Visual Memory Augmentation: Using Eye Gaze as an Attention Filter

Deb Roy, Yair Ghitza, Jeff Bartelma, and Charlie Kehoe
Cognitive Machines Group, MIT Media Laboratory
www.media.mit.edu/cogmac

Abstract

We present two early prototype systems that leverage computer memory to augment human memory in everyday situations. Both experiments investigate the role of eye tracking as a way to detect a person's attention and use this knowledge to affect short and long term memory processes. This work is part of a larger effort underway at the MIT Media Laboratory to develop systems that work symbiotically with humans, leading to increased performance along numerous cognitive and physical dimensions [9].

1. Eye Movements: A Window into Attention

Where we look depends on why we are looking. We consciously make eye contact with others when we want to engage in social interaction and look away otherwise. Less obviously, and often subconsciously, our rapid probing eye movements are influenced by a combination of the "bottom-up" structure of the visual environment and our "top-down" active goals.

Since the classic studies of Yarbus [10], it has been well known that a person's goals significantly influence their eye gaze trajectory. For example, while looking at a picture of a group of people, if a subject is asked "How old are the people in this photograph" versus "How wealthy are the people in this photograph", different goal-dependent gaze patterns will be observed. Based on such findings, we that suggest recording and analyzing eye movements on a wearable computer might lead to interesting data about the user's cognitive and motivational state.

A second line of cognitive experiments that is relevant to Experiment II involves *change blindness* [cf. 8]. Figure 1 shows an example of visual stimuli used to investigate change blindness. Using the *flicker paradigm*, one image is displayed for approximately 240 milliseconds, followed by a blank screen for 40-80ms, then the second image for 400ms, then another 40-80ms of blank screen. This pattern is then cycled until the subject reports that they have detected the difference between images. Some changes are easier to detect than others. Roughly, changes which don't "matter", i.e., which are not central to the main gist of the image, are difficult to detect. An interpretation of such experiments is that even in the case of extremely short term memory encoding (i.e., the time between images in the display), people only remember aspects of a scene that they judge (perhaps subconsciously) to be important or relevant to the situation at hand. This finding also suggests that knowing where a person is looking provides valuable meta-data for annotating visual memories -- a memory augmentation device might only record visual data that the user cares about.

Figure 1: Stimulus used in change blindness experiments (Rensink, O'Regan, and Clark, 1997, images used with permission of authors). When these images are shown using the flicker paradigm (see text for description), subjects often take over a minute to detect the rather large difference between the images (the missing engine).

2. Experiment I: Using Eye-Tracking to Augment Memory for Visual Search

The first experiment is motivated by difficulties people often encounter during visual search such as looking for keys. We often find ourselves searching the same places repeatedly rather than systematically trying new places.

To address this shortcoming in visual search strategy, we have developed a prototype system that tracks where a person looks and provides an extended memory of their search history, guiding the subject towards unsearched portions of a scene. The system has no information about where or what the target is. Target detection is left to the human. The system, instead, augments the human's

memory, keeping track of where the person has already looked.

To test this idea, we worked with a popular children's game, *Where's Waldo* [2], which plays off of people's inability to search effectively. The goal of this game is to find a hand-drawn character named Waldo. Waldo always wears the same colored and textured clothes, but is embedded in scenes full of visual distracters which share elements of his clothing. The result is often an extremely challenging visual search problem.

The memory aid we have developed uses a head-mounted eye tracker (I-Scan Model ETL-500) which includes two miniature video cameras mounted on a headband. The first camera, the *eye camera*, is directed towards the user's right eye. Using infrared illumination, it provides reliable video of the subject's eye movements which are converted into point-of-regard (POR) information by the eye-tracker's firmware. A second *scene camera* is also mounted on the headband. It points outwards to capture the view of the environment as seen from the subject's perspective.

The eye tracker generates x-y coordinates of the subject's visual POR at a rate of 24 samples per second. We explain below how this data is used to feedback visual information to the subject that conveys a history of where in the scene they have already looked. Figure 2 shows the system in use. An image of the game is displayed on a 21" LCD monitor. The subject, wearing the eye tracker, has been searching for approximately 30 seconds. The "burn marks" on the display indicate areas where the subject has fixated for significant periods of time. The assumption underlying the design of the system is that if a subject spends sufficient time looking at a particular region of the image without finding Waldo, it is safe to darken this part of the image since it is unlikely to contain the target. This assumption is of course not always true, but serves as a starting point.

To align the eye tracker data with the image, we implemented a tracking algorithm that finds the LCD monitor image within the video captured by the scene camera. Detection of the location of the monitor image is achieved by searching for the distinctive black border of the monitor. Since the subject's head is in constant natural motion, the location of the monitor within the scene camera's view also shifts. By tracking the location of the monitor, the system dynamically aligns eye tracker data to the original image on the LCD display, enabling precise selective darkening of the image.

Darkening occurs after fixation around the same point on the monitor for 250 ms, where fixation is defined as remaining within a 75-pixel radius. The fixated point is darkened to 25% of the original brightness when the subject looks away from the point.

Based on informal evaluations, the approach appears promising. Some subjects initially felt disconcerted to see parts of their visual field disappear as they fixated on locations, but tended to adapt quickly to the sensation. In trials on several pilot subjects, we found that the feedback mechanism operates as designed, eliminating parts of the scene that do not contain the target, thus guiding the user to the target. We acknowledge that formal evaluations will be necessary before any conclusions can be drawn regarding the effectiveness of the system for search. The parameters for controlling burn rate of the image were set by hand and need to be iteratively tuned as the task is better understood.

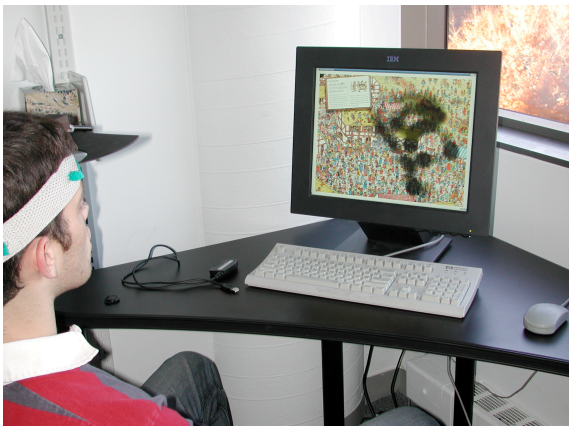

Figure 2: The Find Waldo visual STM augmentation system.

3. Experiment II: Using Eye-Tracking to Filter Personal Visual Diaries

With recent advances in video compression, availability of large memory stores, and cheap, low power cameras, it is practical to consider "always-on" first-person video diaries which capture a person's lifetime experiences, from their point of view (Microsoft Research recently announced its SenseCam project along these lines, although no written account of the work has yet been published; see also [6]). Similar ideas in the audio domain [cf. 7] are more advanced, perhaps due to the much smaller memory footprint of audio compared to video, and more mature tools for speech and audio processing compared to unconstrained video processing.

A key issue for making use of a video diary is effective summarization, indexing, and search of content. Numerous video processing algorithms have been developed for this purpose including detection of scene changes, classification of scene types, detection of faces, and so forth. All of these techniques are based on the content of the video. In contrast, we have explored the use of eye tracking data to find video of interest (in

related work, Healey and Picard [3] used skin conductivity of the user to predict video segments of interest). Based on the visual attention and change blindness experiments discussed earlier, we know that eye gaze provides valuable clues about which parts of a visual experience are found meaningful by the user. Thus, eye gaze data can be used as a salience filter to select portions of video that are likely to provide effective summaries of content and likely targets for later search and retrieval.

We used the same eye tracking hardware described in Experiment I. Although the current configuration of hardware (head worn eye tracker tethered to a workstation) is not immediately suitable for mobile wearable operation, we envision future versions of the hardware in which the eye and scene camera are further miniaturized and integrated into a wearable device resembling a pair of ordinary eye glasses.

The prototype system consists of three main components: an eye fixation detector which predicts salient portions of video in space-time, a hierarchical segmentation algorithm which generates hypotheses of regions of interest in the visual scene based on image analysis, and a visual diary interface for accessing the contents of a video recording.

Fixation Detection

When a subject looks steadily at a particular region of the screen for at least 2 seconds, a "snapshot" of the video stream is taken and passed to the hierarchical image segmentation system. In our system, fixation on a region is defined as staying within a 5-pixel radius on a 640x320 video stream. We acknowledge that a 2-second fixation period is overly restrictive during natural activity [4] -- we started with this setting for the practical concern of reducing the amount of data that is processed by later stages of the system.

Hierarchical Segmentation

Given a fixation point within an image, a hierarchical segmentation algorithm is used to hypothesize salient regions of the image which correspond to the object of interest within the larger visual context. Hierarchical image segmentation proceeds in two stages based on the approach described by Kropatsch and Haxhimusa [5]. The first stage performs connected component analysis based on saturation and value channels in HSV space. For every pixel in the image, the two-channel values are stored in a color map. The color map is scanned to find ten representative colors. The representative colors are added incrementally and are chosen based on maximum distance from each previously found representative color. Each pixel is then set to its nearest neighbor representative color, based on its original color and proximity in SV space. Finally, connected component analysis is performed to group locally adjacent regions of the same color.

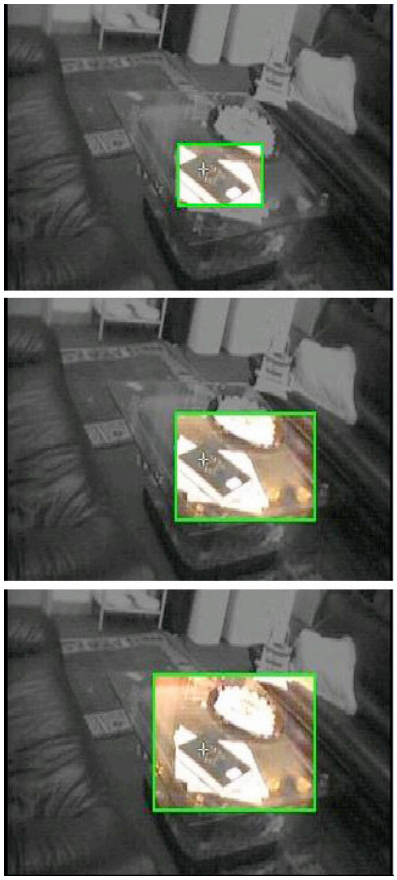

Figure 3: Hierarchical segmentation provides three hypothesized regions of interest containing a fixation point. The hypothesis shown in the topmost image is most appropriate for capturing the single book on the coffee table whereas the third image better captures the entire contents of the table.

The second stage of processing performs graph contraction of the previously found connected components. Each component is treated as a node in a graph, with edges connecting adjacent regions in the graph. For a specified number of iterations, locally adjacent nodes are grouped based on distance in SV color space. Grouped nodes are then contracted to form one node while preserving the adjacency structure of the graph. This iterative approach allows for multiple levels of segmentation. In our current implementation, graph contraction is performed for 3 iterations, providing three levels of segmentation. A sample run of the algorithm is shown in Figure 3.

Visual Diary Interface

A user browses the images comprising the attention-filtered visual memory using the interface shown in

Figure 4. Each frame grabbed by the fixation detector corresponds to a drop-down menu in the interface which contains alternate segmentations of the image. To indicate which segmentation best reflects his/her true region of interest – and hence which image should be saved – the user simply selects an image from the drop-down menu.

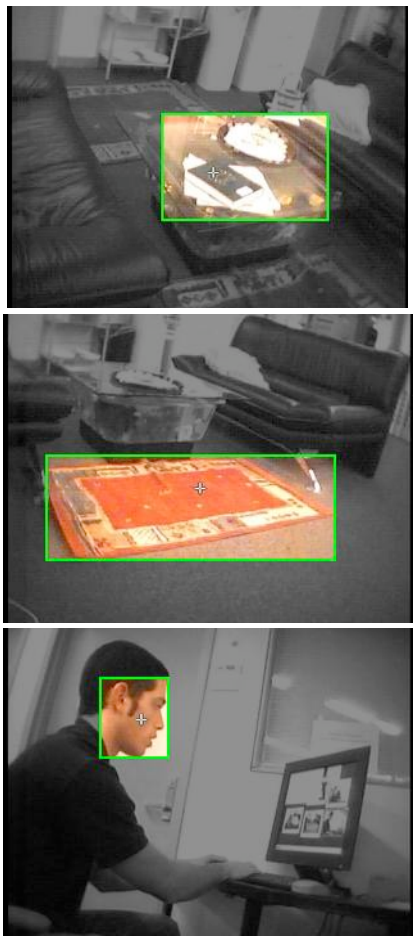

Figure 4: Three frames extracted automatically from an autobiographical video sequence.

The interface displays three video images at a time. The user can scroll through this diary of visual memories by pressing buttons at the left and right sides of the interface, respectively. Thus in essence the interface presents the attentional highlights of a user's raw visual experience as a chronological slide show.

4. Conclusions

We have presented two early experimental prototypes which investigate the role of eye gaze as a form of meta-data for encoding "first-person" video data. The first system addresses the "Find Waldo" problem, serving as an extended visual search memory that keeps track of where a person has already searched and visually guides them to look in new places. The second system analyzes fixation patterns of a person during everyday activities to predict segments of autobiographical video data that is likely to be meaningful to the user, serving as a salience filter for later video indexing and search.

The work presented here is at an early stage. More extensive experimentation will be required to move beyond various hand-coded system parameters, and proper evaluation will be required before any specific claims can be made about the effectiveness of the specific implementations that we have presented. We have recently completed a more detailed evaluation of a related system which combines intentional eye-gaze traces with image segmentation [1], demonstrating the practical potential of the overall approach. In summary, we believe that the idea of using eye gaze to augment wearable video recording is novel, and based on our initial results, worthy of further investigation.

5. References

[1] Bartelma, J., and D. Roy. (In review). Flycatcher: Fusion of Game with Hierarchical Image Segmentation for Robust Detection.

[2] Handford, M. (1997). Where's Waldo? Candlewish Press.

[3] Healey, J. and R. Picard. (1998). StartleCam: A Cybernetic Wearable Camera. Proc.. of the Second Int. Symposium on Wearable Computers (ISWC'98).

[4] Henderson, J. M. (2003). "Human gaze control in real-world scene perception," Trends in Cognitive Sciences, 7, 498-504

[5] Kropatsch, W.G. and Haxhimusa, Y (2004). Grouping and Segmentation in a Hierarchy of Graphs. Proceeding of the 16th IS&T/SPIE Annual Symposium.

[6] Lamming, M., Brown, P., Carter, K., Eldridge, M., Flynn, M., Louie, G., Robinson, P., & Sellen, A. (1994). Computer Journal, 37(3), 153-163.

[7] Lin, W. and A.G. Hauptmann. (2002). A wearable digital library of personal conversations. Proceedings of the Joint Conference on Digital Libraries, pp. 277-278.

[8] Rensink R.A., J.K. O'Regan , and J.J. Clark. (1997). To See or Not to See: The Need for Attention to Perceive Changes in Scenes. Psychological Science, 8:368-373.

[9] Roy, D. (In press). 10x: Human-Machine Symbiosis. BT Technology Journal.

[10] Yarbus, A.L. (1967). Eye Movements and Vision, translated by Basil Haigh, Plenum Press, NY.