METHOD

Prior to commencing the data collection, ethical approval was sought and granted by the University of Wolverhampton Ethics Committee (Faculty of Science and Engineering).

A. Participants

Both Study 1 and Study 2 involved a group of participants with autism and a control group of participants without ASD. Data was collected from a total of 71 unique participants (31 ASD and 40 Control) and retained for a total of 68 unique participants (28 ASD and 32 Control). Six participants with ASD and three control-group participants took part in both experiments. The initial data collected for Study 1 comprised 18 ASD-group participants and 18 control-group participants, whereas Study 2 included 19 ASD-group participants and 25 control-group participants. Details about demographic characteristics are presented in Table I. All participants reported that they used the web daily, except for one in Study 2 who reported that she used the web less than once a month.

Recruitment: All ASD participants were recruited through a UK charity organisation and the Enabling Centre at the University of Wolverhampton. For both institutions, they had to provide a copy of their formal diagnosis to access the provided services. The control-group participants were recruited through snowball sampling from the area of Birmingham, UK.

Inclusion and Exclusion Criteria: The inclusion criteria for the ASD group was a formal diagnosis of autism and all participants met the ADOS diagnostic criteria [20]. Some participants were diagnosed before the introduction of DSM-5 in 2013, so other acceptable diagnoses were "High-Functioning Autism" or "Asperger's syndrome". All participants had to be over 18 years of age and able to use a computer. The exclusion criteria were a formal diagnosis of *any degree of intellectual disability* or a *reading disorder*, as well as *conditions affecting vision* that could not be corrected using glasses or lenses. For the control group, the inclusion and exclusion criteria were similar, except for having a diagnosis of autism. To ensure that no participants with a high incidence of autistic traits were included in the control group, all control participants completed the 50-item Autism Quotient (AQ) test [21]. This test is widely used as a screening tool for autism by general practitioners before providing a referral for expert diagnosis.

Excluded participants: In Study 1, data from five participants (three with ASD and two control) was excluded because of calibration issues and inaccuracies due to head movements. Subsequently, data from one randomly selected control participant was also excluded to have balanced classes. For Study 2, five control-group participants were excluded (one for having a high score at the AQ test, three for reporting school referrals for dyslexia diagnosis, and two for calibration issues).

Level of Independence: All participants were highly independent adults, none of whom relied on a caregiver. Twenty participants with ASD were either employed or enrolled in a higher education degree, while eight lived independently in council housing and received disability benefits. The control-group participants were either employed or in education.

Education: Table I provides details on the number of years spent in education. In Study 2, we also asked about the highest level of completed degree. From the ASD group, eight people had completed high-school (of them, two were enrolled in an undergraduate degree, three were in employment and three relied on benefits), eight had obtained Bachelor's degrees and three had Master's degrees. Of the control group, three people had completed high school (and were enrolled in an undergraduate degree), 10 people had Bachelor's degrees, five people had Master's degrees and one person had a Ph.D.

B. Materials

Web pages are hypothesized to be a particularly suitable stimulus set for several reasons. They offer a variety of both textual and visual stimuli organized for a specific semantic purpose, which allows for a number of visual searching strategies to be employed. Exploring web pages is also a highly familiar and naturalistic task for many people, which is important, as using data from everyday tasks has the potential to unveil differences that could be captured by means which are independent of laboratory equipment. Adult people with autism are also known to focus for longer on images in text-and-image pairs compared to controls [22]. For web pages, the scanpaths of viewers with autism have been shown to exhibit higher variance [23]. These findings lead us to expect that web pages are suitable stimuli for investigating the use of visual processing differences for autism detection.

The materials used in Study 1 were six web pages, two of which had low visual complexity (Apple and Babylon), two had medium complexity (AVG and Yahoo) and two had high complexity (Godaddy and BBC). The details for the selection of materials in Study 1 are presented in [19]. The screenshots of these pages are available in our open repository.

To select the web pages for Study 2 we followed a similar procedure to Study 1, where we analysed the top 100 websites listed by Alexa.com. We first removed any duplicates and pages that were not in English , as well as those designed as search engines and those that required authentication. We then computed the visual complexity scores for the home pages of the remaining websites by using the ViCRAM tool [24] and randomly selected eight websites. We ensured that the home pages of four of them had low visual complexity (WordPress, WhatsApp, Outlook, Netflix), while the home pages of the other four websites had high visual complexity (YouTube, Amazon, Adobe, BBC). The screenshots of these pages are also available in the repository.

C. Tasks

Both Study 1 and Study 2 included two different kinds of tasks, the full list of which can be found in our open repository and Appendix. These tasks were developed based on the "hierarchy of information needs" as defined by [25], which differentiates between basic acquisition of simple facts (e.g., browsing a web page), the ability to look up the answer to a question on the page (e.g., searching for the relevant element), and the ability to combine information from multiple facts in order to arrive at a new piece of information (referred to as "synthesis"). Study 1 included browse and search tasks whereas Study 2 included browse and synthesis tasks. In each study, the two tasks were executed in a counterbalanced order for each participant. All questions were posed verbally, and answering the questions of search and synthesis tasks required the participant to either locate the correct element or say what the answer was, respectively. No interaction with the web pages was involved such as scrolling, clicking, or typing.

Study 1: There were two different kinds of tasks:

- *Browse task:* Participants were instructed to explore the web pages and were free to focus for as long or as little on any given element. The time limit for this task was 120 seconds per page but the participants could move on to the next page when they had finished browsing the current one. This task allowed for capturing differences in visual processing of page elements without the interference of a specific task.
- *Search task:* Participants were required to locate a specific element on the web page in order to answer a question. An example is the question: *"Can you find a telephone number for technical support and read it?"*. This was the most intensive task, as the time limit was 30 seconds for two questions per page.

Study 2: As will be seen, the Browse task from Study 1 gave promising classification results but it was not possible to find out whether this was because of a different approach to browsing in the two groups, or because the ASD group spent longer on the task and had more fixation points (reasons for the longer times could have been related to higher level of conscientiousness or following instructions more strictly). To clarify this, in Study 2 we capped the time limit for the Browse task to 30 seconds per page, which meant that each participant had spent the same time browsing each page. It is important to note that this is not a direct comparison of two time limits for the same task, as we also had a different set of participants and web pages. Instead, the aim was to find out whether a discrimination signal would occur even if the possibility for longer times between participants were removed. With regards to the Search task, we hypothesized that not setting such a short time limit (30 sec) and increasing the level of difficulty could give better results for the classification of the two groups of participants (the time-limited condition had already provided conclusive results about differences in the searching strategies). In Study 2 we allowed longer times for the equivalent of the Search task (120 seconds per page) and added extra difficulty to it, as explained below.

- *Browse task 2:* The Browse task in Study 2 was the same as in Study 1, with the sole difference that time was limited to 30 seconds per page. Each page would change after the 30 seconds had passed.
- *Synthesis task:* Similar to the Search task in Study 1, the participants were required to answer questions about the information provided on the web page. However, these questions required the participants to locate at least two elements on the page and compare them, to arrive at the third piece of information, which was implicit (the correct answer). An example for the Netflix page is *"How much more would you have to pay compared to the basic plan if you wanted to have ultra HD?"*. To answer this question, the participant had to locate the price of the basic plan, the price of the ultra-HD plan and compare the two. Having the added difficulty of inferential thinking was a potentially plausible way to enhance the signal, as it better represents the different cognitive profiles of the two groups. The time limit per web page was 120 seconds, but the participants were free to move on earlier if ready.

D. Apparatus

Both studies used a 60Hz Gazepoint GP3 video-based eye tracker with accuracy of 0.5-1 degree of visual angle. The screenshots were presented on 19" and 17" LCD monitors for Studies 1 and 2, respectively. The distance between each participant and the eye tracker was controlled by using a system-integrated sensor, and was roughly 65 cm. The device recorded data from the right eye. Fixations were extracted using Gazepoint's built-in algorithm based on the position variance technique, where a sequence of gaze data estimates spatially located within a local region are determined to belong to the current fixation, while subsequent data outside of this local region is identified as beginning a new fixation [26].

E. Procedure

All recordings were completed in a quiet room with only the researcher and the participant present. After getting familiar with the purpose and procedure of the experiment, all participants signed their consent forms, filled in the demographic questionnaires and the Autism Quotient test (control participants only), and then performed a nine-point calibration of the device. After a successful calibration, each participant was presented with the web pages in a randomized order, to control for fatigue and memory effects. The two tasks per study were presented in a counterbalanced order for each participant. All questions and answers were presented verbally after the relevant stimulus was presented on the screen.After the completion of the experiment, the participants were debriefed.

III. Classification Experiments

A. Defining the Areas of Interest

In this paper, we extend the exploration of the effects of AOI granularity level on classification performance by adding more extraction configurations than previous work. This allows us to investigate the trade-off between capturing detailed page-specific information on one hand, and not introducing extra

noise through too fine a segmentation on the other. We define the AOIs using two different approaches: (1) generic and (2) page-specific (see Figure 1). Both approaches are systematic and can be replicated with other web pages.

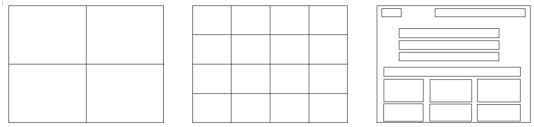

Fig. 1. Types of Areas of Interest (from left to right): 2 x 2 generic grid, 4 x 4 generic grid and page-specific AOIs

Page-specific AOIs: Page-specific AOIs correspond to visual elements on web pages. For identifying page-specific AOIs, we used the extended version of the Vision-Based Page Segmentation (VIPS) algorithm [27]. This algorithm divides web pages into their elements by using both their source code and visual representation. As a result, it arranges the elements in a hierarchical form, where the deeper levels contain more and smaller elements. Since a user study conducted by [27] shows that the fifth level is the most preferred level, we used the elements from that level. However, if a given task could be completed by fixating a single element, then we used the deeper levels. An example is presented in Figure 1. There were a total of 112 page-specific AOIs identified in the first study and a total of 213 page-specific AOIs in the second study.

Generic AOIs: The generic AOIs consisted of simple square grids. We previously experimented with a 2 x 2 grid but now we also added a 4 x 4 grid for the data collected in both Studies 1 and 2. Unlike the page-specific AOIs, this grid-based segmentation ensures that none of the fixations are lost.

B. Features

We experiment with gaze-based and other features, as presented in Table II. We investigate 12 different cases: two eye-tracking studies each with two types of tasks (Browse and Search for Study 1 and Browse and Synthesis for Study 2) and three kinds of AOIs for each task (Page-specific, 2 x 2 Generic, and 4 x 4 Generic). Therefore, we create a different table (i.e. dataset) for each case where each row represents the values of the features for each AOI on each page for each participant.

C. Experimental Setting

Several experiments were performed, taking into account the different tasks and AOI identification approaches. A set of standard classification algorithms were initially tested by using R, where the logistic regression algorithm performed best[1]. The evaluation of performance was based on training and testing sets for 100 random splits of the data[2]. In Study 1, we report results using data from five randomly-selected participants per group for testing and data from the remaining ten participants per group for training. In Study 2, we report results from six random participants per group for testing and the remaining 13 participants per group for training. In both studies the best results were achieved when splitting the training and testing participants in a 70% : 30% ratio for each fold. This randomization was performed for each training and testing iteration. The model was first trained to predict which group a behaviour belongs to. In this context, a behaviour is considered a vector of feature values corresponding to a given AOI for a given participant and web page. After that, it compares the number of times a behaviour from that participant is classified as belonging to each group (ASD or Control) and considers the group with the higher number as the predicted group. Since we performed evaluation for 100 random splits, the accuracy for each split was then averaged and reported together with the 95% confidence intervals (CI).

[1]See Appendix or repository for result comparison to other classifiers

[2]We experimented with different training and testing sizes and report the best performing classifiers; however, the full results for each different data split and each page configuration are available in our repository.

TABLE II
THE LIST OF THE FEATURES ALONG WITH THEIR EXPLANATION

	Feature	Explanation
Gaze-Based	Time to First View	The time in seconds when an AOI was first fixated measured from the moment the web page was presented on screen
	Time Viewed(sec)	The sum of the total durations of all fixations per participant per AOI measured in seconds
	Time Viewed %	The sum of the total durations of all fixations per participant per AOI as a percentage of the total time spent looking at the web page for the specific task
	Fixations	The number of gaze fixations for each participant in a given AOI
	Revisits	The number of times a participant went back to a previously viewed AOI
Non-Gaze	Media ID	The identification number of each web page
	AOI ID	The identification number of each area of interest
	Correct Answer AOI	The area of interest which contains the correct answer to one of the search tasks
	Gender	Previous research has shown that gender may have an impact on the way people look for information on the web [28]
	Visual Complexity	Three levels of visual complexity: low, medium (N/A for the second study) and high.

IV. RESULTS

We trained several logistic-regression classifiers with the default parameters in R using combinations of feature sets. We used all gaze features together with non-gaze ones for all pages and selected combinations of pages. The main results are presented in Table III, with additional results for non-gaze variables that did not provide a clear pattern (namely *Gender* and *Visual Complexity)* presented in the Appendix/repository. The results for individual pages are presented in Table IV for the Browse 2 and Synthesis tasks (the results for individual pages for the Study 1 are presented in the Appendix).

A. Browse tasks

Both the 30-second condition and the 120-second condition provide a discrimination signal, indicating that, indeed, the two groups have different browsing strategies. Best results for the two conditions are achieved when using only the most discriminatory web pages (Apple and AVG for Browse 1 and Outlook for Browse 2; confusion matrix presented in repository). The condition with a longer time limit however, provides better results (i.e. 0.74 compared to 0.65), most

TABLE III
EVALUATION RESULTS FOR ALL TASKS (95% CONFIDENCE INTERVALS GIVEN IN PARENTHESIS)

	Browse, Study 1 (Yaneva et al., 2018)		**Browse, Study 2**		**Search, Study 1 (Yaneva et al., 2018)**		**Synthesis, Study 2**	
Feature set	**All**	**Best**	**All**	**Best**	**All**	**Best**	**All**	**Best**
Page-spec. AOIs	.66 (.64-.68)	.71 (.69-.73) (AP, AVG)	.50 (.48-.51)	**.65 (.63-.67) (OL)**	.70 (.68-.72)	**.75 (.73-.78) (AP, BL)**	.62 (.60-.64)	.69 (.67-.71) (AB, AZ, OL)
2x2 Grid	.60 (.58-.63)	.66 (.63-.68) (AP, AVG)	.50 (.48-.53)	.60 (.57-.63) (OL, WA, YT)	.52 (.50-.54)	.63 (.61-.66) (BL, BBC)	.60 (.58-.62)	.67 (.65-.69) (AB, AZ, OL, WP)
4x4 Grid	.64 (.62-.67)	.71 (.69-.74) (AP, BL, BBC)	.49 (.47-.51)	.61 (.59-.62) (OL, YT	.67 (.65-.69)	.73 (.71-.75) (AP, GD, YH)	.62 (.60-.64)	**.73 (.72-.75) (AZ, OL)**
Page-spec. AOIs + AOI ID	.53 (.51-.54)	.53 (.51-.54) (All)	.50 (.49-.50)	.51 (.50-.53) (BBC, WA, YT)	NA	NA	NA	NA
2x2 Grid + AOI ID	.61 (.59-.63)	.67 (.65-.69) (AP, BL, AVG)	.47 (.45-.49)	.50 (.50-.51) (AB, AZ, WA, WP)	.49 (.47-.52)	.61 (.58-.63) (AP, BL)	.60 (.58-.63)	.60 (.58-.63) (All)
4x4 Grid + AOI ID	.62 (.59-.64)	.62 (.59-.64) (All)	.53 (.50-.55)	.53 (.50-.55) (All)	.56 (.54-.58)	.56 (.54-.58) (All)	.60 (.58-.62)	.60 (.58-.62) (All)
Page-spec. AOIs + Media ID	.66 (.63-.68)	**.74 (.71-.76) (AP, BL)**	.49 (.47-.51)	.62 (.59-.64) (OL, WA)	.68 (.66-.70)	.76 (.73-.78) (AP, BL, AVG)	.60 (.58-.62)	.68 (.66-.70) (AB, AZ, BBC, OL, WP)
2x2 Grid + Media ID	.62 (.59-.64)	.67 (.65-.70) (AP, BL)	.47 (.45-.49)	.58 (.55-.60) (OL, WA, YT)	.49 (.47-.52)	.62 (.59-.64) (BL, BBC)	.62 (.60-.65)	.67 (.65-.70) (AB, AZ, BBC, OL, WA, WP)
4x4 Grid + Media ID	.62 (.60-.65)	.69 (.67-.71) (AP, AVG)	.51 (.49-.53)	.60 (.57-.62) (AZ, OL, WA, YT)	.66 (.64-.69)	.72 (.70-.74) (AP, GD, YH)	.65 (.63-.67)	.72 (.70-.74) (AZ, OL)

Study 1–Apple: AP, Babylon: BL, AVG: AVG, Yahoo: YH, Godaddy: GD, BBC: BBC
Study 2–Adobe: AB, Amazon: AZ, BBC: BBC, Netflix: NF, Outlook: OL, WhatsApp: WA, WordPress: WP, YouTube: YT

TABLE IV
EVALUATION RESULTS FOR INDIVIDUAL WEB PAGES FOR THE BROWSE (STUDY 2) AND SYNTHESIS TASKS WITH PAGE-SPECIFIC AOIS (95% CONFIDENCE INTERVALS GIVEN IN PARENTHESIS)

Task	**Features**	**Adobe (H)**	**Amazon (H)**	**BBC (H)**	**Netflix (L)**	**Outlook (L)**	**WhatsApp (L)**	**WordPress (L)**	**YouTube (H)**
Browse (2)	Gaze	0.45 (.43-.48)	0.43 (.41-.45)	0.5 (.48-.52)	0.47 (.44-.49)	0.65 (.63-.67)	0.49 (.47-.51)	0.47 (0.45-0.49)	0.44 (.42-.46)
Page-specific AOIs	Gaze + AOI	0.48 (.47-.50)	0.47 (.46-.49)	0.46 (.44-.48)	0.5 (.49-.50)	0.48 (.47-.49)	0.48 (.47-.49)	0.47 (.46-.49)	0.51 (.50-.52)
Synthesis	Gaze	0.49 (.46-.51)	0.57 (.55-.59)	0.6 (.58-.62)	0.48 (.46-.51)	0.52 (.50-.54)	0.48 (.46-.50)	0.5 (.48-.52)	0.53 (.51-.56)
Page-specific AOIs	Gaze + AOI	NA	0.5 (.49-.50)	0.5 (.49-.50)	NA	0.5 (.50-.50)	NA	0.5 (.50-.50)	0.5 (.49-.50)

Visual Complexity – Low: L and High: H; NA - no convergence reached

likely due to the longer times spent by the ASD participants[3]. A similar result of 0.71 is achieved both by extracting the data from page-specific AOIs and 4 x 4 grid AOIs, and compares to 0.66 for the 2 x 2 grid, indicating that the AOI granularity level is of higher importance compared to AOI type for a Browse task with a generous time limit. For the 30-second time limit condition, page-specific AOIs captured the differences between the groups better than the generic grids (0.65 compared to 0.60 for the 2 x 2 grid and 0.61 for the 4 x 4 grid). These results are related to specific web pages providing a better signal than others. If we look at results where not all of the "best" pages are included, best performance of 0.662 is achieved by combining AVG + GoDaddy + BBC for Browse 1. The results for the rest of the page combinations can be found in our repository. Other features such as AOI ID and Media ID do not improve the performance and, in many cases, slightly decrease the accuracy, except for Browse 1, where adding Media ID to the page-specific AOIs improves the result. In both tasks there is no clear pattern related to the visual complexity of the pages, as shown in Table IV and consistent with findings from [19]. The Gender variable generally lowered the results.

[3]While this seems a probable reason for the better classification accuracy, it is also possible that the superior results might be related to the different pages and/or participants. While this possibility cannot be ruled out using the presented design, it remains a valid conclusion that there is a discrimination signal for both time conditions, participant groups and stimuli sets.

B. Search and Synthesis Tasks

The best result of 0.75 was achieved for the Search task using the Apple and Babylon web pages (Table III). This means that increasing the time limit and level of difficulty (i.e. the Synthesis task) did not succeed in providing a stronger discrimination signal but instead, it provided a comparable accuracy of 0.73. These results show that even when using a different participant group and stimuli set, eye-tracking data can be used to classify the two groups with a consistent level of accuracy. Notably, the 4 x 4 grid, which was not previously explored in [19], provides the same accuracy for the two conditions. This might be due to 4 x 4 providing the same number of AOIs, while the page-specific AOIs for the different sets of web pages differed in number (discussed in Section V).

Similar to all previous results, best performance is achieved when using data from selected web pages (Apple + Babylon for the Search task and Amazon + Outlook for the Synthesis task). Interestingly, in all other tasks, combining the pages with best individual results led to optimal accuracy, except for the Synthesis task where the best performing individual page (BBC) was not included in any of the optimal combinations.

In most cases, adding variables such as AOI ID and whether or not the AOI contained the correct answer for the Search task decreased the prediction accuracy. Again, no clear pattern was observed with regards to page visual complexity (Table IV), which was consistent with findings in [29]. The Gender variable generally lowered the results.

V. Discussion

As can be seen from the results, i) all tasks provided a discrimination signal, and ii) the classifiers achieve a comparable accuracy to the one presented in the first study when using a different set of stimuli and participants. Best results were achieved using data from the Search task (0.75), followed by data from the first Browse task (0.74), the Synthesis task (0.73) and the second Browse task (0.65 for the 30-second condition). Examination of the 95% confidence intervals reveals that the models are robust in their predictions. Overall, these results confirm that using visual-processing differences as captured through eye-tracking data from web-searching tasks is a promising direction for automatic detection of autism.

Task Effects: The type of task affected the classification, where best performance was achieved using the Search task and the time-unlimited Browse task. Increased task complexity did not amplify the discrimination signal (Synthesis). While the presented design was not intended to test this explicitly, a possible explanation may be that the higher-order inferential processes required to complete this task were executed covertly, without being reflected in the eye-tracking data. Therefore, future research should test whether adding together more data points from simple Search tasks would provide better accuracy, as opposed to designing more complex tasks.

Differences in Browsing Strategies: Study 1 did not provide conclusive results to decide whether the discrimination signal for the Browse task reflected different browsing strategies in the two groups or simply reflected longer times for browsing within the ASD group. Limiting the time to 30 seconds in Study 2 showed that the two groups have different browsing strategies even when the time for browsing is controlled. This finding indicates that the visual processing of web pages, without the interference of a specific task, works differently in the two groups. The significance of this goes beyond the task of automatic autism detection, as it implies that people with autism may scan the elements of a web page in a different order or be drawn to specific elements (e.g., images) more than other elements (e.g., text), as suggested by [29]. Further analysis will be needed to confirm this.

AOI Identification Effects: We compared two types of AOIs: page-specific and generic. The results show that the AOI type has an effect on the performance. In the discussion here we aim to identify whether this significance is in favour of the AOI content (i.e. what the AOIs capture) or AOI granularity. In terms of AOI content, information from the page-specific AOIs seems to offer superior performance, as the best results for Search, Browse 1, and Browse 2 were achieved using page-specific AOIs. At the same time, having generic AOIs for the two Browse tasks did not affect classification accuracy as much as it did for the Search task, indicating that defining the AOIs in a meaningful way with respect to the page is more beneficial for Search tasks as it better captures differences in the visual-processing patterns. When all page elements have equal importance, as was the case with the two Browse tasks, the relevance of the areas to the content is lower. The effects of AOI content do not appear in isolation from granularity effects. While the 2 x 2 grid consistently provided the worst results for all tasks, the 4 x 4 grid had similar performance for Browse 1 and outperformed the page-specific AOIs for Synthesis. The latter could be explained by the different number of page-specific AOIs in the two studies as identified by the VIPS algorithm (112 for Study 1, corresponding to an average of 18.6 AOIs per page, and 213 for Study 2, corresponding to an average of 26.6 AOIs per page). Having more page-specific AOIs for Study 2 may have increased the level of noise and decreased accuracy. These results suggest a trade-off between having sufficiently detailed AOIs and not introducing too much noise. The optimal number of AOIs is best defined empirically.

Effects of Non-gaze Variables: The *visual complexity* of the individual web pages does not provide a clear pattern associated with better accuracy for any of the four tasks. *Participant gender* had mostly negative effect on the results, while *Media ID* generally lowered the accuracy, except for the best results for Browse 1, where it introduced a peak in the level of accuracy. The variable related to *correct answer to a Search task* had a slight positive effect on the Search task classifiers for the individual pages but when added to the best classifier it reduced the accuracy from 0.75 to 0.73.

Comparison with Prior Work: As mentioned in Section I-A, studies using fMRI data represent the state of the art in autism detection using objective markers (as opposed to subjective ones). Our results are comparable to the accuracy range of 71% to 86.7% from the fMRI studies with several important distinctions. First, some of the fMRI studies provide higher accuracy and a few of them test the validity of their classifiers on unseen data collected in other centres. This is currently not feasible for our approach due to the task-specific nature of the data, and the unavailability of comparable data collected independently. Nevertheless, we compare tasks using different stimuli and participant groups and we show that all the main effects from the first study are confirmed by the second one. These results are also slightly lower compared to the accuracy achieved using eye-tracking data for detecting autism in children. This is expected because the comparatively subtler differences between adults with and without high-functioning autism make the task more challenging. It is also possible that the signal extracted using facial stimuli is stronger than the signal extracted using web pages.

Limitations: Some of the limitations of this study are that web-searching tasks are not suitable for very young children and that there is a relatively small number of participants (comparable with those of state-of-the-art fMRI studies for autism detection but smaller than the number needed for large validation studies). The current design is also not able to provide a conclusive explanation of some the results, e.g., what makes some pages more suitable for eliciting larger between-

group differences and why the Synthesis task did not result in a better classification accuracy than the Search task, even though it is more complex. We can only speculate that the reasons for this are related to the covert nature of the inferential process required to solve the Synthesis questions or that the questions indicated too clearly where the participants needed to look, thus masking their natural searching patterns.

Future Work: In spite of the above limitations, the results from this study suggest various possibilities for future work. These include the development of a serious game that does not rely on an eye tracker but logs visual processing differences differently; using behavioural data obtained in a natural environment; and investigating whether gaze data could be used to detect other clinical conditions that exhibit atypical attention patterns such as dementia, schizophrenia, and ADHD, among others. For pursuing all these goals, there are three general questions that need to be explored: i) what makes a good task for eliciting larger between-group differences; ii) how to record these behaviours outside of laboratory conditions; and most importantly, iii) how to do this ethically.

VI. Conclusion

This paper presented two separate studies into automatic autism detection based on visual processing differences, using different participant and stimulus sets. Both studies confirmed the following main effects: i) that visual processing differences could potentially be used as a marker of autism with a comparable accuracy across participants and stimulus sets, ii) that people with autism process the information contained in web pages differently with and without specific information-location instructions and across different time conditions, and iii) that the content and granularity level of the areas of interest have an impact on the classification accuracy, while the visual complexity of the pages and the participant gender do not.

References

[1] American Psychiatric Association, "Diagnostic and Statistical Manual of Mental Disorders (5th ed.)," Arlington, VA, 2013.
[2] U. Frith, *Autism: Explaining the enigma*. Blackwell Publishing, 2003.
[3] A. Bernas, A. P. Aldenkamp, and S. Zinger, "Wavelet coherence-based classifier: a resting-state functional mri study on neurodynamics in adolescents with high-functioning autism," *Computer methods and programs in biomedicine*, vol. 154, pp. 143–151, 2018.
[4] T. Falkmer, K. Anderson, M. Falkmer, and C. Horlin, "Diagnostic procedures in autism spectrum disorders: a systematic literature review," *European child & adolescent psychiatry*, vol. 22, no. 6, pp. 329–340, 2013.
[5] L. D. Wiggins, J. Baio, and C. Rice, "Examination of the time between first evaluation and first autism spectrum diagnosis in a population-based sample," *Journal of Developmental and Behavioral Pediatrics*, vol. 27, no. 2, pp. S79–S87, 2006.
[6] M. Davidovitch, N. Levit-Binnun, D. Golan, and P. Manning-Courtney, "Late diagnosis of autism spectrum disorder after initial negative assessment by a multidisciplinary team," *Journal of Developmental & Behavioral Pediatrics*, vol. 36, no. 4, pp. 227–234, 2015.
[7] C. M. Murphy, C. E. Wilson, D. M. Robertson, C. Ecker, E. M. Daly, N. Hammond, A. Galanopoulos, I. Dud, D. G. Murphy, and G. M. McAlonan, "Autism spectrum disorder in adults: diagnosis, management, and health services development," *Neuropsychiatric disease and treatment*, vol. 12, p. 1669, 2016.
[8] J. S. Anderson, J. A. Nielsen, A. L. Froehlich, M. B. DuBray, T. J. Druzgal, A. N. Cariello, J. R. Cooperrider, B. A. Zielinski, C. Ravichandran, P. T. Fletcher *et al.*, "Functional connectivity magnetic resonance imaging classification of autism," *Brain*, vol. 134, no. 12, pp. 3742–3754, 2011.
[9] L. Q. Uddin, K. Supekar, C. J. Lynch, A. Khouzam, J. Phillips, C. Feinstein, S. Ryali, and V. Menon, "Salience network–based classification and prediction of symptom severity in children with autism," *JAMA psychiatry*, vol. 70, no. 8, pp. 869–879, 2013.
[10] M. Plitt, K. A. Barnes, and A. Martin, "Functional connectivity classification of autism identifies highly predictive brain features but falls short of biomarker standards," *NeuroImage: Clinical*, vol. 7, pp. 359–366, 2015.
[11] S. Ibrahim, R. Djemal, and A. Alsuwailem, "Electroencephalography (eeg) signal processing for epilepsy and autism spectrum disorder diagnosis," *Biocybernetics and Biomedical Engineering*, vol. 38, no. 1, pp. 16–26, 2018.
[12] M. Asgari, A. Bayestehtashk, and I. Shafran, "Robust and accurate features for detecting and diagnosing autism spectrum disorders." in *Interspeech*, 2013, pp. 191–194.
[13] C. Laidi, J. Boisgontier, M. M. Chakravarty, S. Hotier, M.-A. dAlbis, J.-F. Mangin, G. A. Devenyi, R. Delorme, F. Bolognani, C. Czech *et al.*, "Cerebellar anatomical alterations and attention to eyes in autism," *Scientific reports*, vol. 7, no. 1, p. 12008, 2017.
[14] R. S. Hessels, I. T. Hooge, T. M. Snijders, and C. Kemner, "Is there a limit to the superiority of individuals with asd in visual search?" *Journal of autism and developmental disorders*, vol. 44, no. 2, pp. 443–451, 2014.
[15] R. S. Hessels, R. Andersson, I. T. Hooge, M. Nyström, and C. Kemner, "Consequences of eye color, positioning, and head movement for eye-tracking data quality in infant research," *Infancy*, vol. 20, no. 6, pp. 601–633, 2015.
[16] G. Wan, X. Kong, B. Sun, S. Yu, Y. Tu, J. Park, C. Lang, M. Koh, Z. Wei, Z. Feng *et al.*, "Applying eye tracking to identify autism spectrum disorder in children," *Journal of autism and developmental disorders*, vol. 49, no. 1, pp. 209–215, 2019.
[17] W. Liu, M. Li, and L. Yi, "Identifying children with autism spectrum disorder based on their face processing abnormality: A machine learning framework," *Autism Research*, vol. 9, no. 8, pp. 888–898, 2016.
[18] S. Eraslan, V. Yaneva, Y. Yesilada, and S. Harper, "Web users with autism: eye tracking evidence for differences," *Behaviour & Information Technology*, vol. 38, no. 7, pp. 678–700, 2019.
[19] V. Yaneva, L. A. Ha, S. Eraslan, Y. Yesilada, and R. Mitkov, "Detecting autism based on eye-tracking data from web searching tasks," in *Proceedings of the Internet of Accessible Things*. ACM, 2018, p. 16.
[20] C. Lord, M. Rutter, S. Goode, J. Heemsbergen, H. Jordan, L. Mawhood, and E. Schopler, "Austism diagnostic observation schedule: A standardized observation of communicative and social behavior," *Journal of autism and developmental disorders*, vol. 19, no. 2, pp. 185–212, 1989.
[21] S. Baron-Cohen, S. Wheelwright, R. Skinner, J. Martin, and E. Clubley, "The autism-spectrum quotient (aq): Evidence from asperger syndrome/high-functioning autism, malesand females, scientists and mathematicians," *Journal of autism and developmental disorders*, vol. 31, no. 1, pp. 5–17, 2001.
[22] V. Yaneva, I. Temnikova, and R. Mitkov, "Accessible texts for autism: An eye-tracking study," in *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility*. ACM, 2015, pp. 49–57.
[23] S. Eraslan, V. Yaneva, Y. Yesilada, and S. Harper, "Do web users with autism experience barriers when searching for information within web pages?" in *Proceedings of the 14th Web for All Conference on The Future of Accessible Work*. ACM, 2017, p. 20.
[24] E. Michailidou, "Visual complexity rankings and accessibility metrics," Ph.D. dissertation, The University of Manchester, 2010.
[25] G. Marchionini, "Exploratory search: from finding to understanding," *Communications of the ACM*, vol. 49, no. 4, pp. 41–46, 2006.
[26] R. J. Jacob, "Eye tracking in advanced interface design," *Virtual environments and advanced interface design*, pp. 258–288, 1995.
[27] M. E. Akpınar and Y. Yeşilada, "Vision Based Page Segmentation Algorithm: Extended and Perceived Success," in *Current Trends in Web Engineering*, ser. LNCS, Q. Z. Sheng and J. Kjeldskov, Eds., vol. 8295. Springer, 2013, pp. 238–252.
[28] M. Roy and M. T. Chi, "Gender differences in patterns of searching the web," *Journal of educational computing research*, vol. 29, no. 3, pp. 335–348, 2003.
[29] V. Yaneva, I. Temnikova, and R. Mitkov, "Accessible texts for autism: An eye-tracking study," in *Proceedings of the 17th International ACM SIGACCESS Conference on Computers and Accessibility*, ser. ASSETS '15. New York, NY, USA: ACM, 2015, pp. 49–57. [Online]. Available: http://doi.acm.org/10.1145/2700648.2809852

APPENDIX A
TASKS IN STUDY 1 AND STUDY 2

TABLE V
PAGES, COMPLEXITY AND SEARCH TASKS FOR STUDY 1

Task (Complexity)	Search Tasks
Apple (Low)	a) Can you locate the link that allows to watch the TV ads relating to iPad mini?
	b) Can you locate a link labelled iPad on the main menu?
Babylon (Low)	a) Can you locate the link that you can download the free version of Babylon?
	b) Can you find and read the names of other products of Babylon?
AVG (Medium)	a) Can you locate the link which you can download the free trial of AVG Internet Security 2013?
	b) Can you locate the link which allows you to download AVG Antivirus Free 2013?
Yahoo! (Medium)	a) Can you read the titles of the main headlines which have smaller images?
	b) Can you read the first item under the News title?
GoDaddy (High)	a) Can you find a telephone number for technical support and read it?
	b) Can you locate the text box where you can search for a new domain?
BBC (High)	a) Can you read the first item of Sport News?
	b) Can you locate the table that shows market data under the Business title?

TABLE VI
PAGES, COMPLEXITY AND SYNTHESIS TASKS FOR STUDY 2

Page (Complexity)	Synthesis Tasks
Wordpress (Low)	(a) Which of the WordPress plans offers community support instead of email support?
	(b) What is the cheapest plan you can get that offers Email & Live Chat support?
WhatsApp (Low)	(a) Under frequently asked questions, what topic features in both the iPhone and Android columns?
	(b) Under frequently asked questions, which sections feature topics relating to notifications?
Outlook (Low)	(a) According to the text, is Twitter a partner app of Outlook?
	(b) The names of which apps are both mentioned in the text and presented as logos below the text?
Netflix (Low)	(a) Which is the cheapest plan that allows you to watch movies on your laptop, TV and tablet?
	(b) How much more would you have to pay comp. to the basic plan if you wanted to have Ultra HD?
YouTube (High)	(a) Under the American football category, which videos have been posted within the last three months?
	(b) Under the NBA topic, which video has the largest number of views?
Amazon (High)	(a) Which item has the largest price discount measured in percentage?
	(b) Which product has been rated by the largest number of users?
Adobe (High)	(a) Which is the product that is targeted to UX designers and for which students can get a discount?
	(b) How many types of clouds are offered by Adobe within this page?
BBC (High)	(a) According to the schedule, which Grand Prix takes place first: the Australia or the Bahrain?
	(b) Which of the following sports does not have its own tab on BBC Sport: Golf, Cricket or Volleyball?

APPENDIX B
CLASSIFIER SELECTION (STUDY 1)

TABLE VII
CLASSIFIER SELECTION: COMPARISON BETWEEN ALGORITHMS (ACCURACY AND 95% CONFIDENCE INTERVALS)

Algorithm	Search, page-specific AOIs	Browse (1), page-specific AOIs
Logistic Regression	0.70 (0.68, -0.72)	0.66 (0.64, -0.68)
Random Forest	0.65 (0.63,0.67)	0.63 (0.60,0.65)
SVM	0.55 (0.53,0.56)	0.63 (0.61,0.65)
NaiveBayes	0.51 (0.51,0.52)	0.55 (0.54,0.57)

APPENDIX C
RESULTS IN STUDY 1

TABLE VIII
GAZE FEATURE RESULTS - BROWSE (STUDY 1) WITH 95% CONFIDENCE INTERVALS (PAGE-SPECIFIC AOIS)

Page	Gaze Features	Gaze Features + AOI ID
Apple	.63 (.61-.66)	.47 (.46-.48)
Babylon	.54 (.52-.56)	.50 (.48-.52)
AVG	.59 (.57-.62)	.57 (.54-.60)
BBC	.45 (.42-.47)	.49 (.48-.50)
GoDaddy	.55 (.53-.58)	.47 (.46-.48)
Yahoo!	.52 (0.50-.55)	.52 (.50-.53)

TABLE IX
CONFUSION MATRIX FOR THE FIRST BROWSE TASK BEST RESULT: ALL GAZE FEATURES + SELECTED MEDIA (APPLE + AVG). THE NUMBER IS CALCULATED THROUGH DIVIDING 500 (100 TEST TRIALS WITH FIVE PARTICIPANTS PER GROUP) BY THE NUMBER OF CORRECTLY LABELLED PARTICIPANTS FOR EACH CLASS.

	Control - Predicted	ASD - Predicted
Control - Actual	0.722	0.294
ASD - Actual	0.278	0.706

TABLE X
RESULTS FOR INDIVIDUAL WEB PAGES FOR THE SEARCH TASK WITH 95% CONFIDENCE INTERVALS (PAGE-SPECIFIC AOIS)

Page	Gaze Features	Gaze Features + AOI ID	Gaze + Correct Answer AOI
Apple	.69 (.67-.72)	.51 (.50-.52)	.69 (.67-.72)
Babylon	.63 (.60-.65)	.49 (.48-.50)	.61 (.58-.63
AVG	.39 (.37-.42)	.48 (.47-.50)	.43 (.41-.46)
BBC	.57 (.55-.59)	.50 (.50-.50)	.55 (.52-.57)
GoDaddy	.60 (.58-.63)	.48 (.47-.49)	.58 (.55-.60)
Yahoo!	.48 (.46-.50)	.50 (.49-.52)	.47 (.46-.49)

TABLE XI
CONFUSION MATRIX FOR THE SEARCH TASK BEST RESULT: ALL GAZE FEATURES + SELECTED MEDIA (APPLE + BABYLON).

	Control - Predicted	ASD - Predicted
Control - Actual	0.698	0.192
ASD - Actual	0.302	0.808

APPENDIX D
RESULTS FOR GENDER AND COMPLEXITY IN STUDY 2

TABLE XII
RESULTS FOR GENDER AND VISUAL COMPLEXITY (ACCURACY AND 95% CONFIDENCE INTERVALS)

	Browse, Study 1		**Browse, Study 2**		**Search, Study 1**		**Synthesis, Study 2**	
Feature set	**All**	**Best**	**All**	**Best**	**All**	**Best**	**All**	**Best**
Page specific AOIs + Gender	.53 (.50,.56)	.65 (.62,.68) (AP, AVG)	.59 (.56,.61)	.66 (.64,.69) (AB, NF, OL, WA, WP, YT)	.51 (.49,.54)	.61 (.57,.64) (AP)	.58 (.56,.60)	.64 (.62,.66) (BBC, WP)
2 x 2 grid + Gender	.49 (.46,.51)	.69 (.66,.72) (AP, BL)	.62 (.60,.64)	.65 (.63,.67) (AB, WA, WP, YT)	.58 (.55,.60)	.57 (.54,.60) (BL, BBC)	.59 (.57,.61)	.62 (.59,.65) (AB, OL, YT)
4 x 4 grid + Gender	.60 (.57,.63)	.67 (.64,.69) (AVG, GD)	.63 (.60,.65)	.66 (.63,.68) (AB, BBC, NF, WA, WP, YT)	.53 (.50,.56)	.57 (.54,.59) (GD)	.6 (.58,.62)	.63 (.60,.65) (AB, AZ, BBC, OL, WA, YT)
Page specific AOIs + VC	.66 (.64,.68)	.75 (.73,.78) (AP, AVG)	.61 (.59,.63)	.61 (.59,.63) (AB, NF, OL)	.7 (.68,.72)	.76 (.74,.78) (AP, BL, AVG)	.61 (.58,.63)	.67 (.65,.69) (AB, AZ, OL)
2 x 2 grid + VC	.56 (.53,.58)	.70 (.68,.72) (AP, BL, AVG, GD)	.49 (.46,.51)	.60 (.58,.62) (BBC, WP, YT)	.56 (.53,.58)	.67 (.65,.70) (BL, BBC)	.63 (.61,.65)	.67 (.65,.69) (AZ, NF, OL)
4 x 4 grid + VC	.66 (.64,.68)	.71 (.69,.73) (AP, AVG)	.48 (.46,.49)	.60 (.58,.63) (OL, WA, YT)	.66 (.64,.68)	.73 (.71,.75) (AP, GD, YH)	.64 (.62,.66)	.72 (.70,.74) (AZ, OL)

Study 1–Apple: AP, Babylon: BL, AVG: AVG, Yahoo: YH, Godaddy: GD, BBC: BBC
Study 2–Adobe: AB, Amazon: AZ, BBC: BBC, Netflix: NF, Outlook: OL, WhatsApp: WA, WordPress: WP, YouTube: YT