Searching for threat: factors determining performance during CCTV monitoring

Christina J. Howard[1], Tomasz Troscianko[1], Iain D. Gilchrist[1], Ardhendu Behera[2], & David C. Hogg[2]
[1]University of Bristol
[2]University of Leeds
UK

Abstract

Monitoring closed-circuit television (CCTV) for security purposes is a task requiring sustained attention and the processing of many complex, constantly changing visual elements. Studies of performance in such tasks reveal a high level of workload and rapid loss of performance as workload is increased. Similarly, laboratory based experimental paradigms suggest that performance in CCTV monitoring is extremely dependent on the complexity and number of video screens monitored. We suggest that measuring eye movements during CCTV monitoring might provide a novel and rich source of data to illuminate the question of how CCTV monitoring is performed. In the psychological literature, two influences on attention are traditionally considered: the ability of events in the world to capture our attention regardless of our current goals (stimulus-dependent salience) and the ability to direct our attention towards stimuli relevant to the task we are trying to perform (goal-based relevance). Stimulus-dependent salience and goal-based relevance together determine the human fixation priority assigned to scene locations (Fecteau and Munoz, 2006, TICS 10, 382-390). Tests of the stimulus-dependent salience component of this process tend to look for regions in the image that are consistently fixated and link this to the underlying image properties. However, when the task is common across observers, consistent fixation location can also indicate that that region has high goal-based relevance. By examining the eye movements of multiple expert observers, we may start to characterise features of the moving video stimulus that are predictive of events likely to be judged as suspicious.

Eye movements in the real world

Studies of naturalistic task performance have used eye movements as a measure of how attention moves around the visual field, or 'attentional deployment' (e.g. Hayhoe & Ballard, 2005; Land, 1999; Underwood, Chapman, Brocklehurst, Underwood, & Crundall, 2003; Findlay & Gilchrist, 2003). The way people use eye movements to do various everyday tasks has been investigated; including steering a racing car (Land & Tatler, 2001), hitting a cricket ball (Land & MacLeod, 2000), making tea (Land, Mennie, & Rusted, 1999) and sandwiches (Hayhoe, 2000). The

In D. de Waard, J. Godthelp, F.L. Kooi, and K.A. Brookhuis (Eds.) (2009). *Human Factors, Security and Safety* (pp. 1 - 7). Maastricht, the Netherlands: Shaker Publishing.

pattern of eye movements recorded suggest that the eyes do not perform a pure 'forward-planning' role for actions, nor do they wait for actions to occur before fixating relevant parts of a scene. Instead they provide visual information 'just-in-time' for action and lead actions by between around half a second to a second (Land & Hayhoe, 2001). In addition, where observers look in a scene is not determined solely by saliency derived from low level features of the kind described by Itti and Koch in their model (Itti & Koch, 2000). Instead, fixation 'priority' is instead determined by the dual influences of stimulus-dependent low-level salience and goal-based relevance i.e. whether the stimulus relates directly to the task at hand (Fecteau & Munoz, 2006). In the current programme of work we measure eye movements to investigate such attentional deployment for closed-circuit television (CCTV) monitoring. We hypothesise that eye movements are likely lead responses by a buffer of the approximate magnitude suggested by Land and his colleagues (e.g. Land & Hayhoe, 2001). It remains an open question, however, whether the level of cognition required for a task such as CCTV might extend the length of this temporal buffer while cognitive processing of fixated events takes place.

Closed-Circuit Television (CCTV)

In CCTV control rooms, the principal duty of operators is to monitor a bank of screens through which about 50-60 images selected from up to 600 cameras are streamed, with the task of anticipating and detecting critical events. Each operator also has a desktop screen (spot monitor) through which, when something from the wall-mounted screens attracts attention, they can select chosen video images for display on their spot monitor and over-ride the pre-set programme. Despite the complexity of this task, the prediction of incidents by the operators appears to be possible (Troscianko et al. 2004). The main difficulty in this task is the degree of cognitive and visual overload both in terms of monitoring what may be complex and constantly changing scenes, and in terms of allocating attention effectively to many screens at once. Despite the familiarity of operators with the scenes and likely occurrences presented to them, as well as their ability to group images in terms of their geographical locations, this is still clearly a very demanding task. The paucity of research, and therefore, guidance on this topic has long been a subject of concern to CCTV operators and managers, and calls for an investigation into this issue are still being made by the CCTV surveillance community (e.g. Donald, 2005). Tickner and Poulton (1973) found that typical levels of perceptual load are greater than the average individual can process effectively. They had observers monitor 4, 9 or 16 screens, and recorded detection rates of 83%, 84% and 64%, respectively. More recently in a survey, 82% of operators reported feeling able to monitor a maximum of 16 cameras or fewer, with more than half of operators reporting a maximum of as little as one to four cameras (Wallace, Diffley & Aldridge, 1997). Clearly the task is a demanding one even for trained operators, and a deeper understanding of the way in which CCTV monitoring is performed may help us to design and implement systems that maximise human monitoring performance, bearing in mind that the number and type of camera images will depend on the type of surveillance task.

Figure 1. Scenes similar to those typically captured by CCTV. Even in this relatively sparse display of only four static scenes, cognitive and perceptual load is clearly high, and may result in missed incidents or late detection. Colour images can be viewed at http://extras.hfes-europe.org

Laboratory tasks relevant to CCTV monitoring

The literature on theoretically-driven tasks in the laboratory predicts that CCTV will place a heavy load on the perceptual and cognitive systems, with performance decreasing rapidly as the complexity and number of video streams is increased. This is certainly in accordance with the reports of CCTV operators and goes some way to explaining the nature of the task faced during CCTV monitoring.

From the laboratory, there is a wealth of evidence pointing towards the detrimental effects of load of performance. In visual search tasks, the speed with which search targets are detected is typically dependent on the number of objects in the search array. Although search for some features appears to exhibit 'pop-out', or to display relatively little effect of set size on reaction time (such as a target defined by its unique colour in the display), search times for targets defined by conjunctions of features, or search for an object defined by the presence of both of two or more features (e.g. a red, rightward-tilted object), exhibits large decrements with set size (Treisman & Gelade, 1980;Treisman, Sykes & Gelade, 1977). This is of particular relevance to CCTV monitoring tasks since even in a single CCTV monitor, there are many objects, people and environmental features all competing for attention. The target, i.e. potentially suspicious behaviour, appears likely to be defined by a conjunction of features such as suspicious patterns of movement, location in the scene and physical appearance. Furthermore, search for disjunctions of features (e.g. a red object or a rightward-tilted object) has been shown to cause a performance cost

compared to search for a single feature in applied contexts. Manneer, Barrett, Phillips, Donnelly and Cave (2007) showed that for airport security screening, searching for two potential target types frequently caused a loss in accuracy, and in some cases an increase in search time, compared to searching for a single object type. Clearly, CCTV monitoring is likely to involve some similarity both to conjunction and disjunction searches, since there can be no single prototypical search target. Rather, operators must search simultaneously for many potentially criminal events, as well as accidents or other incidents requiring intervention.

Another field of laboratory work is particularly relevant to the task of CCTV monitoring, namely the multiple object tracking (MOT) task and its variants. The MOT task is reminiscent of the traditional fairground 'shells' game, whereby a shell is hidden underneath one of a number of visually identical cups. The cups are then shuffled in a manner requiring close and sustained attention. The observer must keep track of the cup that contains the shell to win a prize. In a typical MOT task, a number of identical moving objects are presented on a screen, some of which are designated as targets for tracking. Observers must attempt to keep track of the target objects and to continue to distinguish them from non-targets, so that at the end of the trial, when queried, they may say whether a probed object was a target or not. Typically in these tasks, a capacity limit of around four objects is reported (e.g. Pylyshyn & Storm, 1988) although more recently a flexible limit has been proposed that depends on objects' speed (Alvarez & Franconeri, 2007) and the degree of precision required (Franconeri, Alvarez & Enns, 2007; Howard & Holcombe, 2008). These results are likely to be applicable to the applied task of CCTV monitoring since the motion of many individuals around one or many screens is likely to be a powerful cue to their potential level of perceived suspiciousness. Many factors which are extremely prevalent in CCTV images have been shown to increase the difficulty of MOT tasks. For instance, performance decreases if objects change their shapes in a non-rigid fashion (vanMarle & Scholl, 2003) as real people would do as their form appears on the screen, or if they pass behind occluders (Scholl & Pylyshyn, 1999) such as would be expected from people moving behind each other, behind trees or buildings, for instance.

A third field directly pertinent to CCTV monitoring is that of divided attention or 'dual' tasks. In such experiments, observers are required to perform two tasks at once, and this has been repeatedly shown to cause a marked decrement in performance in both tasks (e.g. Pashler, 1995; Schneider & Shiffrin, 1977). It seems likely that these results apply at least on occasions to CCTV monitoring, since operators may need to make a cognitive or perceptual judgement about activity in two or more areas of screens simultaneously. Indeed they may also need to continue to monitor many events whilst making decisions about the level of appropriate action required.

Future directions

We believe that measuring eye movements will provide a new window into our understanding of how CCTV monitoring is performed. This is likely to have implications for the ways in which CCTV monitoring is managed and carried out, as

well as for control room systems design and for training programmes. The method can be used to study systematically how the number, arrangement and content of screens affect performance. We hope that this method will also create a new synergy between laboratory-based models and their applications.

The results will help to test and constrain laboratory-based models of human performance in CCTV monitoring and in similar tasks. Most laboratory studies use simplified and carefully controlled tasks and stimuli in order to make claims about the underlying perceptual processes involved. The advantage of these methods is of course the ease with which conclusions can be drawn about perceptual processes. Testing in applied scenarios, however, is likely to help generalise these results to tasks that are more relevant to the real world with its inherent complexity. In particular, the visual complexity of CCTV images, and their dynamic nature are likely to influence performance in a manner that is very different from the static, simple stimuli most frequently used in the laboratory.

In addition, we hope that measuring eye movements during monitoring will provide an invaluable new method for studying responses to video stimuli drawn from real-world scenarios outside of CCTV monitoring. Eye movement data will provide a novel and rich source of data for a variety of applied tasks. There are many cases in which a moving image must be consistently attended, and this method could be applied to any scenario, perhaps relating to other fields of safety and security such as air traffic control, or relating to navigation in the real world, to sports behaviour or to social perception.

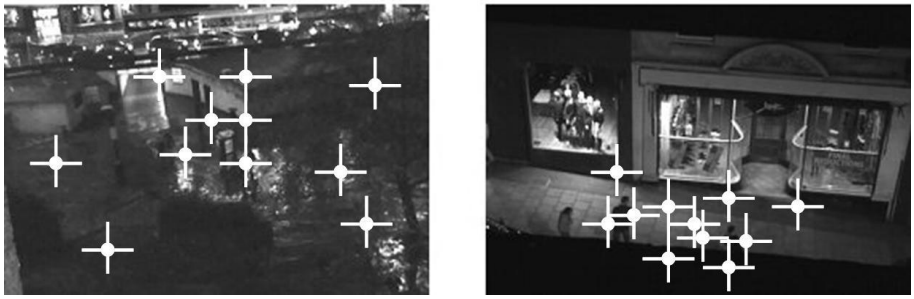

Figure 2. Sample fixation data shows the position of the eyes for eleven observers at a single point in time. The left-hand panel shows a time at which the between-observer consistency is low. The right-hand panel shows a time at which between-observer fixations are highly consistent, indicating an event of significant interest for observers. Colour images can be viewed at http://extras.hfes-europe.org

We plan to use this human eye movement data in future to inform and constrain a computer leaning algorithm, and to work towards automatic detection of elements of CCTV footage that are likely to be judged as potentially suspicious by human observers. We will study two distinct forms of the CCTV viewing situation. Firstly, to establish a ‘ground truth’, we will study the decisions of expert observers when looking at a single screen in a rested state, with no competing visual distracters. Secondly, we will study a realistic control-room situation in which there is a bank of

multiple screens and a high perceptual and cognitive load. In this multiple screen scenario, we will be able to study the interactive effects of multiple screens on eye movements and attention, as well as strategies used by operators to spread their attention over multiple screens. Eye movement data from the single screen scenario will be used to extract features from the moving video image which are likely to be judged as suspicious. Eye movements are likely to precede behavioural responses and are therefore well placed to best inform artificial detection processes. In addition, when eye movement data is combined across different observers (see Figure 2) and with behavioural responses, it provides a measure of the direction of goal-directed attention.

Acknowledgements

This work is supported by an EPSRC Cognitive Systems Foresight grant. Many thanks to Kate Rennicks, David Walsh and Ian Townley at Manchester City Council CCTV Control Room for their invaluable contributions to this project.

References

Alvarez, G.A. & Franconeri, S.L. (2007). How many objects can you track? Evidence for a resource-limited attentive tracking mechanism. *Journal of Vision, 7*(13), 1-10.

Donald, C. (2005) How many monitors should a CCTV operator view? In *CCTV Image*. STL Publishing, London, England, pp. 355.

Fecteau, J.H. & Munoz, D.P. (2006) Salience, relevance, and firing: a priority map for target selection. *TRENDS in Cognitive Sciences*, *10*, 382-390.

Franconeri, S.L., Alvarez, G.A., & Enns, J.T. (2007). How many locations can be selected at once? *Journal of Experimental Psychology: Human Perception and Performance, 33*, 1003-1012.

Findlay, J.M. & Gilchrist, I.D. (2003) *Active vision: The psychology of looking and seeing.* Oxford: Oxford University Press.

Hayhoe, M. (2000) Vision using routines: A functional account of vision. *Visual Cognition*, *7*, 43-64.

Hayhoe, M. & Ballard, D. (2005) Eye movements in natural behavior, *TRENDS in Cognitive Science*, *9*, 188-194.

Howard, C.J. & Holcombe, A.O. (2008) Tracking the changing features of multiple objects: Progressively poorer perceptual precision and progressively greater perceptual lag. *Vision Research*, *48*, 1164-1180.

Itti, L. & Koch, C. (2000) A saliency-based search mechanism for overt and covert shifts of attention. *Vision Research*, *40*, 1489-1506.

Land, M. (1999) Motion and vision: why animals move their eyes. *Journal of Comparative Physiology*, *185*, 341-352.

Land, M.F. & Tatler, B.W. (2001) Steering with the head: The visual strategy of a racing car driver. *Current Biology*, *11*, 1215-1220.

Land, M. & Furneaux, S. (1997) The knowledge base of the oculomotor system. *Phil. Trans. R. Soc. Lond. B., 352*, 1231 – 1239.

Land, M.F. & Hayhoe, M. (2001) In what ways do eye movements contribute to everyday activities? *Vision Research*, *41*, 3559-3565.

Land, M.F. & MacLeod, P. (2000) From eye movements to actions: how batsmen hit the ball. *Nature Neuroscience*, *3*, 1340-1345.

Land, M., Mennie, N. & Rusted, J. (1999) The roles of vision and eye movements in the control of activities of daily living. *Perception*, *28*, 1311-1328.

Menneer, T., Barrett, D. J. K., Phillips, L., Donnelly N. & Cave, K.R. (2007) Costs in searching for two targets: dividing search across target types could improve airport security screening. *Applied Cognitive Psychology*, *21*, 915-932.

Pashler, H. (1995). Divided visual attention. In S. Kosslyn (Ed.) *Visual Cognition: Invitation to Cognitive Science* (pp. 71-100). Cambridge, MA, USA: MIT Press.

Pylyshyn, Z.W., Storm, R.W. (1988). Tracking multiple independent targets: evidence for a parallel tracking mechanism. *Spatial Vision*, *3*(3), 1-19.

Schneider, W. & Shiffrin, R.M. (1977). Controlled and automatic human information processing: 1. Detection, search, and attention. *Psychological Review*, *84*, 1-66.

Scholl, B.J. & Pylyshyn, Z.W. (1999). Tracking multiple items through occlusion: clues to visual objecthood. *Cognitive Psychology*, *38*, 259-290.

Tickner, A.H. & Poulton, E.C. (1973) Monitoring up to 16 synthetic television pictures showing a great deal of movement. *Ergonomics*, *16*, 381-401.

Treisman, A.M. & Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology*, *12*, 97-136.

Treisman, A., Sykes, M. & Gelade, G. 1977. Selective attention and stimulus integration. In *Attention and performance VI*, S. Dornic (Ed.), Hillsdale, NJ, USA: Lawrence Erlbaum, pp. 333-361.

Troscianko, T., Holmes, A., Stillman, J., Mirmehdi, M., Wright, D., & Wilson, A. (2004). What happens next? The predictability of natural behaviour viewed through CCTV cameras. *Perception*, *33*, 87-101.

Underwood, G., Chapman, P., Brocklehurst, N., Underwood J., & Crundall, D. (2003) Visual attention while driving: sequence of eye fixations made by experienced and novice drivers. *Ergonomics*, *46*, 629-646.

vanMarle, K. & Scholl, B. J. (2003). Attentive tracking of objects versus substances. *Psychological Science*, *14*, 498-504.

Wallace, E., Diffley, C., & Aldridge, J. (1997). Ergonomic Design Considerations for Public Area CCTV Safety and Security Applications, *International Ergonomics Association Congress*, July 1997.