Visual saliency as an aid to updating digital maps

C. Davies*, W. Tompkinson, N. Donnelly¹, L. Gordon and K. Cave²


Research & Innovation, Ordnance Survey,
Romsey Road, Southampton SO16 4GU, UK.
Telephone: +44 (0) 23 8079 2936
Fax: +44 (0) 23 8030 5072
Email: Clare.Davies@ordnancesurvey.co.uk

(1) School of Psychology
University of Southampton
Southampton SO17 1BJ, UK

(2) Department of Psychology
University of Massachusetts
Amherst, MA 01003, USA

* (author for correspondence)
Abstract
Visual attention is considered in the context of a professional computer-based task, using aerial photography for updating topographic mapping data (photogrammetry). There is potential for using visual attention models to help develop various semi-automated 'attention-aware' support systems for this task, and these are discussed. An experimental study is described which examined the potential influence of expertise, image type and exposure duration on the role of visual saliency or salience (as calculated by Itti & Koch (2000)'s saliency maps) in the distribution of visual attention with such imagery. Using a non-intrusive, low-resolution and low-cost method to determine the approximate distribution of visual attention, effects of expertise and landscape type were found. Unexpectedly, saliency appeared to be more relevant to visual attention among expert users than novices, and potential reasons for this are explored. Implications and further research plans are discussed.

Keywords: visual attention; aerial photography; top-down/bottom-up processing; expertise; visual saliency
1. Introduction

Probably every reader has seen a ‘remotely sensed’ image such as an aerial or satellite photograph. Photogrammetry is the science of interpreting these remotely sensed images to help produce or update topographic maps. Typically the images will be inspected to identify changes in topographic objects that appear in the landscape, such as the location of a new building or a change in the shape of a field boundary. Once these changes have been identified, a photogrammetric operator (also called a photogrammetrist) will use computer software to digitise (i.e. trace) the new outlines of topographic objects into the digital map. Figure 1 illustrates an example of aerial photography and the corresponding topographic mapping that has been derived from it.

Figure 1: A comparison of aerial photography with its corresponding topographic mapping. A building has been circled in each of the images to show how the outline of a building depicted in the photograph appears when it is digitised in the corresponding map. Ordnance Survey data © Crown Copyright, 2003.

The goals of any mapping agency include the rapid collection, maintenance and distribution of up-to-date spatial information, and partial automation of existing processes can help achieve these goals. Indeed, maintaining national mapping is a very involved task; at Ordnance Survey (Great Britain’s national mapping agency) at least 5000 changes are made daily to the national topographic database (Ordnance Survey, 2004). As a result, continual investment is required to ensure that its maps can continue to be relied upon.

When maintaining the currency of mapping, photogrammetric data capture is one area where efficiency gains may be made. However, despite using powerful image display and mapping software on high-end computer workstations, photogrammetry is still a largely manual procedure, relying heavily on the interpretation skills of the human photogrammetric operator. There is obvious potential to improve the efficiency of this process. For example, improvements may be possible through producing tools that are modelled upon, and complement, the operator's basic psychological skills whereby visual attention is allocated across images. Towards this end, the research reported here begins to investigate the ability of a computational model to predict the characteristics of human visual attention when
interpreting aerial photography.

This paper begins by describing why full automation of photogrammetry is still not achievable. Focussing on the human aspects of this problem, a brief section reviews the recent progress in computationally modelling human visual attention, with particular regard to the task context of photogrammetry. Top-down influences on attention have recently come to the fore, challenging visual saliency-based models. An experiment was conducted to examine the apparent relevance of these two aspects of visual attention, paying particular attention to the role of photogrammetry expertise and of different types of depicted landscape. The final section discusses the implications of this and further research for partially automating the computer-based photogrammetry task.

2. Background

2.1 Automated mapping

In understanding why the above procedures are still very manually intensive, it needs to be appreciated that landscapes are inherently complex, to the extent that current machine intelligence does not begin to match human intuition in the extraction of useful information (Sowmya & Trinder, 2000). It is very difficult to apply automated techniques to mapping to allow objects to be automatically digitised from remotely sensed imagery. This level of object recognition and manipulation has been a goal of the photogrammetric and computer vision research communities for over 20 years. Yet there are no universal automatic techniques that can recognise and extract topographic objects in all contexts with sufficient success (Agouris, Mountrakis & Stefanidis, 2000).

Bellman & Shortis (2004) suggest why there has been limited progress towards fully automatic mapping solutions. If automatic mapping can be separated into the steps of object identification (through image interpretation) followed by the extraction of those objects, Bellman & Shortis assert that photogrammetric research has focussed upon the second of these steps. That is, advancements in photogrammetric research have concentrated on producing algorithms that digitise the outlines of specifically defined known objects in the landscape, and not upon applying technology to understand how new objects are initially detected. As a result, the entire process still relies on the image interpretation skills of the human operator, to target the execution of specific algorithms to extract the desired objects. However, we know little about exactly how operators find and identify target objects, or how we can help them to do so.

2.2 Visual Attention

Visual attention is the mechanism that enables “people to select the information that is most relevant to ongoing behaviour” (Chun & Wolfe, 2001). The crucial role that attention plays in human visual perception has been well demonstrated, along with its benefit to computer vision tasks that, in particular, are applied to natural scenes (Hamker, 2004).

How might we use an understanding of visual attention processes to help photogrammetrists in computer-aided mapping tasks? A number of possible scenarios could be envisaged. It might eventually be possible, for instance, for a preliminary machine vision process to identify apparent objects within the image (in geographic terms, 'features'). The computer could then edit the image to emphasise object groupings or structure, to encourage the
operator to quickly assimilate the visual structure of the image and to direct her or his attention between major areas rather than processing individual items. Since the operator is often trying to compare the imagery to an existing map, which also tends to simplify features but preserve shape and structure, this highlighting of the overall structure might make it easier for the operator to define key areas for further examination.

However, such a pre-processing solution would only be worthwhile if we knew that deriving the structure of an image was in fact difficult for an operator. Research dating back to the Gestaltists of the early twentieth century (e.g. Koffka, 1935) suggests that at present it is actually far harder for computers to do this than for humans. In fact, we seem to find it hard not to extract structure and shape, whereas deriving vector graphics from images such as photographs is still a painful process for a computer.

Alternatively, imagine if low-level visual saliency aspects (e.g. brightness or contrast) were shown to be the primary driver of human visual attention. This might imply that operators would be too easily distracted away from spotting a change in the landscape that was perhaps not visually salient in itself. To increase operator efficiency, we might envisage a pre-processing algorithm that identified areas where map revision was required, and placed a visually salient marker on each such area to automatically attract the operator's attention. However, if a trained photogrammetrist is able to ignore visual saliency and follow more appropriate strategies for directing attention, again this may not be worthwhile.

A computational model of visual attention could also be used to underpin a system that ‘drives’ a human interpreter to areas of interest within aerial imagery that covers a large expanse of the landscape. If the model demonstrated that most changes were occurring in regions where a human was unlikely to attend, such a system might 'zoom in' to areas of interest for a particular task that the operator would not have found so quickly if left to their own devices.

However, if top-down influences proved to be more critical in directing attention in photogrammetry, then understanding those influences may suggest a completely different kind of solution – perhaps one based on the semantics of the landscape and of the task at hand, which would require a much more intelligent approach to any automation efforts. Therefore, it is important to understand the process of visual attention in image interpretation for mapping.

To move towards this goal, we need to be clear about the applicability of current, laboratory-derived, visual attention models to this real-life computer-based task. The most popular current model is Itti and Koch’s model of selective visual attention (Itti & Koch 2001). Therefore a primary goal of the current research was to compare this model against real human participants viewing remotely sensed images, to investigate the relevance of the model's measure of saliency to people's actual attention under varying circumstances. Although several methods for computationally modelling visual attention have been published (Walther, Rutishauser, Koch & Perona, 2004), the Itti & Koch model was chosen as a starting point due to its relatively straightforward bottom-up design, and to evidence of its application to natural scenes (Itti & Koch 2000, 2001; Parkhurst, Law & Niebur, 2002). Moreover it is popular, being cited in 14 of the 18 papers presented at the 2nd International workshop on attention and performance in computer vision (Palletta, Tsotsos, Rome & Humphreys, 2004).

Inspired by the responses of primates' early visual systems within the complex problem of
scene understanding, Itti & Koch's model is a system that builds a series of neural 'feature maps'. Each of these maps extracts local spatial discontinuities in the different modalities of colour, intensity and orientation. These low level features are extracted from an original colour image at several spatial scales using linear filtering. The result is a biological model that makes the system sensitive to local spatial contrast in a given feature, rather than its amplitude (Itti & Koch, 2000). The input from these multiple, quasi-independent feature maps are then combined to give rise to a single output, the saliency map. In this, the relative “saliency” at every location in the visual field is given by a number on a scale from 0 to 255.

As someone looks at the image, his or her visual attention targets the most salient locations first, and then the next most salient locations that have not yet been viewed. A dynamic neural network models this, using the saliency map as an input and iteratively selecting each attended location in the order of decreasing saliency (Itti, Koch & Niebur, 1998). Competition among neurons in the saliency map give rise to a single winning location that corresponds to the next attended target. Inhibiting this location automatically allows the system to attend to the next most salient location (Itti & Koch, 2000).

For the purpose of this work, the last stage (the use of the neural network) is not used. Instead the saliency map is the basis of the analysis for this study, to test whether salience, as determined by the Itti & Koch model, is related to the actual visual attention patterns of expert photogrammetrists and matched novices when viewing aerial images.

2.3 Previous work

2.3.1 Tests of the model

Other researchers have recently tested the Itti & Koch model for human participants with computer-based imagery. For instance, Parkhurst et al. (2002) tested the model against four students' visual fixations with home interior scenes, natural landscapes, buildings and city scenes, and computer-generated fractals. The students were not given any task instructions or reasons to study the images, beyond being asked to look around the scenes. Each image appeared for five seconds, and an eye movement tracker recorded participants’ gaze direction. The Itti & Koch saliency map showed best predictive power for the very first fixation, but became less relevant over time. Colour and intensity (contrast) appeared to play a bigger role in this than orientation, although the features' respective roles varied greatly by image type. This could help explain why contrast was found to be much less important in another study by Einhäuser & König (2003). The fundamental differences in saliency attributed to luminance, as opposed to orientation, are also explored further in Nothdurft (2000). Taken together, all of these studies suggest that applying the model may not be straightforward, as its different components may have variable effects.

2.3.2 Top-down influences

Bottom-up models like Itti & Koch's predict attention purely from visual features. However, as reviewed by Henderson (2003), human gaze control is 'smart' in using cognitive information as well. Henderson lists the sources of such information as:

- Short-term episodic knowledge in working memory (where have I just been looking?)
- Longer-term episodic knowledge (have I seen this image before? What happened then?)
Henderson argued that visual saliency is less relevant to visual attention when people are looking at meaningful scenes within the context of an active task. Therefore, we might find Itti & Koch's model working less well for photogrammetrists actively inspecting images than for Parkhurst et al.'s untrained and goal-free students. Consistent with this view, recent studies (Ludwig & Gilchrist, 2002; Oliva, Torralba, Castelhano & Henderson, 2003; Wolfe, Horowitz, Kenner, Hyle & Vasan, 2004) have shown that top-down influences can actually affect attention within the first 200ms, i.e. the very first fixation. Thus we could hypothesise that expert photogrammetrists may have greater reason to diverge from the patterns of attention predicted by a bottom-up model like Itti & Koch's.

Equally, though, we might expect that top-down influences of photogrammetrists' usual tasks with images may be suspended in a situation where they know they are not doing those tasks; instead, being highly familiar with this type of image, they may fall back on allowing saliency to guide their attention for want of any other strategy. In such a circumstance we might see no expert-novice difference due to top-down effects, or we might even see a reverse situation where the confusion and unfamiliarity of aerial imagery to novices might disrupt the usual attentional mechanisms so that they are less predictable by visual saliency (or any) models.

Updating maps involves a semi-focused visual search or change detection task, where the operator is looking for anything unusual or different from expectations. Unlike many stimuli used in visual attention research, aerial photographs are meaningful, especially to photogrammetry experts. A highly-trained photogrammetrist would be aware of different types of image within aerial photography, in terms of the landscape content (for example, agricultural versus urban); the nature of the map update task also tends to vary with landscape type. Therefore, we hypothesised an additional potential 'top-down' effect of image type within our database.

However, by 'top-down' most researchers seem to imply quite low-level and artificial goals within the experimental context, rather than the higher-level context of the task. Oliva et al. (2003) suggested that in a more naturalistic task, general knowledge could also play a part (for example, knowing that people tend to be found on the pavement in a street scene). It is not clear whether the specific nature of users' longer-term expertise, or their usual (as opposed to experimental) task experience with a specific image type, may also play a part in early visual attention in computer-based tasks. If not, then no differences between experts and novices would be found.

3. Method

3.1 Participants

The experiment had 54 participants; all were employees of Ordnance Survey. Four participants' data had to be dropped: two failed to respond to any of the 500msec trials; one admitted to trying to 'sabotage' the experiment by deliberately looking at featureless areas; the data for the fourth was lost in a system crash. Ages of the remaining 50 participants ranged from 21 to 60 years (mean = 36, median = 34). There were 19 females (12 novices and 7 experts).
To test for expertise effects, the sample of 50 included 24 photogrammetrists ('experts'), and 26 comparable staff who rarely looked at aerial imagery at all ('novices'). There were no significant age ($t = -0.66, p = 0.5$) or gender (Pearson's chi-square = 1.53, $p = 0.3$) differences between the two groups.

Group membership was initially defined by the department in which people worked (novices were recruited largely from administrative departments with little contact with the mapping database). We also double-checked this via a questionnaire checking the recency and frequency of people's viewing of aerial photography. For example, 17 of the 24 experts had already viewed such imagery earlier that day, 6 within the past week, and 1 within the past two weeks. By contrast, none of the novices had viewed imagery that day, 6 not for at least a fortnight, and 15 not for at least a month.

Staff took part in the experiment voluntarily, within paid work time.

3.2 Design

Given the possibility of top-down influences, we were concerned about the physical and demand-characteristic effects of presenting an experiment to professionals with no experience of scientific method. This was an initial proof-of-concept experiment to see whether visual attention measures could distinguish anything meaningful with aerial imagery and with older, non-student users, prior to investing in further studies and expensive equipment. With this in mind, we employed a conventional number-probe task rather than physical eye-movement tracking, to test for visual attentional foci within each image at the end of three different exposure times (so that we obtained three sample points along the time course of early visual attention).

The number-probe task involves briefly flashing onto the screen a randomised and pseudo-randomly displaced grid of 25 two-digit numbers (ranging from 11 – 99, omitting all those with zeroes which are much easier to identify). The numbers appeared immediately after the test image was removed from the screen, and were followed by a mask of stars to disrupt the immediate visual after-image. Participants then had to recall as many numbers as possible. Piloting showed that with a 100ms duration for the display of the numbers, people usually recalled less than three numbers, and usually only one. To ensure that people tried to encode the images, a two-choice recognition task followed the number probe task. After recalling the numbers, participants were presented with two smaller images side-by-side, one of which was the test image. They had to press a key with the left or right hand to say which image they had just seen.

So that we could measure three separate points along the time-course of initial attention, each participant needed to see each image three times. This was achieved via three blocks of 50 trials (one for each of 10 images in each of five categories - see below). The order of the three blocks was randomly assigned for each participant, and the order of trials was randomised within each block.

To enable comparison between the output of the saliency model and the attention responses of human participants, each image was analysed as a grid of 5x5 squares, with an average saliency computed across the pixels in each square. While number probes provide a coarser index of attentional fixation than measuring eye-movements, they are sufficient to demonstrate whether saliency in the Itti & Koch model differentially influenced visual attention for experts and novices, and for different image types.
3.3 Procedure

The experiment took place in a small quiet meeting room within Ordnance Survey headquarters. Each participant was tested alone. The first 10 trials of each block were practice trials and were not recorded.

Participants were seated and completed a brief questionnaire, which recorded their age and gender and asked a few questions on their experience with imagery. The reverse of the questionnaire sheet contained ethical reassurances and overall instructions for the experiment. Participants then put on headphones, before reading more detailed task instructions on the computer screen. The experiment began when they were ready.

Each trial began with a 500ms tone warning that an image was about to appear. The test image was displayed for 500, 1250 or 2000 ms, followed immediately by the number grid (different for each trial) for 100 ms, and then the star mask for 500ms. A dialog box then asked participants to type any numbers they recalled. When they could recall no more they pressed return to continue, and were presented with smaller versions of the test image (randomly assigned to right or left) and another image, side by side. Participants had to press either the 'F' or 'J' keys on the computer keyboard to correspond to the left or right image, to say which one they had previously seen.

Between blocks of trials participants were encouraged to rest their eyes and take a break for as long as they wished, although few took more than a minute. At the end of the experiment participants were asked to talk about how they had found it and what they thought it was about, and debriefed about its purpose.

3.4 Materials and equipment

The test images consisted of randomly-generated samples from Ordnance Survey's current aerial photography database. The sample images were then selected and classified into clear categories that reflected their landscape type, rejecting those where there appeared to be a mixture of classes (for example, urban and rural or suburban and industrial). Five broad categories were used: agricultural, urban, suburban, uncultivated wilderness (mountains, moors, forests) and industrial. Examples are shown in Figure 2 (shown in colour in the experiment). The experiment was run on a normal PC with CRT screen.
4. Results

For each of the 50 images, a visual saliency map was calculated according to Itti & Koch's (2000) model. This map was then simplified to match the resolution of the collected responses, by averaging the saliency values within each square of the 5x5 grid. The mean saliency value was determined for each square that had been responded to by each participant, for each image, at each time duration. On the rare occasion that a participant reported more than one number, the average saliency across the two squares was used.

This saliency variable was analysed by expertise, sex, age, exposure time and image type, to determine whether these factors affected its relevance to the distribution of visual attention by expert and novice participants. A repeated-measures ANCOVA model was used (expertise and sex as 2 x 2 between-subjects factors; exposure time and image type as 3 x 5 within-subjects factors; age as a continuous covariate). Expertise affected the reporting of numbers such that experts' responses showed significantly higher salience values than those of novices (F(1, 42) = 7.61, p < 0.01).

The role of saliency appeared to decrease with exposure time, so that it was less at 2000ms than at 500ms; however, this did not reach significance (F(2, 42) = 1.77, p = 0.17). This was not surprising though, since previous studies (for example, Parkhurst et al., 2002) have shown that this decrease over time is quite a shallow slope. Image types did significantly differ, however (F(4, 42) = 3.48, p < 0.01). The relative saliency of people's responses with different image types is shown in Figure 3.
Expertise did not significantly interact with either time (F (2, 42)<1) or image type (F(4, 42)<1). However, there was an interaction between expertise, sex and age (F(1, 42) = 6.05, p = 0.01). This appeared to suggest that any effect of gender is cancelled out by experience with the task. Female novices (and female experts, who were almost identical) attended to more visually salient locations than male novices, regardless of age. However, male experts showed the highest level of saliency in their attended locations, once they had some years of experience. This is shown in Figure 4.

---

1 We assumed, in checking for expertise, that image viewing recency and frequency were more important than length of experience. Yet Figure 4 also suggests a role for the latter, at least for male photogrammetrists.
5. Discussion

The results suggest, unexpectedly, that experience with aerial imagery leads experts to be more responsive to visual saliency than novices. One possible explanation for this is that remotely sensed images are complex. There was some anecdotal evidence that, particularly among novices, the 500ms exposure was perceived to be too fast for conscious strategic viewing. In such situations the novices may have been more likely to default to looking at the centre (or a random area) of the image, or scanning around it randomly, rather than behaving in the ordered fashion typical of visual search studies. However the effect of expertise was found across all exposure durations and so this is unlikely to fully explain the experimental effect reported.

The fact that saliency was more relevant to experts' attention is useful in enabling us to use and extend models like Itti & Koch's, towards predicting and understanding professional imagery users' behaviour. However, as suggested earlier, it also suggests that visual saliency could have a distracting impact on photogrammetrists' work, in situations where a landscape change was not visually salient from the air. If further research confirms this, then this might (for instance) strengthen the case for computational image pre-processing that would adjust the saliency levels of key areas where change was expected to have occurred.

What happens when a novice becomes an expert? In other words, what defined the difference between these two groups in our study? As shown earlier, the involvement of staff from within Ordnance Survey for the 'novices' group allowed us to discount issues such as age, gender, or familiarity with the general geographic mapping domain. The key variable noted earlier was the recency with which participants had viewed an aerial photograph, and this in
turn is a partial indicator of their familiarity with this type of image. Familiarity with such images may have a number of subtle effects: one could be an ability to mentally parse and interpret the image, so that the expert would not suffer any confusion or curiosity about the nature of its content (bearing in mind that some geographic objects are quite difficult to recognize from the air, e.g. the last example in Figure 2 which is actually part of a coal-fired power station).

Alternatively, another effect of familiarity might be an ability to actually ignore that content, treating the image not as a real-world photograph but merely as another piece of data for processing. The expert would therefore be open to simply following the guidance of low-level visual patterns, and in some circumstances this might actually be beneficial (e.g. in matching those patterns to a pre-existing map). Previous work (Davies, 2002) has considered this ability of users to almost simultaneously view and interpret the same computer-based geographic image as either a representation of the real world, or a complex abstract geometric figure. It may be that experts are better able to 'turn on and off' the deeper semantic interpretation, depending on the task at hand, and this may help them to deal more efficiently with each new image.

The results also show that saliency appeared to have different levels of impact for different image types. This could be due to at least two very different issues: at the low level, the distribution of strongly-salient points varied between images, and Itti & Koch's model does not help us to account for this since the saliency scale for each image is calculated relative only to that image itself. At a higher level, there are obvious semantic differences between the images' contents: both experts and novices would have understandable reasons to respond differently to suburban areas (where people are quite likely to have extended or built new houses), and wilderness areas (where little may have changed, and there is less man-made structure). With the present data it is hard to tease out the differences between these two factors, although further work will do so.

For the present initial investigation, these findings clearly justified the use of a relatively simple, physically unintrusive and low-cost method for measuring the distribution of visual attention, rather than full eye tracking. This shows that such methods still have value in circumstances where the precise location of gaze is not critical to hypothesis testing. Obviously there are drawbacks, particularly the uncertainty surrounding the true contribution of saliency to the variance in gaze fixations, and hence the exact effect size and predictive power. Further work is using detailed eye-movement tracking to tease this out.

It should be noted that this study used a considerably larger participant sample than some previous tests of Itti & Koch's (and other) visual attention models. This was necessary due to the explicit between-subjects comparison (experts versus novices). However, it was noticeable that the distribution of visual attention varied greatly even among members of the same participant group. These individual differences may be of some concern when examining the predictive power of computational models for human attention.

The continuing theme of a top-down role in early visual attention suggests that a bottom-up model like that of Itti & Koch may never optimally predict people's visual attention. Alternative models such as that of Tsotsos, Culhane, Wai, Lai, Davis & Nuflo (1995), which included a role for task-specific attentional bias, may need supplementing further with the factors suggested by Henderson (2003). However, it is not yet clear how factors such as episodic scene knowledge or scene-schema knowledge can be formally predicted, nor how they would be incorporated into the mathematical model to produce an overall 'personal
saliency map'. Since higher-level cognitive factors are even more likely to vary between individuals than low-level perceptual strategy, far more sophisticated models may be needed.

From the point of view of a mapping agency attempting to use these factors to understand and help photogrammetrists, more work is obviously needed. While a perfect automatic saliency prediction model may be some way off, the constraints of the photogrammetry task and the contents of aerial imagery (at least of the British landscape) are finite, which makes further research more manageable. Ongoing work is now defining the exact cognitive processes of this task, to help define the factors which can realistically be built into an integrated model of attention within this context. Such a model will in turn help us to focus realistic attempts to make semi-automated photogrammetry support systems that are, at least partly, 'attention-aware'.

6. Acknowledgements

Code for implementing Itti & Koch’s visual saliency model was obtained from Itti (2003).

The authors are very grateful for the assistance of Natalie Lock and Helen Sockett for running the visual attention experiments with participants at Ordnance Survey. Also to Luke Phillips for developing the software, and to David Holland for his assistance in preparing the images used in these experiments.

This article has been prepared for information purposes only. It is not designed to constitute definitive advice on the topics covered and any reliance placed on the contents of this article is at the sole risk of the reader.

7. References


