Eye tracking as a new interface for image retrieval

O K Oyekoya and F W M Stentiford

Different modes of human-computer interaction will play a major part in making computing increasingly pervasive. More natural methods of interaction are in demand to replace devices such as the keyboard and the mouse, and it is becoming more important to develop the next generation of human-computer interfaces that can anticipate the user's intended actions. Human behaviour depends on highly developed abilities to perceive and interpret visual information and provides a medium for the next generation of image retrieval interfaces. If the computer can correctly interpret the user's eye gaze behaviour, it will be able to anticipate the user's objectives and retrieve images and video extremely rapidly and with a minimum of thought and manual involvement.

1. Introduction

The best interfaces are the most natural ones. They are unobtrusive and provide relevant information quickly and in ways that do not interfere with the task itself.

The disappearance of technologies into the fabric of everyday life is as a result of human psychology rather than technology [1]. There are many challenges in computing that have to be overcome before the dream of integrating information technology with human users can be achieved.

This is still a distant vision, mainly because of hardware constraints [2], but also because there are serious human-computer interaction (HCI) issues to consider. Social and cognitive factors are just as important in making computers inconspicuous [3].

Eye tracking and other natural methods such as voice and body gestures will play an important part in solving the cognitive issues of pervasive computing. Eye tracking offers a new way of communicating with human thought processes.

This paper addresses the problem of retrieving images using a natural interface for search. Understanding the fixations and saccades in human eye movement data and its validation against a visual attention model suggests a new image retrieval interface that uses new eye tracking technology. Such a model will not only have to identify the items of interest within the image but also characterise it according to its relative importance.

Section 2 describes related research and the key issues that are addressed. This is followed in section 3 by a description of the current research with results from preliminary experiments. Section 4 discusses some outstanding issues including the cognitive factors that have to be overcome. The final section provides some conclusions and an indication of the future work.

2. State of the art

Research activity in eye tracking has increased in the last few years due to improvements in performance and reductions in the costs of eye-tracking devices. The research is considered under three headings:

- eye tracking technology,
- human behaviour,
- current applications.

2.1 Eye-tracking technology

A number of eye-gaze detection methods have been developed over the years. Invasive methods that required tampering directly with the eyes were mostly used before the 1970s. The search coil method [4] offers high accuracy and large dynamic range but requires an insertion into the eye! Non-invasive methods such as the DPI (Dual Purkinje Image) eye tracker [5] require the head to be restricted and are relatively expensive.

More recently systems have appeared that use video images and/or infrared cameras. The FreeGaze system

[6] attempts to limit errors arising from calibration and gaze detection by using only two points for individual personal calibration. The position of observed pupil image is used directly to compute the gaze direction but this may not be in the right place due to refraction in the surface of the cornea. The eyeball model corrects the pupil position for obtaining a more accurate gaze direction. ASL [7], Smarteye [8], IBM's Almaden [9], Arrington's Viewpoint [10], SR's Eyelink [11] and CRS [12] eye trackers are examples of recent commercial eye trackers. A typical commercial eye tracker tracks the pupil and the first Purkinje image (corneal reflex) and the difference gives a measure of eye rotation. Section 3 describes LC Technology's Eyegaze system [13] used in this research in more detail.

Several methods of improving the accuracy of estimating gaze direction and inferring intent from eye movement have been proposed. The Eye-R system [14] is designed to be battery operated and is mounted on any pair of glasses. It measures eye motion using infra-red technology by monitoring light fluctuations from infra-red light and utilises this as an implicit input channel to a sensor system and computer. Mulligan and Beutter [15] use a low-cost approach to track eye movement using compressed video images of the fundus on the back surface of the eyeball. A technical challenge for these types of trackers is the real-time digitisation and storage of the video stream from the cameras. Bhaskar et al [16] propose a method that uses eye-blink detection to locate an eye which is then tracked using an eye tracker. Blinking is necessary for the tracker to work well and the user has to be aware of this.

2.2 Human behaviour

Experiments have been conducted to explore human gaze behaviour for different purposes. Privitera et al [17] use ten image processing algorithms to compare human identified regions of interest with regions of interest determined by an eye tracker and defined by a fixation algorithm. The comparative approach used a similarity measurement to compare two aROIs (algorithmically-detected regions of interest), two hROIs (human-identified regions of interest) and an aROI plus hROI. The prediction accuracy was compared in order to identify the best matching algorithms. Different algorithms fared better under differing conditions. They concluded that aROIs cannot always be expected to be similar to hROIs in the same image because two hROIs produce different results in separate runs. This means that algorithms are unable in general to predict the sequential ordering of fixation points.

Jaimes et al [18] compare eye movement across categories and link category-specific eye-tracking results to automatic image classification techniques. They hypothesise that the eye movements of human observers differ for images of different semantic categories, and that this information can be effectively used in automatic content-based classifiers. The eye tracking results suggest that similar viewing patterns occur when different subjects view different images in the same semantic category. They suggested that it is possible to apply the Privitera's fixation clustering approach [17] to cluster gaze points.

Pomplun and Ritter [19] present a three-level model, which is able to explain about 98% of empirical data collected in six different experiments of comparative visual search. Pairs of almost identical items are compared requiring subjects to switch between images several times before detecting a possible mismatch. The model consists of the global scan path strategy (upper level), shifts of attention between two visual hemifields (intermediate level) and eye movement patterns (lower level). Simulated gaze trajectories obtained from this model are compared with experimental data. Results suggest that the model data of most variables presents a remarkably good correspondence to the empirical data.

Identification and analysis of fixations and saccades in eye-tracking protocol is important in understanding visual behaviour. Salvucci [20] classifies algorithms with respect to five spatial and temporal characteristics. The spatial criteria divide algorithms in terms of their use of velocity, dispersion of fixation points, and areas of interest information. The temporal criteria divide algorithms in terms of their use of duration information and their local adaptivity. Five fixation identification algorithms are described and compared in terms of their accuracy, speed, robustness, ease of implementation, and parameters. The results show that hidden Markov models based on the dispersion threshold fare better in terms of their accuracy and robustness. The Minimum Spanning Tree uses a minimised connected set of points and provides robust identification of fixation points, but runs slower due to the two-step approach of construction and search of the minimum spanning trees. The velocity threshold has the simplest algorithm and is thus fast but not robust. Areas of interest are found to perform poorly on all fronts. These findings are implemented in the EyeTracer system [21], an interactive environment for manipulating, viewing, and analysing eye-movement protocols.

Stone and Beutter (from NASA) [22] focus on the development and testing of human eye-movement control with particular emphasis on search saccades and the response to motion (smooth pursuit). They conclude that current models of pursuit should be modified to include visual input that estimates object motion and not merely retinal image motion as in current models.

Duchowski [23] presents a 3-D eye-movement analysis algorithm for binocular eye tracking within virtual reality. Its signal analysis techniques can be categorised into three — position-variance, velocity-based and ROI-based, again using two of Salvucci's criteria. This is easily adapted to a 2-D environment by holding head position and visual angle constant.

2.3 Current applications

Eye-tracking equipment is used as an interface device in several diverse applications. Schnell and Wu [24] apply eye tracking as an alternative method for the activation of controls and functions in aircraft.

Dasher [25] is a method for text entry that relies purely on gaze direction. The user composes text by looking at characters as they stream across the screen from right to left. Dasher presents likely characters in sizes according to the probability of their occurrence in that position. The user is often able to select rapidly whole words or phases as their size increases on the screen.

Nikolov et al propose [26] a system for construction of gaze-contingent multi-modality displays of multi-layered geographical maps. Gaze contingent multi-resolutional displays (GCMRDs) centre high-resolution information on the user's gaze position, matching the user's interest. In this system, different map information is channelled to the central and the peripheral vision giving real performance advantage.

Nokia [27] conducted a usability evaluation on two mobile Internet sites and discovered the importance of search on mobile telephones contrary to the initial hypothesis that users would not like to use search because of the effort of keying inputs. The research also showed that customers prefer any interface that produces a successful search. This evaluation confirms that users do have a need for information retrieval for mobile usage.

Xin Fan et al [28] propose an image-viewing technique based on an adaptive attention-shifting model, which looks at the issue of browsing large images on limited and heterogeneous screen zones of mobile telephones. Xin Fan's paper focuses on facilitating image viewing on devices with limited display sizes.

The Collage Machine [29] is an agent of Web recombination. It deconstructs Web sites and re-presents them in collage form. It can be taught to bring media of interest to the user on the basis of the user's interactions. The evolving model provides an extremely flexible way of presenting relevant visual information to the user.

Cognitive interest is hard to measure and so any steps taken to suggest user selection will improve performance and allow users to change their mind. Farid [30] describes the implementation and initial experimentation of systems based on a user's eye-gaze behaviour. It was concluded that the systems performed well because of minimal latency and obtrusiveness. A zooming technique is adopted with a magnified region of interest and multiple video streams.

Eye tracking is being used successfully for various applications and experimental purposes, but outstanding issues include accuracy and interpretation. The research described in this paper uses an attention model to assist in the interpretation of users' eye-gaze behaviour when conducting a visual search and it is hoped that this will lead to a more intimate and rapid interface for content-based image retrieval (CBIR).

3. Current research objectives

The aim of this research is to provide a rapid and natural interface for searching visual digital data in a CBIR system (Fig 1). A pre-computed network of similarities between image regions in an image collection can be traversed very rapidly using eye tracking providing the users' gaze behaviours yield suitable information about their intentions. It is reasonable to believe that users will look at the objects in which they are interested during a search [31] and this provides the machine with the necessary information to retrieve plausible candidate images for the user. Such images will contain regions that possess similarity links with the gazed regions, and can be presented to the user in a variety of ways. Dasher's text entry [25] and Kerne's Collage Machine [29] both provide promising CBIR interfaces for future investigation.

Initial experiments have investigated the gaze behaviour of participants, and compared it with data obtained through a model of visual attention (VA) [32]. This enabled possible differences in behaviour to be detected arising from varying image content. Regions of interest are identified both by human interaction and prior analysis and used to explore aspects of vision that would not otherwise be apparent. Images with and without obvious subjects were used in this work to accentuate any behaviour differences that might be apparent.

3.1 System overview

The system design is broken down into two major components as shown in Fig 2:

- algorithmic analysis of image to obtain visual attention scores,
- human identification of region of interest.

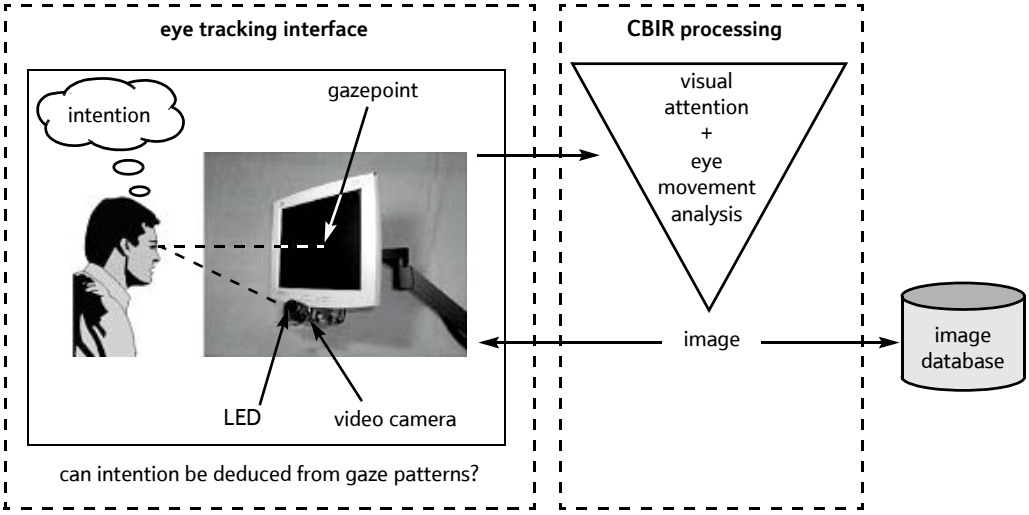

Fig 1 Proposed system architecture.

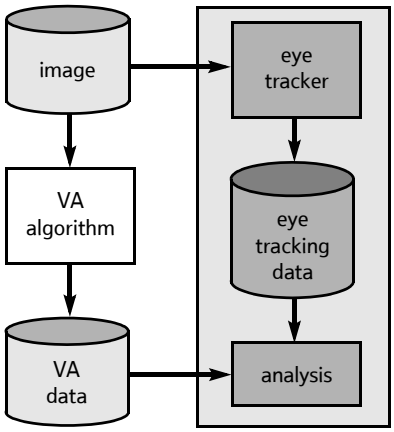

Fig 2 Experimental process.

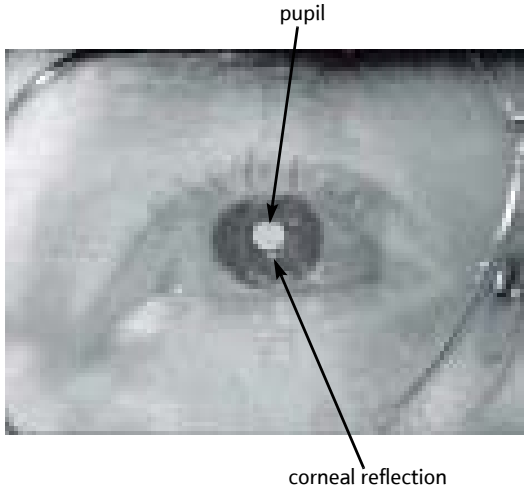

Fig 3 Camera image of eye, illustrating bright image pupil and corneal reflection.

It should be noted that the analysis process is grouped with the eye-tracking process because the main goal is to carry out real-time analysis to identify objects of interest for use in real-world applications.

3.2 Eye tracking equipment

The Eyegaze system [13] is an eyetracker designed to measure where a person is looking on a computer screen. The Eyegaze system tracks the subject's gazepoint on the screen automatically and in real time. It uses the pupil-centre/corneal-reflection (PCCR) method to determine the eye's gaze direction. A video camera located below the computer screen remotely and unobtrusively observes the subject's eye (as shown in Fig 1). No attachments to the head are required. A small, low power, infra-red light emitting diode (LED) located at the centre of the camera lens illuminates the eye. The LED generates the corneal reflection and causes the bright pupil effect, which enhances the camera's image of the pupil (Fig 3).

Specialised image-processing software in the Eyegaze computer identifies and locates the centres of both the pupil and corneal reflection. Trigonometric calculations project the person's gaze-point based on the positions of the pupil centre and the corneal reflection within the video image. The Eyegaze system generates raw gaze-point location data at the camera field rate of 60 Hz. The procedure to calibrate the Eyegaze system is robust yet fast and easy to perform. The calibration procedure takes approximately 15 seconds and is fully automatic; no assistance from another person is required. The procedure does not accept full calibration until the overall gaze prediction accuracy and consistency exceed desired thresholds. To achieve high gazepoint tracking accuracy, the image processing algorithms in the Eyegaze system explicitly accommodate several common sources of gaze-point tracking error such as nonlinear gaze-point tracking equations, head-range variation, pupil-diameter variation and glint that straddles the pupil edge. A chair with head rest provides support for chin and forehead in order to minimise the effects of head movements, although the eye tracker does accommodate head movement of up to 1.5 inches (3.8cm).

3.3 Overview of the visual attention model

The model used in this work [32] employs an algorithm that assigns high VA scores to pixels where neighbouring pixel configurations do not match identical positional arrangements in other randomly selected neighbourhoods in the image. This means, for example, that high scores will be associated with anomalous objects, or edges and boundaries, if those features do not predominate in the image. A flowchart describing this process is given in Fig 4.

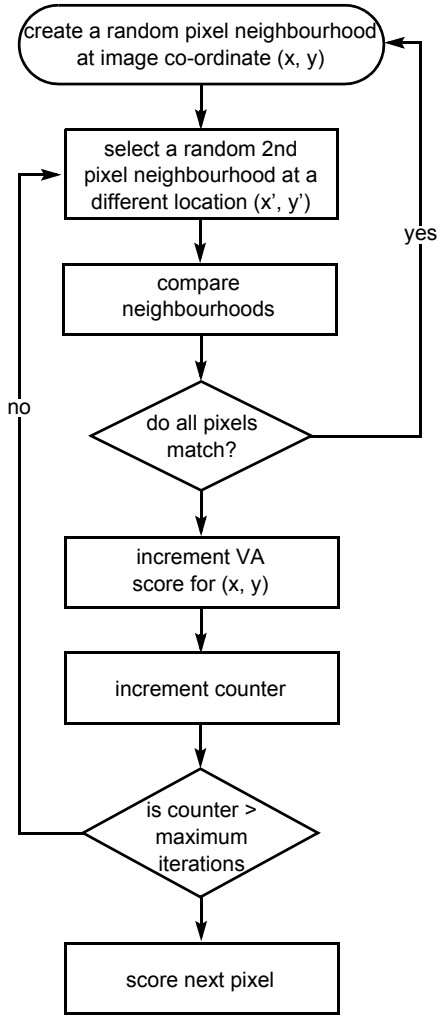

Fig 4 Visual attention model.

For display purposes the VA scores for each pixel are displayed as a map using a continuous spectrum of false colours with the scores being marked with a distinctive colour or grey level (as in Figs 6 and 7 in section 3.5).

3.4 Experiment design

In this experiment the VA algorithm is applied to each image to identify regions of interest and obtain VA scores for each pixel. It should be noted that the parameter settings are the same for all the images used. The images are viewed by a human participant and eye-tracking data is gathered using the Eyegaze eye tracker. The VA and eye-tracking data is then combined and analysed by identifying the co-ordinates of the gaze points on the image and obtaining the VA scores from the corresponding pixel position. VA scores are then plotted against time for each image and subject as illustrated later in Figs 6 and 7.

All participants had normal or corrected-to-normal vision and had no knowledge of the purpose of the study. Participants included a mix of graduates and administrative staff.

Over the course of the experiment, 4 participants were presented 20 images for 5 sec each separated by displays of a blank screen followed by a central black dot on a white background (Fig 5). These images were displayed on a 15 inch LCD flat panel monitor at a resolution of 1024 $\times$ 768 pixels. All participants were encouraged to minimise head movement and were asked to focus on the dot before each image was loaded.

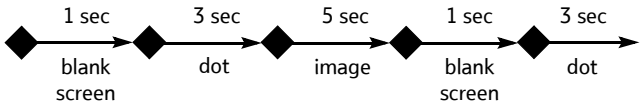

Fig 5 Display sequence.

3.5 Results

Figures 6 and 7 show two of the images used in the experiments together with corresponding VA maps and graphs for four subjects. The locations of saccades and fixations performed by the subjects on each of the images are recorded by the eye-tracking system. The VA score that corresponds to the pixel at each fixation point is associated with the time of the fixation and plotted as graphs for study in units of 20 ms.

It is observed that there is considerable variation in behaviour over the four subjects, but all viewed the regions with the highest VA scores early in the display period.

The variance ν of the VA score (x) over time was calculated as follows:

$$\nu = \frac{n\sum x^2 - (\sum x)^2}{n(n-1)}$$

The variance ν measures the average spread or variability of the data series x. The variances of the VA scores for the duration of the display over the six images for each subject are shown in Table 1 and Fig 8.

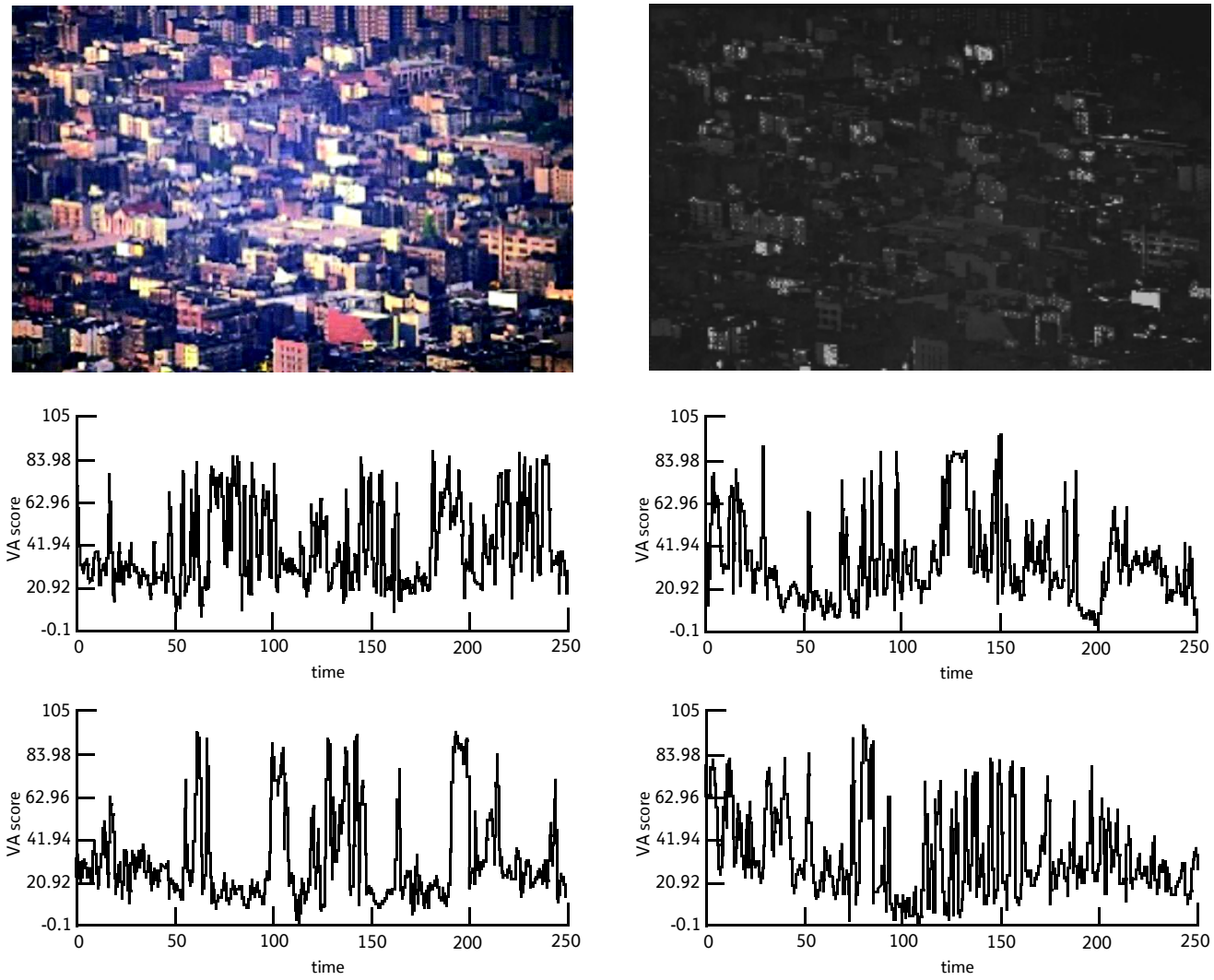

Fig 6 No obvious subject image — VA map and plots for 4 subjects.

Table 1 Variance of VA score.

		Subjects			
		1	2	3	4
Unclear ROI	Image1	325	193	333	532
	Image2	479	496	328	629
	Image3	389	175	365	197
Obvious ROI	Image4	741	687	1094	857
	Image5	1432	1453	1202	1466
	Image6	1246	1226	862	1497

4. Discussion

The goal of the initial experiment was to explore the relationship between gaze behaviour and the visual attention model in determining eye-movement patterns during different stages of viewing.

Results indicate that regions with high VA scores do attract eye gaze for those images studied. However, it was apparent that individual behaviours varied considerably and it was difficult to identify a pattern over such a small amount of data. Nevertheless the results did show that there was a higher variance in VA score over time on images with obvious regions of interest due to gaze patterns shifting between areas of high visual attention and the background.

This would seem reasonable in view of a natural inclination to make rapid visual comparisons between anomalous material and a relatively predictable background.

Accurate interpretation of interest is necessary for a successful interface. Fixations above a certain threshold and pursuit movement above a set velocity are just some of the factors that can be interpreted as an indication of interest. The findings by Jaimes et al [18] suggest that similar viewing patterns occur when different subjects view different images in the same semantic category. Hence, discrimination within an image might yield useful interpretation of interest.

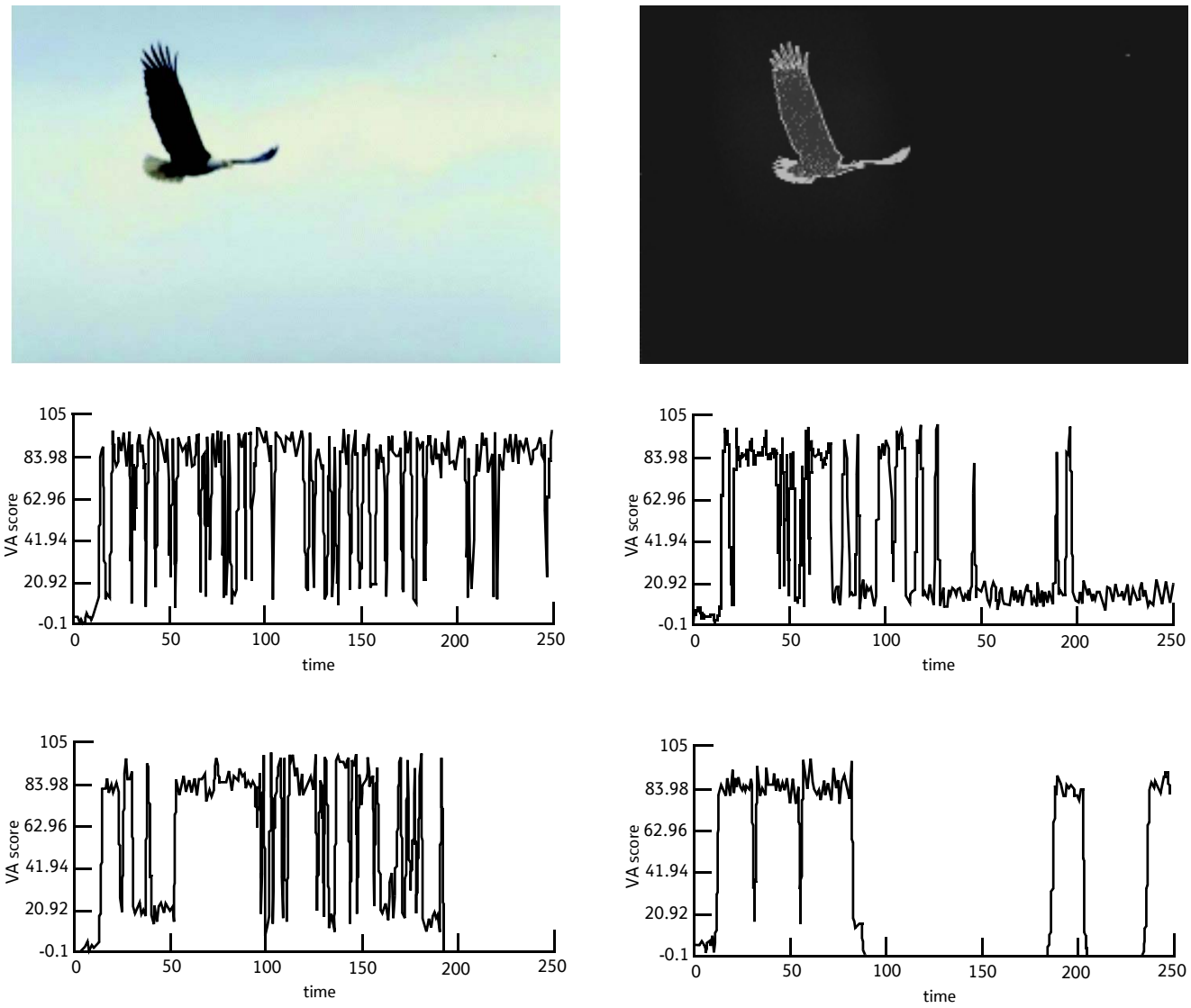

Fig 7 Obvious subject image — VA map and plots for 4 subjects.

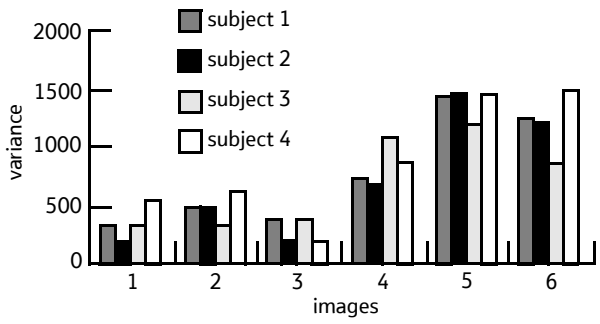

Fig 8 Variance histogram.

The accuracy of gaze location is an important factor in the results and some of the noise may be due to head and body movement as well as the basic accuracy of the equipment.

There is clear evidence [33] that users do not need to look directly at objects during covert attention.

This means that gaze direction does not necessarily indicate a current region of interest, only a general direction and could confound some conclusions.

Overall, the technical challenges still facing eye-tracking approaches include accuracy, simultaneous tracking and capturing of a visual scene, and most importantly interpreting gaze behaviour.

5. Conclusions and future directions

Preliminary experiments have confirmed that clear regions of interest in images lead to the attraction of eye gaze, which are not inconsistent with the visual attention model.

This gives credence to the belief that eye trackers can provide a new and exciting interface technology that promises to inspire a new range of computational tools which react to our thoughts and feelings rather than our hands.

Experiments are planned to investigate gaze behaviour in more constrained conditions in which users are focused on specific visual search tasks. This will reduce (but not eliminate) the confounding effects of users' prior interests and associated behaviours. The attention graphs should reveal details of gaze behaviour that can be utilised during CBIR operations.

Acknowledgements

The authors would like to thank the EPSRC, BT Exact, SIRA and the Imaging Faraday Partnership for their support in this work.

References

1 Weiser M: 'The computer for the 21st century', Scientific American (2003).

2 Satyanarayanan M: 'Pervasive computing: vision and challenges', IEEE Personal Communications, pp 10—17 (2001).

3 Lueg C: 'On the gap between vision and feasibility', Proceedings of the International Conference on Pervasive Computing, Springer Lecture Notes in Computer Science (LNCS 1414), pp 45—57 (2002).

4 Robinson D A: 'A method of measuring eye movement using a scleral search coil in a magnetic field', IEEE Transactions on Biomedical Electronics, BME-10, pp 137—145 (1963).

5 Crane H D and Steele C S: 'Accurate three-dimensional eye tracker', Applied Optics, 17, pp 691—705 (1978).

6 Ohno T, Mukawa N and Yoshikawa A: 'FreeGaze: a gaze tracking system for everyday gaze interaction', Proceedings of the Eye Tracking Research and Applications Symposium, pp 125—132 (2002).

7 ASL — http://www.a-s-l.com/

8 Smarteye — http://www.smarteye.se/

9 Almaden — http://www.almaden.ibm.com/cs/blueeyes/

10 Viewpoint — http://www.arringtonresearch.com/

11 SR Research — http://www.eyelinkinfo.com/

12 CRS — http://www.crsltd.com/

13 LC Technology — http://www.eyegaze.com/

14 Selker T, Lockerd A and Martinez J: 'Eye-R, a glasses-mounted eye motion detection interface', CHI '01, extended abstracts on Human Factors in Computer Systems, Seattle, Washington (2001).

15 Mulligan J B and Beutter B R: 'Eye movement tracking using compressed video images', Vision Sciences and its Applications: Optical Society Technical Digest Series, 1, pp 163—166 (1995).

16 Bhaskar T N, Foo T K, Ranganath S and Venkatesh Y V: 'Blink detection and eye tracking for eye localization', IEEE Tencon, India (2003).

17 Privitera C M and Stark L W: 'Algorithms for defining visual regions of interest: comparison with eye fixations', IEEE Transactions on Pattern Analysis and Machine Intelligence, 22, No 9, pp 970—982 (2000).

18 Jaimes A, Pelz J B, Grabowski T, Babcock J and Chang S-F: 'Using human observers' eye movements in automatic image classifiers', Proceedings of SPIE Human Vision and Electronic Imaging VI, San Jose, CA (2001).

19 Pomplun M and Ritter H: 'A three-level model of comparative visual search', in Hahn M and Stoness S C (Eds): 'Proceedings of the Twenty First Annual Conference', Cognitive Science Society, pp 543—548 (1999).

20 Salvucci D D and Goldberg J H: 'Identifying fixations and saccades in eye-tracking protocols', Proceedings of the Eye Tracking Research and Applications Symposium, pp 71—78, New York, ACM Press (2000).

21 Salvucci D D: 'An interactive model-based environment for eye-movement protocol analysis and visualization', Proceedings of the Symposium on Eye Tracking Research and Applications, pp 57—63, Palm Beach Gardens, Florida, USA (2000).

22 Stone L and Beutter B et al: 'Models of tracking and search eye-movement behaviour', NASA (2000).

23 Duchowski A T, Medlin E, Cournia N, Murphy H, Gramopadhye A, Nair S, Vorah J and Melloy B: '3D eye movement analysis', Behavior Research Methods, Instruments and Computers (BRMIC), 34, No 4, pp 573—591 (2002).

24 Schnell T and Wu T: 'Applying eye tracking as alternative approach for activation of controls and functions in aircraft', Proceedings of the 5th International Conference on Human Interaction with Complex Systems (HICS), Urbana, Illinois, USA, p 113 (2000).

25 Ward D J and MacKay D J C: 'Fast hands-free writing by gaze direction', Nature, 418, p 838 (2002).

26 Nikolov S G, Bull D R and Gilchrist I D: 'Gaze-contingent multi-modality displays of multi-layered geographical maps', Proc. of the 5th Intl Conf on Numerical Methods and Applications (NM&A02), Symposium on Numerical Methods for Sensor Data Processing, Borovetz, Bulgaria (2002).

27 Roto V: 'Search on mobile phones', Nokia Research Centre (2002).

28 Fan X, Xie X, Ma W-Y, Zhang H-J and Zhou H-Q: 'Visual attention based image browsing on mobile devices', IEEE International Conference on Multimedia and Expo, I, pp 53—56, Baltimore, MD, USA (2003).

29 Kerne A: 'CollageMachine: an interactive agent of Web recombination', Leonardo, 33, No 5, pp 347—350 (2000).

30 Farid M, Murtagh F and Starck J L: 'Computer display control and interaction using eye-gaze', Journal of the Society for Information Display, 10, No 3, pp 289—29 (2002).

31 Oyekoya O K and Stentiford F W M: 'Exploring human eye behaviour using a model of visual attention', International Conference on Pattern Recognition, Cambridge UK (2004).

32 Stentiford F W M: 'An estimator for visual attention through competitive novelty with application to image compression', Picture Coding Symposium, Seoul (2001).

33 Itti L and Koch C: 'A saliency-based search mechanism for overt and covert shifts of visual attention', Vision Research, 40, No 10—12, pp 1489—1506 (2000).

Oyewole Oyekoya is undertaking a PhD research programme at University College London working with the Content Understanding Group, based at BT's Adastral Park.

He was awarded a scholarship and graduated in June 2001 with a BSc (Hons) in Computer Science at Kingston University.

He undertook his work placement at Nortel Networks, gaining valuable experience in the telecommunications industry and won company sponsorship to produce an interface for test automation as his final dissertation.

He is a Faraday Associate sponsored by the Imaging Faraday Partnership and BT Exact. He is now researching eye tracking for improved image search and retrieval.

He will shortly be presenting his research at the forthcoming International Conference on Pattern Recognition in Cambridge.

Fred Stentiford won a scholarship to study Mathematics at St Catharine's College, Cambridge, and obtained a PhD in Pattern Recognition at Southampton University. He first joined the Plessey Company to work on various applications including the recognition of fingerprints and patterns in time varying magnetic fields. He then joined BT and carried out research on optical character recognition and speech recognition. From 1983 he led a team developing systems employing pattern recognition methods for the machine translation of text and speech. This work led to the world's first demonstration of automatic translation of speech between different languages. He led research into the design of new dialogues for telephone services and managed the government funded collaborative Dialogues 2000 project which aimed to promote common standards in the spoken user interface in UK industry. He then returned to vision research and led a group developing new systems for analysing and delivering multimedia content over the telecommunications networks. He now holds a chair in Telecommunications with UCL and leads a team at the Adastral Park Campus researching new technologies for understanding visual content, in close collaboration with BT's Broadband Applications Research Centre.

He is a corporate member of the IEE and the BCS and has published over 50 papers and filed 15 patents on pattern recognition techniques.