An Interface for Visual Inspection based on Image Segmentation

James Mardell
Mark Witkowski
Robert Spence
Department of Electrical and Electronic Engineering
Imperial College London
London SW7 2BT
{james.mardell, m.witkowski, r.spence}@imperial.ac.uk

ABSTRACT
If a person is lost in the wilderness it is increasingly normal for the area in question to be over-flown by an Unmanned Aerial Vehicle (UAV) whose on-board video camera transmits a view of the terrain below. It is then the task of a human operator to visually inspect that view by means of a visual interface specifically designed to enhance the likelihood of the missing person being located.

We investigate a novel approach to the visual inspection of the terrain image: that of presenting small segments of that image for very short periods of time, though commensurate with the speed at which the UAV flies.

Participants took part in an investigation in which the challenging task was to identify the presence, in typical terrain images, of human beings. Six representative terrain maps were involved, and the six degrees of segmentation explored were such as to provide individual terrain image viewing times between 3.9 s and 108 ms.

We report the result of investigating the proposed segmentation approach to visual inspection in the demanding and realistic context of Wilderness Search and Rescue. Our investigation reveals a clear and distinctive change of visual inspection performance through appropriate visual interface design can have important consequences.

Specifically, we investigate whether it is better to search for the targets within the context of a large image area for a correspondingly shorter time, or to look instead for those targets within smaller segments of the same image area, but for a correspondingly shorter time. We adopt a laboratory-based approach, using simulated (but photo-realistic) search targets embedded in actual aerial terrain imagery taken from the Montana State Library Natural Resource Information System (http://nris.mt.gov/gis/).

Figure 1 shows how this trade-off between image area and presentation time is achieved within the interface design. Images from six different fly-past terrain areas are each divided into six different segmentation degrees. At the top is a terrain image of one complete strip of land to be searched for potential targets (corresponding to a swept ground area of $2341 \times 117$ m, at a ground resolution equivalent to six inches (15 cm) per pixel). The segmentation degrees 1–6 successively divide the total image area into smaller image tiles, each shown at a correspondingly faster rate. A number of visual targets are randomly embedded in each terrain strip.

The key question is how many of these targets are correctly identified at each segmentation level/speed combination: are some combinations more effective than others?

1. INTRODUCTION

If someone is lost in extensive and inaccessible terrain such as a mountainous or sparsely populated region, one approach to the search for that person is to fly an Unmanned Aerial Vehicle (UAV) over the area in which the person may be found, and arrange for images taken by a camera on the UAV to be transmitted to a base station where they can be interpreted by trained human operators. This activity is generally referred to as Wilderness Search and Rescue (WiSAR).

The question arises as to what is the best way to present those images to the human operator in order to maximize the likelihood that the missing person or relevant evidence may be identified. Self evidently, even small improvements in identification performance through appropriate visual interface design can have important consequences.

The WiSAR motivation is to locate a missing person within the limits of a plausible terrain search area [7]. Where large terrain areas must be searched, aerial survey represents the normal, and in some cases the only feasible, approach. The primary measure for success in this task is the correct identification of these visual rescue “targets”, with a secondary criteria of minimizing false identifications which might lead to the wasteful misdirection of ground based resources.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

AVI '12, May 21–25, 2012, Capri Island, Italy
Copyright 2012 ACM 978-1-4503-1287-5/12/05 ...$10.00.
3. THE WISAR TASK

We consider an idealized abstraction of the WiSAR task suitable for laboratory investigation. We assume a downward facing camera free of various flight artefacts such as vibration and banking effects. We have also assumed that the UAV flies at a height commensurate with the resolution of the imaging device, to give adequate detail to secure target identification. These assumptions are consistent with currently available UAVs such as the IAI Heron (http://www.iai.co.il/) or the General Atomics Predator (http://www.ga-asi.com/).

Following launch the UAV performs a search pattern of the designated area [3, 7]. The search pattern adopted depends on the circumstances of the emergency. Where a “last known position” can be estimated, a spiral or box search may be employed, starting at that position and working outwards. Where a known flight path is available, a search parallel to that path may be adopted in the first instance.

4. EXPERIMENTAL DESIGN

To address the research questions outlined in section 2, we prepared six “over-flight” images strips of terrain (referred to as “A”–“F”) representing various types of “wilderness areas”. An example (strip “D”) is shown in figure 1. Each of the six strips was then prepared by being segmented into each of the six degrees (a total of 8190 images) and software prepared to present the sequences at the required presentation rates.

The experimental design consists of showing each volunteer participant each of the six strips, each at a different segmentation level-speed combination, 1–6. Each participant therefore observed every image strip only once, and saw each of the six segmentation degrees only once. In order to minimize the potential for learning effects and interaction between image and segmentation, combinations of image and segmentation degree were presented in random order and combination according to a conventional balanced block design. The experimental design repeats after six participants.

Each of the six strips of terrain to be searched for potential targets corresponds to a swept ground area of 2341 × 117 m, or 15360 × 768 px at a ground resolution of six inches per pixel. For segmentation degree 1, figure 1, top left tile, the strip is divided into 15 non-overlapping images (at the full 1024 × 768 px screen resolution), each presented for 3878 ms. At the other end of the scale, segmentation degree 6, figure 1, bottom right tile, the total strip is divided into 540 (15 × 6 × 6) much smaller images, but these are shown for only 108 ms (171 × 128 px at screen resolution). Any unused screen area was set to a neutral gray. All the intermediate segmentations follow a similar pattern: the smaller the area, the higher the presentation speed.

Table 1 summarizes the essential details for each of the segmentation degrees. The subtended visual angles assume a nominal 72 cm viewing distance on a standard 15 inch (38 cm) LCD monitor. At all six segmentation degrees, 1–6, every pixel of the total strip is presented during a fixed total overall time of 58.2 s. This corresponds to a simulated UAV “over-flight” ground speed of 90 mph (~145 km/h).
Each of the six terrain image strips was chosen to provide a range of terrain types, from light scrubland to extensively forested areas. Some man-made features appear in the strips, as they would in a typical over-flight situation. Each terrain strip contained between three and five individual targets contained in exactly three groups. Photo-realistic targets were inserted into the terrain strips using standard photograph manipulation techniques. These targets were correctly scaled, and were distinct. Participants were asked to indicate each individual target by pressing the keyboard space bar.

During each of the image sequence presentations we recorded the eye gaze position of each participant using an LC Technologies eyegaze tracker system (http://www.eyegaze.com/). This equipment records the gaze position relative to the display screen coordinates every 16.6 ms (60 Hz) and this data was saved for later analysis. The unit requires calibration for each individual participant prior to recording and this was done as part of the experimental procedure (section 5). Once calibrated, accuracy is given as about 15 screen pixels.

The equipment also records eye blinks and times when the user was looking away from the test display.

5. EXPERIMENTAL PROCEDURE

Twelve volunteer participants agreed to take part; they were drawn from the student and staff population of the University Electrical and Electronic Engineering Department. The average age of the participants was 26.5 years. All were asked to declare uncorrected vision deficiencies and that they were not sensitive to rapidly changing or flashing visuals. No such problems were reported and all signed consent forms.

Once seated at the workstation each participant was shown a brief animated presentation (~90 s) outlining the purpose of the test and procedure to be followed. The animated presentation introduced the task and motivation for using UAVs in aerial search and rescue, gave an example of the terrain over-flight scenario and presented an example showing multiple human search targets, which were then highlighted.

Figure 2 shows a still from this video, showing four people in a terrain scene: two individuals and a group of two. Participants were instructed to press the space bar once for each person they located (four presses in this example).

In the next phase of the experiment eye gaze calibration was performed. This was successfully completed for all twelve participants. Following this, each of the six terrain image strips were shown. Space bar presses were recorded for later comparison to actual target appearances. Participants were free to make additional comments, which were noted. Participants were offered a doughnut as thanks for their assistance. Overall time for completion for a participant was under 20 minutes.

### Table 1: Segmentation degree details

<table>
<thead>
<tr>
<th>Segmentation</th>
<th>Tiles</th>
<th>Time (ms)</th>
<th>Tile Size (px)</th>
<th>Subtended Visual Angle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Degree</td>
<td></td>
<td></td>
<td>Width</td>
<td>Height</td>
</tr>
<tr>
<td>1</td>
<td>15</td>
<td>3878</td>
<td>1024 768</td>
<td>22.62° 17.06°</td>
</tr>
<tr>
<td>2</td>
<td>60</td>
<td>3705</td>
<td>512 384</td>
<td>11.42° 8.58°</td>
</tr>
<tr>
<td>3</td>
<td>135</td>
<td>431</td>
<td>342 250</td>
<td>7.64° 5.72°</td>
</tr>
<tr>
<td>4</td>
<td>240</td>
<td>242</td>
<td>256 192</td>
<td>5.73° 4.30°</td>
</tr>
<tr>
<td>5</td>
<td>375</td>
<td>155</td>
<td>205 154</td>
<td>4.59° 3.45°</td>
</tr>
<tr>
<td>6</td>
<td>540</td>
<td>108</td>
<td>171 128</td>
<td>3.83° 2.86°</td>
</tr>
</tbody>
</table>

Figure 2: Example frame with four indicated targets

6. ANALYSIS OF RESULTS

The main focus of this study is the success, or otherwise, of visual search target identification across the six segmentation degrees. These can be separated into three distinct categories:

**Identified:** the subject reported the presence of the target by means of a space bar press when a target was shown on the screen (a True Positive, TP, result).

**Unidentified:** the subject failed to report the presence of a target (a False Negative, FN, result).

**Mistakenly Identified:** the subject reported a target by space bar press when none was shown on the screen (a False Positive, FP, result).

The Identified (TP) condition is the ideal, while the Unidentified (FN) condition is very serious, corresponding to a missed target and potential failure to rescue the victim. The Mistakenly Identified (FP) condition can lead to a waste of ground search resources, and should clearly be minimized. The percentages of Identified (TP) and Unidentified (FN) responses must sum to 100% within each terrain sequence presentation, whereas the number of Mistakenly Identified false positives (FP) is unbounded.

In the context of the experimental procedure defined, some care needs to be taken over the definition of an “Identified” (TP) target. It is clear from many previous studies [e.g., 4], that there is a significant time interval between a person identifying an on-screen target and the actual moment at which the physical response (pressing the space bar) occurs.

With shorter image presentation times, characterized by pre-attentive recognition (section 1), this delay could result in multiple images replacing the target before the user’s response is recorded. During longer image presentation periods, characterized by active search, there may also be a significant time interval between a target appearing and its identification. Therefore we define the “Identified” period as starting with the appearance of the target to the time the target disappears from the screen plus an additional 2 s.

6.1 Target Identification Rates

With the above interpretation in mind, figure 3 shows the effect of segmentation degree on both Identified (and so Unidentified) and, below, the Mistakenly Identified rate. These are averaged over all participants and all trials.

There is an apparent trend between segmentation levels 1–3, showing a substantial reduction in target identification (and corresponding increase in unidentified targets). There is an apparent stabilization of identification for the higher segmentation degrees, 4–6. Care needs to be applied to this
700

interpretation, as an ANOVA\(^1\) shows that none of these differences by segmentation is statistically significant to \(p < 0.05\) (\(F(5, 55) = 1.907, p = 0.108\)). We note that the lack of difference across segmentations is in itself interesting.

For Mistaken Identification, any apparent correlation between a segmentation degree and number of mistakenly identified targets should be treated with care. An ANOVA analysis\(^2\) of these data reveals a \(p\) value of 0.895 (\(F(5, 55) = 0.327\)).

7. DISCUSSION AND FUTURE WORK

The main, and unexpected, finding from this study is the apparent constancy of target recognition across the range of segmentation degrees. This is perhaps more so because there is a clear shift in visual search strategy from segmentation degree 1–6; a transition from a generally “serial” (attentive search) to a forced “parallel” (pre-attentive) strategy [8].

The observation that target identification can take place with a presentation rate as short as 108 ms is surprising, as takenly Identified differences among segmentation degrees.

transformation applied to the Identification results to satisfy differences amongst segmentation degrees with an arcsine transformation. As presentation rate decreases, from degrees 4–1, the screen area become increasingly significant at higher rates. As presentation rate decreases, from degrees 4–1, the significance of these issues reduces.

We have only tested identification to a maximum full screen size (1024 × 768 on a 15 inch monitor). There is an improvement trend in recognition scores (figure 3) across segmentation degrees 3 = 45%, 2 = 72% and 1 = 77%, and there might be further recognition performance gains to be made in testing yet larger and slower presentations.

Every tile is displayed at its native resolution (six inches ground area per pixel). There is a well-established correlation between visual acuity (distance from fovea) and the size of a target [e.g. 1]. It would be informative to determine whether magnifying (using standard image scaling techniques) the higher segmentation degrees to artificially change observable size would affect target identification. While no extra resolution would arise it would provide insight into this size/acuity trade-off. Mardell et. al. [5] have reported on using selective magnification based on gaze position.

There is a reduction in gaze movement from segmentation degree 1 (average of 1443\(^\circ\) per experiment trial) to degree 6 (540\(^\circ\)), suggesting that reduced gaze travel does not markedly affect recognition rates. This tends not to support the notion that improvements in reading rates in text RSVP is primarily due to this reduction in gaze travel.

8. SUMMARY AND CONCLUSIONS

We have investigated the effects on visual search of a straightforward trade-off between image size and presentation time, using a simulated WiSAR scenario. We observed that small targets can be effectively recognized within complex photo-realistic scenes with very short exposure times (~100 ms), when presented in small segments (in the style of a Rapid Serial Visual Presentation). We also observed that the recognition rate did not differ markedly from when the same targets were to be identified in equivalently larger image areas, but for a correspondingly longer time. Using eye gaze recording equipment we confirmed that this equivalence is maintained across two quite distinct visual search strategies, active search and parallel (“pre-attentive”) recognition.

References