Visual saliency from image features with application to compression

Harding, P. J.; Robertson, Neil

Published in:
Cognitive Computation

DOI:
10.1007/s12559-012-9150-7

Publication date:
2012

Cognitive Computing manuscript No.
(will be inserted by the editor)

Visual Saliency from Image Features with application to Compression)

P. Harding · N. M. Robertson

Received: date / Accepted: date

Abstract Image feature point algorithms and their associated regional descriptors can be viewed as primitive detectors of visually-salient information. In this paper, a new method for constructing a visual attention probability map using features is proposed. (Throughout this work we use SURF features yet the algorithm is not limited to SURF alone.) This technique is validated using comprehensive human eye-tracking experiments. We call this algorithm "Visual Interest" (VI) since the resultant segmentation reveals image regions that are visually salient during the performance of *multiple* observer search tasks. We demonstrate that it works on generic, eye-level photographs and is not dependent on heuristic tuning. We further show that the descriptor-matching property of the SURF feature points can be exploited via object recognition to modulate the context of the attention probability map for a given object search task, refining the salient area. We fully validate the Visual Interest algorithm through applying it to salient compression using a pre-blur of non salient regions prior to JPEG and conducting comprehensive observer performance tests. When using the object contextualisation, we conclude that JPEG files are around 33% larger than they need to be to fully represent the *task-relevant* information within them. We finally demonstrate the utility of the segmentation as a Region of Interest in JPEG2000 compression to achieve superior image quality (measured statistically using PSNR and SSIM) over the automatically-selected salient image regions while reducing the image file-size by down to 25% of that of the original. Our technique therefore delivers superior compression performance through the detection and selective preservation of visually-salient information relevant to multiple observer tasks. In

P. Harding
Heriot-Watt University
E-mail: patrick.harding.is@gmail.com

N. M. Robertson
Heriot-Watt University
E-mail: N.M.Robertson@hw.ac.uk

contrast to the state-of-the-art in task-directed visual attention models, the Visual Interest algorithm reacts only to the image content and requires no prior knowledge of the scene nor of the ultimate observer task.

Keywords Feature Points · Visual Saliency · Task-directed Viewing · Object Recognition · Region of Interest Compression

1 Introduction

In manual image analysis there are regions of an image that an observer is naturally drawn to because of the nature of the image content. In combination with the observer's purpose or "task" this information is used to guide the deployment of visual attention. A range of contextual qualities may trigger pre-cognitive visual attention: colour, orientation, curvature, size, motion, depth cues and aspects of shape. This pre-attentive vision is commonly referred to as "bottom-up" attention. Passive viewing (exploring the image content with no prior idea about the content or what is being searched for within the image) of an image is guided by this process [33]. The reaction time to bottom-up stimuli of this kind is approximately 25-50 ms in comparison with visual search which is competitive with the eye-saccade time (200ms) [13]. There exist reliable computational models of visual saliency, which typically locate a narrow focus of attention looking for the most salient regions by rank outputting probability maps matching the grid of the original image [11,8]. Such models assesses centre-surround differences in *Colour*, *Intensity* and *Orientation* across scale and assign values to feature maps based on outstanding attributes [11]. Cross scale differences are also examined to give a multi-scale representation of the local saliency. The maps for each channel (*Colour*, *Intensity* and *Orientation*) are then combined by normalizing and weighting each map according to the local values. In this way homogenous areas are ignored and "interesting" i.e. visually-salient areas are found and highlighted.

Task-directed viewing - in which the observer has some pre-conception of the features and layouts for which he is searching - is a far more complicated. A human has spent many years learning associations among objects and their context. This prior knowledge of object shapes and their appropriate contexts is a "top down" process that guides visual attention in combination with the pre-attentive "bottom up" information available to provide efficient search under task [34]. The bottom-up response is clearly very important to the formation of the overall understanding and ultimate interpretation of the image. The top-down processes are governed by the "task" the observer is performing, whether a memory task or an object search. How these two factors combine is unknown, althought it is generally accepted that the top-down process builds upon the bottom-up contextual information to provide the overall scene understanding [33]. The top-down information about the identity of a target acts quickly to configure the visual system to look for that target (in less than 200ms in some cases [34]). Top down processes involve voluntary control in

multiple visual cortical areas, resulting in selective sensory processing of relevant visual targets [9]. In addition to volition, observers under task select visual features optimally [21]. Therefore the top down processes are strongly under the the control of the observer and are efficient from an information processing perspective.

Computing task-directed attentive processes is not straightforward: this is intuitively understandable when one considers the nature of tasking. The range of possible tasks, top down processes and observer experience levels is enormous. Most of the computational-modelling work is inspired by what is known of the biological system. They generally work on the basis of a bottom up attention map in combination with some priors about the nature of the task that introduces an element of learned target representation that biases the map as an approximation to top-down processes. The top-down correction is necessary because observers under task will look in regions that lie outwith the "bottom-up" regions of predicted attention if alternative regions make more semantic sense for the task, as described by Brockmole *et al.* [3].

Navalpakkam and Itti propose a top-down saliency metric that maximises the signal-to-noise ratio between a search target and distractors [20]. However, this approach requires knowledge about bottom-up saliency in order to optimise the signal to noise ratio. Peters and Itti combine bottom up with a low-level signature of the entire image, and learns to associate different classes of signatures with the different gaze patterns recorded from human subjects performing a task of interest [22]. Torralba *et al.* look at a more sophisticated scheme for attention prediction that is based upon the contextualisation of objects under search [29]. The technique used to create the top-down correction to the bottom up surface associates object-class recognition with semantically-plausible locations defined by scene gist via a prior learning process. This is a clever model because it can reliably analyse the overall gist of a scene and discover the likely object class location therefore providing effective eye-fixation prediction for task *even in the absence of any actual objects of the class.* Torralba *et al.* [29] exploit horizontality (see also [6]) in the distribution of objects to produce their contextual modulation. The context modulation process is described in some detail in other work of Torralba [28]. The difficulty of this technique is the learning phase of objects and their contexts within the scene gist, which could be hard to generalise.

Task attention prediction models can provide predictions of observer eye-fixations under task that significantly outperform the bottom-up only maps. However, they are complicated and require a reliable learning phase of database object class recognition and object-gist/context association assessment. This is still difficult to do reliably to give consistent results in general scenarios. Furthermore, the techniques here described are looking to find regions of focused attention. That is, they look to narrow down to the most salient locations relevant to the task in hand.

There are therefore a few notable issues with extant saliency models. Firstly, the biological models are either presented for unrealistic constraints (e.g. the bottom up models look at the passive case: this is a rare situa-

tion) or are specifically tuned to task by modulating a bottom up map with some learned information. This is a problem because we want generality in a model. An observer should be able to change their task during analysis and our model should still be able to predict their attention patterns. Secondly, there is typically not a thorough probability analysis presented with the saliency maps produced by the models. Typically, the threshold applied to segment the maps is found iteratively to give particular image areas, which are then labelled as "salient". This is a problem, since the salient area of an image will vary strongly between images since it is highly dependent on the image content.

Computer vision feature points (basically, local features at which the signal changes three-dimensionally in space and scale) have many attractive properties, such as robust invariant descriptor matching over scale, rotation and affine offset that could be useful in combination with their use as a primitive saliency detector. Feature matching has been used in estimating inter-frame homography mis-matching as an estimate of temporal saliency in video [36], but not as a measure of spatial saliency. Harding and Robertson compare the co-occurrence of a set of computer vision feature points with predictive maps of visual saliency [7]. The authors studied the co-occurrence of six prominent feature detectors (SIFT [15], MSER [16], Harris-Laplace [17], SURF [2], FAST [24,25] and Kadir-Brady Saliency [12]) with visual attention maps (modelled using the algorithms of of Itti *et al.* [11] and Harel *et al.* for passive viewing and utilising task directed eye fixations to construct a map for the "task"). For all of the visual attention surfaces examined the interest points were strongly distributed towards the more visually salient regions of the images, with SURF and SIFT most strongly so. This is not an altogether surprising notion, since the construction of computer vision feature point algorithms act cross-scale and look for local differences in some way, much like the "bottom up" visual saliency algorithms. However, Harding and Robertson further note that where visual attention shifts away from bottom-up visually salient regions due to task, the *attention shift is directed towards regions rich in feature points.* It is therefore plausible that these feature point algorithms could be used to construct a *general* surface of visual attention: not one limited to bottom up or a specific task, but rather one that can cover all regions likely to attract attention under different viewing conditions. Importantly, the point-to-point descriptor matching capability of these features can be used for object recognition, such as described in [15] in which the distribution of matched features is checked between images. Such object recognition can be built-in to the probability surface construction to modulate the visual attention map towards the object relevant context of the image [29] with the advantage of utilising the dual properties of a single algorithm: salient feature detection and point description.

(a) VI(*General*) segmentation (b) GBVS segmentation to match area of (a)

Fig. 1 A demonstration of Visual Interest's advantages relative to the state of art. (a): VI applied to an image from "Aberdeen" (38% background). (b): GBVS applied to the same image to get the same excluded area. Note that the GBVS algorithm is selecting salient material relevant to *some* tasks but it is retaining the sky as salient while missing the people. In contrast, the VI segmentation successfully retains a lot more detail in the image. (See Figure 9 for probability map visualisation of these images.)

1.1 Contributions of this work

We propose in this paper an algorithm to construct such a general visual attention segmentation algorithm map based on the points generated by the SURF algorithm [2] , this being fast, having a robust descriptor and having high coincidence with visually salient regions [7]. As such, the aim of the proposed algorithm is to segment the image area containing information that is actually and also *potentially* salient to a range of search tasks. We use the robust matching properties of SURF to achieve object recognition. This is then exploited to refine the horizontal context similarly to the modulation carried out in [29]. We use comprehensive eye-tracking under task to validate the parameters both for the construction of a probability map and for the segmentation of it into "salient" and "background" regions. We create a set of rules with parameters set such that the segmentation is expected to be successful on generic eye-level surveillance-type imagery in RGB or black & white. By way of qualitative illustration we show the comparison of the VI algorithm with the Graph-Based Visual Saliency model [8] in Figure 1, this being the best-performing visual saliency algorithm without the requirement for a machine learning phase.

1.2 Paper roadmap

The paper is split into two main parts. In the first we discuss the refinement of a set of SURF feature points into a probability surface which represents a visual-saliency map (Section 2). This introduces the idea of features implying visual saliency, the calibration of the number of features and their spacing in the image and the validation via experiments with human eye-tracks. Then, in Section 3 we apply the visual interest segmentation to the JPEG2000 algorithm

as a predefined ROI. The concept of salient ROI compression is not new. Yu et al. [35] present an elegant case that observers attend one scale at a time and that this information can be used to filter images by salient priority to give impressive results for comparative compression over short observer assessment times. This approach is not favoured here, since we do not wish to modify the image to tune it to its salient features prior to compression, but rather to select and preserve the naturally salient regions over all scales which may apply to both short and detailed inspection by many observers. We do use blurring of detail away from the ROI such as in [10] for the validation phase but our end results integrate the ROI segmentation into the JPEG2000 international standard of compression designed specifically for ROI definition. In that section we validate the results on a public dataset and use Structural Similarity and Peak Signal-to-Noise Ratio as the measures of performance.

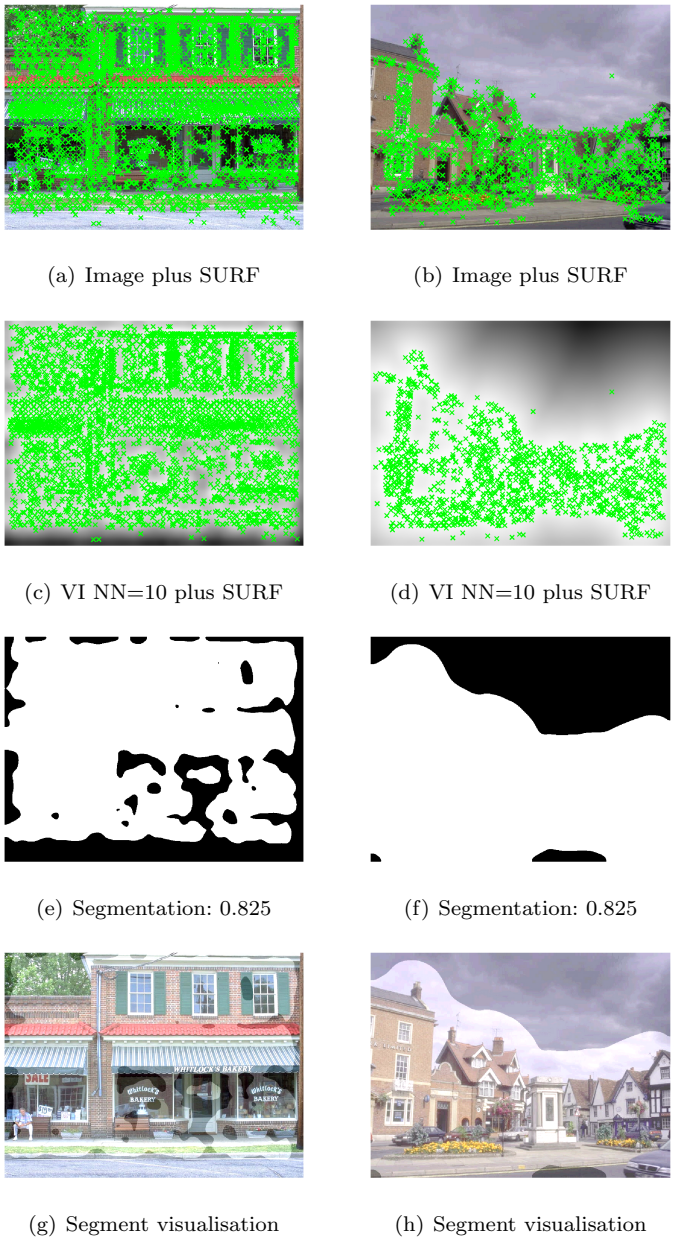
2 Computing Visual Interest from a generic image

To construct a probability-of-attention map from interest points, there are three factors that must be analysed to give a consistent result. The first is the threshold of the interest point algorithm. This parameter generally governs the number of points generated for a given image, but is image-content dependent. The second is how to join the points together. This will be expressed as the extent to which spatial neighbour points are included in the local density calculation. Then we wish to segment the image into regions of foreground (i.e. potentially valuable to an observer performing a task) and background using a threshold. Figure 2 illustrates the procedures described in sections 2.1 to 2.3 of the text.

2.1 Choice of threshold for SURF points

SURF has a threshold parameter that governs the number of points that are detected in an image. The algorithm looks for regions of high local variation over space and scale, filters these points for robustness and then applies the threshold to determine the number of points that are allowed through to the output.The relationship between the number of SURF points and the threshold depends on the image content. The threshold parameter filter acts on the absolute strength of the points, based upon the local variation. Therefore an image with bland content is expected a priori to return a lower number of points for a given threshold than would an image containing much detail.

We seek a set of interest points that will naturally cluster around the busy regions of an image and leave plain, featureless regions barren of points. We could do this by imposing a fixed number of points per image and iterating, changing the threshold until we reach that number. This is not an ideal situation for in the case of highly cluttered or sparse images we risk rationing or over-fitting weak points to the data using this method. We therefore seek



(a) Image plus SURF (b) Image plus SURF

(c) VI NN=10 plus SURF (d) VI NN=10 plus SURF

(e) Segmentation: 0.825 (f) Segmentation: 0.825

(g) Segment visualisation (h) Segment visualisation

Fig. 2 Visual Interest Surface (SURF threshold = 1.5, NN=10) applied to two different images. (NN=*number of nearest neighbour features. See Equation 3 later.*) (a) & (b): Image plus SURF points. (c) & (d): Derived NN=10 Visual Interest Surface plus SURF points. (e) & (f): Segmentation. Surface thresholded using threshold 0.825. (g) & (h): Illustration of image occupancy above threshold.

a measure to adapt the threshold to cope with general images reliably to give a suitable distribution of points on the image, based on the image content alone. Electro-optic sensors are often calibrated to balance the received signals in some way to reflect the degree of belief about the image quality as the output. There are many such operations available, as attested by the array of optional settings on any digital camera. Usually image compression is applied as a part of the capture process. This is often JPEG for high visual quality. The JPEG algorithm generally reduces the high frequency components in the image. While these processes undoubtedly help to produce high quality photos, traded off against storage capacity, the potential for variation between different camera-algorithm combinations is large and will impact the image content, and therefore the expected output of the SURF algorithm.

We make use of two datasets to inform our choices of threshold. The first is that of Torralba *et al.* (see url in [29]) and the second is our own. The Torralba dataset is 72 images heavily compressed using JPEG (inspection of the set shows JPEG blocking artefacts). Each image is RGB and 800 x 600 pixels in dimension. Our dataset "Aberdeen" is a series of 36 images containing people, street name signs and number plates. The SURF algorithm is applied with thresholds 0.05 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0 5.5 6.0 6.5 7.0 7.5 8.0 for all images in these datasets. The resultant number of points generated on each image was counted. This is shown in Figure 3, left. The right side of this figure shows the same data, normalised by the respectful image areas of each set.

The Y-axis is effectively a density value of the points generated for a given SURF threshold. The density function can be translated into a global (pan-image) average point separation by taking the inverse of Figure 3 (Top Left). E.g. a density of 0.005 corresponds to there being a SURF point for every 200 pixels in the image. We take the square root of this value to get a *linear average pixel separation*. The result of this is plotted in Figure 3, (Bottom Left), showing that while there may be some overlap between the extremes of the real datasets, generally they are distinct from each other in how they react to SURF, and there is possibly cross over between datasets. In order to get the same average linear spacing for the two datasets we need to apply different thresholds. The lower the threshold, the more points get through and so the average spacing decreases. Both sets tend towards a lower limit of circa 11 pixels of linear separation if *all* points are let through.

The linear separation values are global (pan-image) spacings. In general it can be expected that the SURF points will cluster towards the visually-salient regions in an image thus creating different local spacings. Table 1 shows the data from Harding and Robertson [7] in entabulated form. This data is the percentage of SURF points overlapping the different salient map categories by percentage area. The map categories are that of Itti [11], that of Harel [8] and the "task" surface generated from eye-fixations under tasking. The errors have been combined using the standard form, $\sigma_{mean} = \sqrt{{\sigma_1}^2 + {\sigma_2}^2 + {\sigma_3}^2}$. The study looked at the coincidence of 200 surf points with the visual saliency

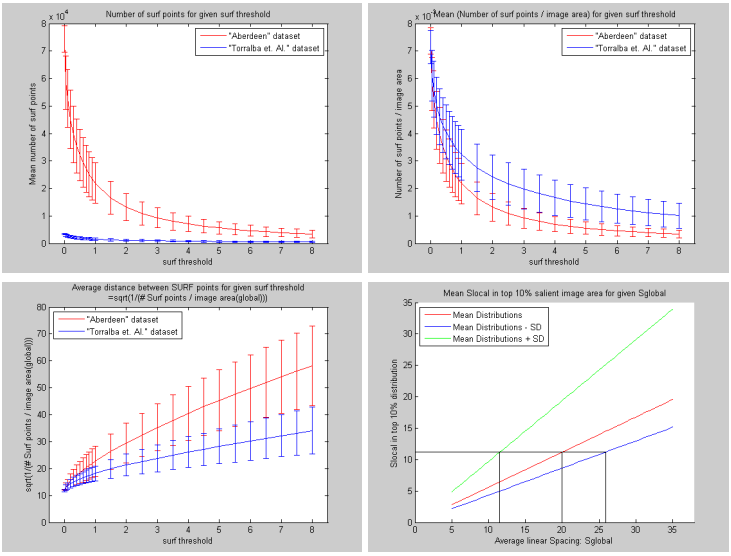


Fig. 3 *Top Left:* Average number of SURF points generated for a given SURF algorithm threshold. *Top Right:* Average number of SURF points generated for a given SURF algorithm threshold/image area. *Bottom Left:* Average linear spacing of SURF points generated for a given SURF algorithm threshold. Standard deviation at each value is shown. *Bottom Right:* $\overline{s}(90-100)$ generated for a set of possible $\overline{s}(global)$ (see Equations 1 and 2). This is the average local spacing in the top 10% saliency slice for a given global linear spacing. Standard deviation at each value is plotted seperately. Inteception lines with the minimum linear density of 11 (as seen in *Bottom Left*) are plotted. The mean global spacing required to achieve a top-10% saliency slice local density of 11 is 20 pixels.

Area Slice	Itti	GBVS	Task	Mean(All)	sd(All)
50-100%	86 ± 11	95 ± 5	82 ± 11	87.7	16.34
60-100%	69 ± 15	89 ± 9	73 ± 11	77	21.79
70-100%	58 ± 17	81 ± 12	61 ± 11	66.67	24.02
80-100%	40 ± 17	67 ± 16	49 ± 11	52	27.75
90-100%	20 ± 13	43 ± 18	30 ± 11	31	25.24

Table 1 Table of SURF point distributions towards ranked salient image area percentages of three distinct visual attention maps and their combined averages.

schemes at different image area percentages ranked in saliency. This clustering effect is quite a strong one.

We can take the *distributions* from Table 1 and find out what values of the *local average linear spacing* are generated (in pixels) for each ranked slice of saliency for a given specified *global average linear spacing*. Let us denote the global average linear spacing as $\overline{s}(global)$ and the local average linear spacing as $\overline{s}(local)$. The global and local spacing calculations and the corresponding number of SURF points generated are shown in equations 1 and 2, respectively.

$$\overline{\#_{sp}}(global) = \frac{Area(I(global))}{(\overline{s}(global))^2} \Rightarrow \overline{s}(global) = \sqrt{\frac{\overline{\#_{sp}}(global)}{Area(I(global))}} \quad (1)$$

$$\overline{\#_{sp}}(local) = \frac{Area(I(local))}{(\overline{s}(local))^2} \Rightarrow \overline{s}(local) = \sqrt{\frac{\overline{\#_{sp}}(local)}{Area(I(local))}} \quad (2)$$

Using the mean values of the percentage distribution in each saliency slice in Table 1, above, and the known image area occupancy of the points, it is possible to calculate the values of $\overline{\#_{sp}}(local)$ for each image region. These are:

$$\begin{gathered}
\overline{\#_{sp}}(0-50\%) = 0.1233 \times \overline{\#_{sp}}(global) \\
\overline{\#_{sp}}(50-60\%) = 0.1067 \times \overline{\#_{sp}}(global) \\
\overline{\#_{sp}}(60-70\%) = 0.1033 \times \overline{\#_{sp}}(global) \\
\overline{\#_{sp}}(70-80\%) = 0.1467 \times \overline{\#_{sp}}(global) \\
\overline{\#_{sp}}(80-90\%) = 0.2100 \times \overline{\#_{sp}}(global) \\
\overline{\#_{sp}}(90-100\%) = 0.3100 \times \overline{\#_{sp}}(global)
\end{gathered}$$

Where the percentages represent the ranking of image pixels by area. Thus $90-100\%$ represents the top 10% of image pixels ranked by saliency.

Given that we know the percentage area of each slice, we can substitute this data into equation 2 to get the expected local average linear separation. Since we are interested in looking at the highest local linear density available, we choose to look at the values of $\overline{s}(90-100)$ generated for a set of possible $\overline{s}(global)$. These values can be compared with Figure 3 (Bottom Left) to choose a suitable threshold.

We are looking for the value of the $\overline{s}(global)$ value that produces a value of $\overline{s}(local(top10\%))$ that approaches the minimum allowed value. This sets the density of points in the most salient regions to be at the limit of around 11, as shown in Figure 3 (Bottom Left). Figure 3 (Bottom Right) shows plots of $\overline{s}(90-100)$ for given $\overline{s}(global)$ for mean, upper and lower values. This plot gives the average local linear spacing in the top 10% saliency slice for a given forced global linear spacing. Intersection lines are plotted showing the values of global linear spacing required to give a local linear spacing of 11 for the mean, lower and upper values of the saliency distribution. The mean value occurs at an average global linear spacing of 20 pixels. The outlier plots show that there will likely be some overclustering or underclustering from this value. The distribution of the points above, is of course, approximate but is based on real dataset overlaps for saliency categories under different viewing tasks.

Our choice of $\overline{s}(global) = 20$ should allow for strong feature point clustering around the fundamental limit for the most salient regions and a tail-off away from such regions. We can iterate the SURF threshold until the average linear point spacing of $\overline{s}(global) = 20$ is achieved within a certain margin of error. Achieving a fixed density allows for the algorithm to react to images of different physical dimensions and facilitates the joining up of the points into a probability surface.

2.2 Creating a probability surface from feature points

Having now considered the appropriate threshold to get the point density in an image we can now consider the density or cluster method that will be used to combine the interest point array into an attention probability surface. The term "probability" indicates that where the points are at the maximal density, there is a high likelihood of attention and there will be a tail-off away from these high density regions. The goal is to compare possible surfaces with real observer eye fixations and to choose parameters values to produce a map that is fit for purpose: one that we have high expectation of successfully segmenting the visually salient and background material in an image.

One method would be to count points over a local window that passes over the image and from this a sum of local densities map can be built. The disadvantage of this method is that the points need to be constrained within the window and the choice of window size is not obvious. Another method would be to use a diffusion method, such as the application of a Gaussian function to each SURF keypoint followed by their normalised summation, (see de Campos *et al.* [4]). In this case however, tuning would be required to set the Gaussian sigma value, while we wish to choose a fixed set of parameters to retain a reasonably large image area even in the case that the image is sparse in points. A better method that maintains awareness of the points at any distance is a sum of the nearest neighbour distances. The map is constructed by calculating the distance to the nearest n points for each pixel in a mask the size of the image. This approach holds the advantage of retaining the distance relationship over all nearest n points irrespective of range and normalises the attention map to the highest and lowest regions of point density irrespective of that density. This is realistic since an observer would be likely to attend the regions of greatest content variation even in uncluttered, bland images.

In order to avoid a collapse in the distance function at locations coinciding with the SURF points, the distance function excludes the nearest neighbour distance, leaving only second-nearest and above neighbour SURF points to contribute to the probability map. Equation 3 is used to calculate the map value at pixel (i,j) for the n nearest neighbours.

$$d_{2:n}(i,j) = \sum_{m=2}^{n} \sqrt{[(i - nn_m(i))^2 + (j - nn_m(j))^2]} \tag{3}$$

in which nn_m is the m^{th} nearest neighbour to location (i,j). This equation is simply the sum of the Euclidean distance from the current point in the mask to the n nearest neighbours. This gives us an inverted density map in which the low values will be near clusters of points. To convert this surface into a "probability" surface we need to normalise the data and invert the map. We normalise to the pixel values [0 1] using equation 4:

$$Norm(d_{2:n}) = \frac{d_{2:n} - min(d_{2:n})}{max(d_{2:n}) - min(d_{2:n})} \tag{4}$$

We then invert the normalised map in order to get the most-likely region attention at the highest probability using equation 5:

$$VisualInterestMap = max(Norm(d_{2:n})) + min(Norm(d_{2:n})) - Norm(d_{2:n}) \quad (5)$$

Finally, we rescale the pixel values of the map from the interval [0 1] to the interval [0.1 1] to reflect the fact that even the lowest priority region may receive some degree of attention, even if only to understand the context of the more salient material in the scene. (This is important since we are going to combine horizontal search contextualisation from object recognition later on.)

Towards the corners of an image, points will only be distributed within an arc tending downwards towards a right angle, thus forcing the corner points to have higher distance value than centre values due to the bounding of the image, thus reducing their probability. Usually, we could expect that an image would be centred towards a region of interest in the wider scene, so the extreme periphery being slightly less weighted is not a harmful thing. In fact, this property has some advantage since the map is being scaled and normalised. Since the depth of the map depends on the maximum to minimum values, the high likelihood of a dip improves the surface stability in the high-cluster regions by reducing the local depth variation.

The choice of the n value is not obvious. Since the SURF point algorithm is looking at salient local regions it is desirable that the surface as a detector of visual interest should not tail off too quickly. While in regions of approximately uniform point density the number needs to provide a stable output since the aim is to preserve content above a certain level of "probability". At boundary regions the surface must not tail-off too quickly at first, but should be allowed to collapse fast once the immediate surroundings of the boundary point is reached. Finally "hole" regions lying in between high density regions should not diminish too quickly as these regions are likely to be contextually valuable to observers scanning between highly visually salient regions around a hole. After some initial examinations it was decided to test out the viability of surfaces with nearest neighbour inclusion = 5, 10 and 15, validated against observer eye-fixations.

2.3 Segmenting the surface into high and low interest regions - choice of threshold

The proposed "Visual Interest" probability maps are two dimensional image areas with an amplitude set in the interval [0.1 1]. We want to threshold this amplitude consistently in order to segment the image into high and low regions of probability over a range of tasks. Our aim is not to tune the map to a particular task, but rather to preserve all detail that could be of interest to an analyst under different instructions. This method is viable because the SURF points are image reactive, hence the variations in N shown in (e.g.) Figure 3

(Top Left) for fixed threshold. In an image that has a large bland space, e.g. such as the sky above the horizon, the SURF points will cluster towards the detailed part of the image and leave the empty part free. In turn, this causes the surface to attain high value around the detail and low value around the bland parts that can be cut off by the threshold. Similarly, an image that has detail over a large area will retain that detail above the same threshold level.

Figure 2 is an illustration of this effect on two images chosen from the Torralba *et al.* dataset with NN=10 used in the surface construction. In this example, the SURF threshold used is 1.5, this being the value that gives a global linear spacing of 20 pixels over the dataset (refer to Figure 3 (Bottom Left)), and the surface threshold applied is 0.825 (from the interval [0.1 1]). The image on the left generates 3553 SURF points and the image on the right generates 1308 SURF points, both taken at the fixed SURF threshold of 1.5. The top row shows that the SURF points are clearly clustering towards the detail in the image. The SURF application threshold is fixed and so the fact that the number of points in each image is different is attributable to the image content. The second row shows the derived visual attention 'probability" map with the surf points superimposed. The surfaces tail-off away from the most dense regions. The third row shows the thresholded probability maps using the value 0.825 (the maps have values [0.1 1]). The white part is above threshold, black below. The left image has 75.6% area above threshold and the right image has 70.95% above threshold. The last row shows the above threshold areas highlighted in the images. This shows that the surface is acting as desired. The detailed regions in the left image have been brought out pretty evenly. In the right image the clusters of SURF points e.g. around the windows have been joined to each other above threshold leading to an elegant segmentation.

Figure 2 is an illustration. What we now do is validate the choice of surface threshold using thorough observer testing to justify the choice of a general pair of parameters, NN and the segmentation threshold, that can be applied to general, eye-level imagery.

2.4 Validation of the Parameters

We want a segmentation that can accurately select image regions likely to attract visual attention over many different tasks. We must therefore ultimately gauge the performance of our surface against observers performing different tasks. To do this we used eye-tracker data by observers performing different search tasks and analysed how their fixations coincided with our generated probability surfaces (NN=5,10,15) at different segmentation thresholds.

Another measure of the surface viability is to compare the surface with the ground truth information in images. Object-category objects consistent with the observer tests were manually extracted from images and coincidence of the probability map with the ground truth data was recorded.

Finally, since the surface is designed to take advantage of the content-adaptive nature of the SURF algorithm, the area above each threshold for each image was collected.

This data allows for judgement to be made on how effective the algorithm is at capturing observer eye-fixations and task-relevant objects for a given nearest neighbour count and cut-off threshold. This can then be compared with the background area that each threshold provides. The idea is to choose a combination of the nearest neighbour and threshold that will detect task-directed eye fixations from many tasks while excluding a substantial image area.

2.5 Experimental Procedure

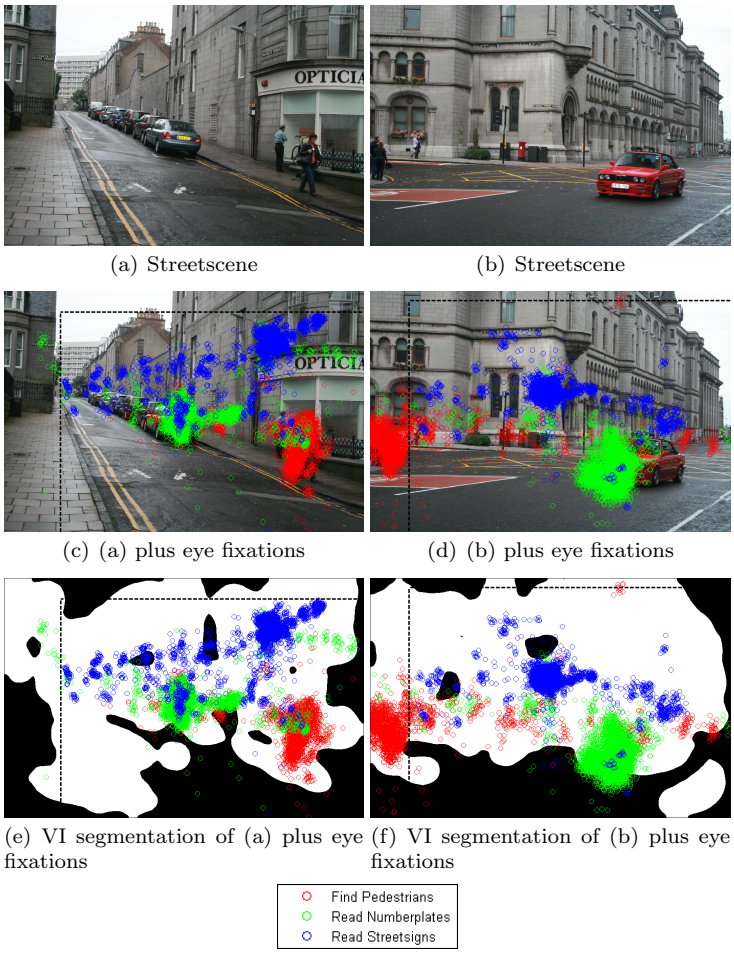
The visual attention probability maps (NN = 5, 10, 15) were designed to segment important regions for *many tasks* in eye-level imagery. We therefore built a dataset of images suitable for performing multiple tasks. The final dataset consisted of 36 images taken at street level in the city of Aberdeen. The images were collected with multiple tasks in mind. Each image had at least one "object" present in the scene of the following class: numberplate, streetname sign, pedestrian carrying an object.

The formal tasks chosen for the experiments were 1) find the street names and read the letters. 2) find the vehicle registrations and read the letters. 3) find a previously-seen cut-out of a "person" in the full image and identify what they are carrying. This led to 108 tasks over the image set.

The nature of the tasking is perhaps not too important, but the tasks chosen represent normal surveillance tasks within an urban environment at eye-level. The tasks chosen focus the observer attention towards the objects of interest, guided by the scene context. The first two tasks provide hard performance data on the ability of the observers to read alphanumerics on the original images. The third task provides some level of discrimination ability of the observers. The eye-tracker data cannot tell us about human performance, only about fixation location. The data collected was used to check for participant consistency within the group of observers.

The experimental set up consisted of a high resolution screen (1200×1600 pixels, but note that the original SLR had 2592×3888 pixels) placed at a distance from the observer of 60cm, such that the angular resolution of the screen was one pixel. A head mount was employed to keep the position of the head bounded during the experiment. A Tobii eye-tracker was set beneath the screen, pointing towards the head mount.

The images were scaled for each task. The images for the alphanumeric readability tasks were scaled to 3.5 cycles (line pairs) across each letter, just above the supposed limit of resolution for reading letters displayed orthonormally to the sensor of 3 cycles [19]. The images for the "find and identify" task set were scaled to fit on the screen.



(a) Streetscene (b) Streetscene

(c) (a) plus eye fixations (d) (b) plus eye fixations

(e) VI segmentation of (a) plus eye fixations (f) VI segmentation of (b) plus eye fixations

Fig. 4 Task directed viewing: Top row shows two cluttered street scenes with pedestrians, car registration plates and street signs. Centre row shows the eye fixations of observers performing multiple tasks on these images: red: count pedestrians, green: read car registrations, blue: read streetsigns. (The alphanumerics were scales to 3.5 cycles over their width. The black dashed line delimits the part of the image displayed for the read streetsign task.) Note the strong general horizontal distribution of eye fixations for each given task and also note the strong differences between the distribution of the eye fixations for each task. The bottom row shows the VI segmentation's ability to pick up fixations relevant to multiple observer tasks.

Each participant was shown a 36 image slideshow per task in random order. They were instructed to carry out the task in silence at first. They would then press a keyboard button and they would then give their response orally which was written down by an overseer. The button press removed the image from the screen and presented a mid-intensity grey level image to the observer with a fixation cross in the centre. During the time that the image was present on the screen eye-fixations were recorded. The eye-fixations are an interpolated

average of the saccade patterns of the eyes. The eye-tracker was calibrated to record a fixation if the eyes fixated around an area for 30 ms within a radius of 30 pixels. The differently scaled images for the next task were then assigned a random order and the next task was analysed.

In total there were 7 participants that carried out the 108 tasks, giving valuable information, both about task performance and the associated eye-tracker fixations. Each participant was given a bottle of wine after the experiment in thanks for their contribution. During post-experimental discussion, the participants all expressed surprise that the image set only consisted of 36 images, albeit scaled for 108 viewing tasks. The effect of scaling and task-directing of attention was clearly such that the non-task relevant parts of the image were not attended or registered by observers for many of the images presented. Thus, even though the images were repeated in content, if not in scale and quadrant, the observers did not notice that the images had repeated themselves, implying that they were very task focussed during the experiment.

In addition to our own dataset, we used that of Torralba *et al.* which is supplied with eye-tracker data of participants performing object count in three tasks. 36 images have eye fixation points from 8 observers performing a "count people" task and 36 images have eye-fixation points from 8 observers performing a "count paintings" task and a "count cups" task, leading to 108 tasks by 8 participants.

In total the eye fixation data available contained 6 tasks. Each observer performed each task over a set of 36 images.

2.6 Observer performance

The observer group for the Aberdeen dataset were very consistent with their objective task *performance* in the "find and identify object" task and the "find and read streetnames". This is perhaps to be expected, since there is a degree of prior knowledge about what objects are likely to be carried and inference regarding the identity of obscure letters can be made if the overall word in a streetsign can be comprehended. There was a fall-off in performance for some of the observers for the "find and read numberplate" task, which can be explained by the complication of distinguishing pseudo-random characters close to the resolution limit. At the resolution limit it is expected that there will be a range of participants that will fail to read the numbers successfully due to information-smearing and confusion generated between generally similar characters such as for example "a" and "o", especially in a bland font such as that used in British numberplates. Since the data was not ortho-rectified to the camera head it is possible that the true resolution limit for some numberplates was far above 3.5 cycles for some observers.

The readability results only differed significantly for the numberplates. Each participant had to read 285 characters in the 36 image set. For the purposes of assessing readability, there were two error classes defined. The first was a minor error; that is a failure to distinguish between characters that are

similar in structure[1]. This class of error was awarded 0.5 counts, since character confusion and contextual information (such as car colour) could be taken into account in a more detailed study later without too much trouble. The second was a major error: that is, omission of a character or the misclassification of normally distinguishable characters. This class of error resulted in 1 count. The cumulative errors of each observer were summed to give a total error for each of the seven experimental participants. These total cumulative error scores were then averaged and the standard deviation calculated. The error rate for the set of participants was 45.4 ± 21.7 out of a total of 285 alphanumeric characters. This equates to an error rate of $15.9\% \pm 7.61\%$

All participants were able to detect the numberplates successfully and read notably more than 50% of the letters sucessfully. This was the expected outcome by setting the cross-character resolution to 3.5 cycles rather than the theoretical limit of 3 cycles [19].

Overall, the participants performance levels indicated that they understood the instructions of the test and performed the test in a fashion consistent with each other.

2.7 Probability maps vs. human attention from eye-tracker data

For each individual image presented to the observers a SURF based probability map was calculated, with the parameters as previously discussed. The SURF threshold was adjusted by iteration until the average global spacing equalled 20 ± 0.1. Where the image had been scaled and cropped for display, the probability map had the same scaling and cropping applied.

The probability maps had values in the range [0.1 1] and the test was to find out how well the probability maps at each level of threshold predicted observer eye-fixations under the different task conditions.

For each *task* slide the eye-fixation coordinates collected from *all* participants were used. Due to the scaling of the images to the above specified resolution some images were smaller and some images were larger than the dimensions of the screen. In the case that the images were larger than the display screen, only fixations within the screen were counted. For images that were smaller than the screen size, only eye-fixations centred within the boundary of the image were counted. There were eye-fixations recorded outwith the image boundaries. This could have been due to a natural lapse of concentration from the participant or from the 30 pixel accuracy of the equipment.

There were three possible classifications for the eye-tracker points. 1. On screen, on image. 2. On screen, off image (in the case where the image is smaller than the screensize). 3. Off screen (e.g. participant blinking). 95% of all eye fixations lay within the image boundary on the screen and only these points were included. A demonstration of the display and the collected eye-tracker

[1] Such confusions could arise from letters at low resolution including the following sets of confusions: $F \Leftrightarrow R$, $W \Leftrightarrow M$, $W \Leftrightarrow N$, $G \Leftrightarrow 6$, $G \Leftrightarrow C$, $B \Leftrightarrow 8$, $V \Leftrightarrow Y$, $X \Leftrightarrow A$, $H \Leftrightarrow K$, $5 \Leftrightarrow 6$, $F \Leftrightarrow P$, $H \Leftrightarrow A$, $G \Leftrightarrow D$, $O \Leftrightarrow D$, $B \Leftrightarrow E$, $6 \Leftrightarrow 8$ and $G \Leftrightarrow 6$

points for all seven observers for the three tasks in the Aberdeen set is shown for a pair of images in Figure 4. Note the consistency of the results between the participants, shown by the close clustering of the fixation points for each task. Note also the wide difference in the distribution of the fixations for each task. Finally, see the ability of the VI segmentation to capture image regions relevant to multiple tasks.

For the attention map of each image for the task, the threshold levels were tuned over the whole threshold range of [0 1] at intervals of 0.01 and the percentage of inlying/outlying eye-fixations was recorded. (Note that the maps were set to the interval [0.1 1], so the perfect overlap in the interval [0 0.1] is unremarkable.)

The data is output as ROC curves in Figure 5 showing the overlap between the task eye-fixations and the surfaces thresholded to different degrees, for both the Aberdeen and the Torralba datasets. The surfaces are all set to between [0.1 1], but for ease of display this has been multiplied by 100 on the x-axis. We only include NN = 10 here, since this value was the overall best fit.

The task ROC curves are all similar in profile. Notably the two indoor tasks have poor performance for a given threshold relative to he other tasks which are performed in an outdoor environment. Typically, indoor images are considerably more cluttered and since the tasks on the indoor set were open ended "count object" tasks, the search pattern for the inside scene is more extensive with people looking in more and more unlikely locations to complete the task. In contrast, the outdoor tasks are consistent: there is less clutter outdoors and objects and their contexts for this task set are obviously better defined. Accordingly, the accuracy of eye fixation prediction is dependent upon the difficultly of the task in hand. In our outdoor scenes, the objects of interest are not generally hidden in confusing clutter. Accordingly, if we set our threshold at 0.825 we have the expectation of capturing 85-95% of eye fixations for outdoor scenes and possibly a bit less for structured indoor scenes.

2.8 Probability maps vs. object position ground truth

The other key performance measure of the surface is its overlap with potential objects of interest. To this end all numberplates, streenames and pedestrians in clear view were manually extracted from the Aberdeen dataset. (The Torralba dataset does not have multi-category objects in every image and not every image has the object present.) In total, 46 streetsigns, 50 numberplates and 122 pedestrians were extracted over the 36 images. The mean percentage overlaps of these object categories lying within the surface NN=10 at each threshold is shown in Figure 6.

2.9 Percentage area thresholded for given area

As illustrated in Figure 2, the area occupied by the surfaces varies by design for a given surface threshold. This area occupancy is relevant to image compression

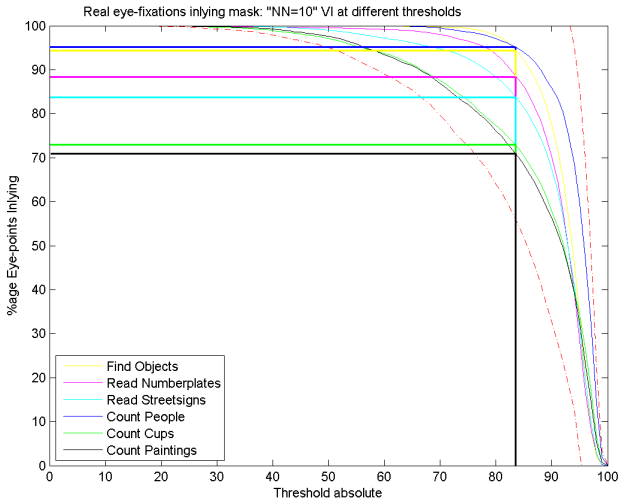


Fig. 5 Overlap of eye fixations collected for all tasks at different surface thresholds of the proposed visual interest surface, using NN = 10. Intersection lines at a fixed threshold of 0.825 are plotted to show the percentage of eye fixations counted above threshold for each task. The red dot-dash lines are the upper and lower standard deviation limits for *all* tasks.

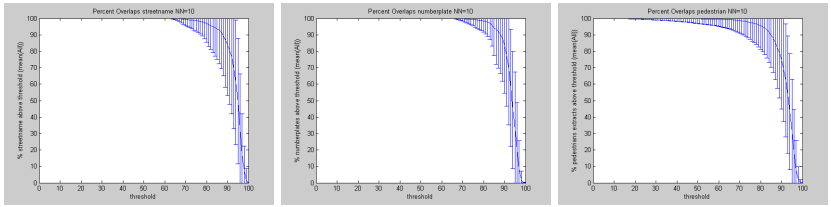


Fig. 6 Overlap of surfaces (NN = 10) at different thresholds with objects manually extracted from the "Aberdeen" dataset. Objects Extracted from left to right: Streetsigns, Numberplates, Pedestrians. Whole objects extracted where possible.

since the more image area that is expendable, the greater the potential savings that can be made. The percentage area from across the two image sets is shown in Figure 7.

2.10 Interpretation

Figure 5 shows the overlaps of the task eye-fixations with the thresholded new surfaces with NN = 10. (The overlaps are similar but slightly inferior for NN = 5 and 15.) The shapes of these responses are excellent for our purposes and perform consistently at predicting eye-fixation regions over the multiple tasks. The approximate turning point where the eye-fixation prediction power starts to collapse is at a threshold of around 0.8 where approximately 90% of eye-fixations are included.

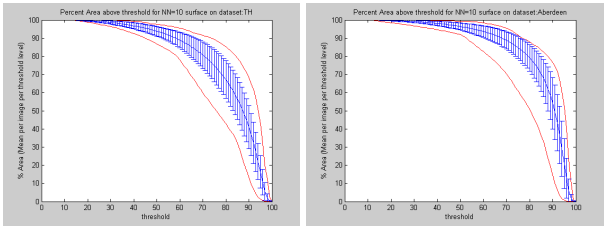
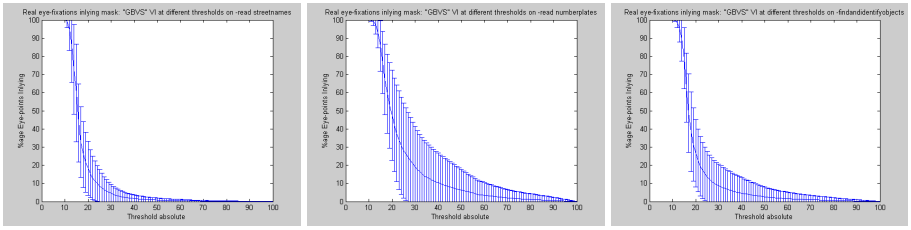


Fig. 7 Mean percentage of image areas above threshold for given thresholds and NN = 10. The error bars show the standard deviation of the image sets, while the red lines indicate the absolute minima and maxima of the image set.

The ground truth data shown in Figure 6 shows the mean area of manually extracted objects in the Aberdeen dataset lying above threshold at all visual interest threshold levels. Again, the Visual Interest surface performs well with the approximate turning point being at a threshold of circa 0.8 where 95% of object areas over the set are preserved. Note that not all of an object needs to be seen for an observer to classify it correctly.

Finally, Figure 7 shows the percentage of image areas lying above threshold for the NN = 10 surface acting on the "Torralba" (left) and "Aberdeen" (right) image sets. This is very important information because the smaller the area of the image that can be considered potentially of interest under analysis, the greater the expendable area. The algorithm performs differently for each image set, but again at around same turning point of approximately threshold = 0.8, the surfaces exclude approximately 20% of image area. This is however image content adaptive, with variation between 10 and 40% of image area potentially excludable at the extremes of the Aberdeen dataset. The fast fall-off beyond threshold = 0.8 offers opportunity to increase the excluded area sharply, although at the risk of losing eye-fixation points and therefore regions of analyst interest.

Not all eye-fixation points need to be included in the core, of course. Many are contextual or transient fixations collected as a part of the tasking process. It is feasible that these fixations could pick up the necessary contextual detail on a reduced-content representation of the out-of-core image regions. If the task is very focused, the narrow field of view of the human visual system is likely to miss such lack of detail as long as it does not in itself become distracting in the image context. On this basis we are interested in the range of surface thresholds [0.75 0.875] and will perform subsequent validation at a threshold of 0.825, which is a conservative threshold that is chosen empirically as one of the last points of slow variation so that we can expect a good trade off between eye fixation prediction (Figure 5), object location inclusion (Figure 6) and selected background area (Figure 7). The results in the NN=5, NN=10 and NN=15 sets were similar, so we choose the set NN=10. Thus we now have fixed the parameters of the segmentation algorithm "Visual Interest" (or VI) which are: mean linear spacing = 20, number of nearest neighbours = 10, Threshold for segmentation = 0.825.



(a) Read Streetnames (b) Read Numberplates (c) Find & Identify

Fig. 8 Mean percentage of overlapping task-directed eye fixations for a given threshold of GBVS maps of all the images in the Aberdeen dataset. The error bars show the standard deviation of the counts between images.

At this point it is worth comparing the fixation-capture of a bottom-up attention map at given thresholds, such as GBVS. Figure 8 shows the percentage of captured fixations for a given threshold of GBVS for the three sets of task-directed eye fixations recorded for the Aberdeen dataset. (The same test using the algorithm of Itti *et al.* gives similar plots.) Note the shape of the curves in comparison with the Visual Interest overlaps shown in Figure 5: GBVS has an overlap convexity that is the opposite of the Visual Interest algorithm. In Figure 5, the thresholded map has an initial long flat region with a wide turning-point region, allowing for a threshold value to be selected that, by expectation, should capture a large proportion of the eye fixations in an image of a similar scene. In contrast, Figure 8 demonstrates a sharp initial decline in eye fixation capture, with the turning point region situated between threshold 20 to 30 only capturing around 10 to 15% of all fixations. These two Figures show the difference between a typical bottom-up model and the Visual Interest algorithm. The bottom up models typically generate sharp-peaked maps tuned towards finding the "first few" passive fixations, while the Visual Interest algorithm generates a flatter attention map that can be thresholded to give a high expectation of selecting image regions relevant to multiple tasks.

2.11 Summary of the "visual interest" surface analysis

We have proposed, constructed and tested a "Visual Interest" surface from SURF points. By experimentation we have found that it is possible to find a threshold that can be set to predict attended regions of an image accurately under multi-tasking. We find that for our chosen parameters, a surface threshold of 0.825 will detect nearly 90% of all observer eye fixations while leaving approximately 20% of image area classified as "background". This allows for the possibility of high levels of compression to take place over these regions, or for us to reduce the search space in automatic target recognition. The algorithm "Visual Interest" is not dependent upon further tuning or parameters.

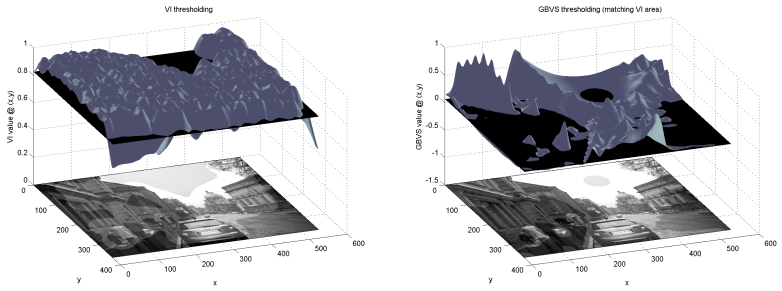


Fig. 9 A demonstration of Visual Interest's advantages relative to the state of art. *Left:* A 3d visualisation of the VI probability map applied to an image from the "Aberdeen" set. Threshold intersection shown giving 38% background. *Right:* A 3d visualisation of the GBVS probability map (thresholded to give equal excluded area to VI). Note that the GBVS algorithm is selecting salient material relevant to *some* tasks but it is retaining the sky as salient while missing the people. In contrast, the VI segmentation successfully retains a lot more detail in the image. (See Figure 1 for segmented image examples.)

With the parameters fixed based on the eye-tracker information, it is now ready for final validation based on observer experiments.

The 'VI' output is a general algorithm for the detection of *potentially salient* image material. The closest comparison that this algorithm has are bottom up models, since it reacts only to the image content and requires no foreknowledge of image content. However, the advantage of this technique is that the algorithm selects information relevant to both the bottom up problem *and* task performance and requires no prior learning, in contrast to the state-of-art in task modelling [29]. This is illustrated in Figures 1 and 9 comparing GBVS and VI.

2.12 Validating the segmentation algorithm using compression

We validate the VI segmentation using observer testing on compressed imagery.

We segment original images to context-compress them from general sensors and compare the utility of these "Visual Interest" treated images against global compression in observer experiments. The technique proposed is to pre-filter the background regions before compression using a degree of Gaussian blur. This approach was taken by Itti for video analysis in [10]. This operation destroys high frequency information in an image and allows for efficiencies in coding since a bland region is relatively easier to encode than a busy one. It is also a suitable way of degrading the image for human interpretation since the application of a Gaussian blur effectively changes the viewing scale of the image [15,14]: thus is is a "natural" distortion which should cause little distraction to the observer, for example through artefacts.

For passive viewing of video, typically with a narrow attentional focus, a sharply-peaked bottom-up attention map is probably adequate for ROI com-

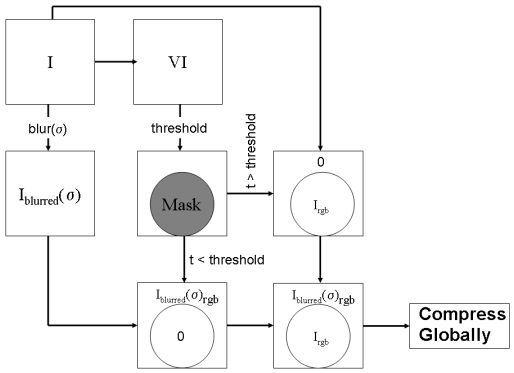


Fig. 10 Compression circuit for adding Gaussian blur to the images. The parts of the image above threshold are kept at original resolution. The parts of the image below threshold are blurred to degree sigma. The segmented regions are then combined into a single image before compression.

pression. We note, though, that using a bottom-up attention model for still-frame compression is a poor idea. The multi-task eye fixation capture expectation of circa 15% at the turning points in Figure 8 shows that if we utilise a bottom-up model for compression, we are very unlikely to keep image areas related to task. If we choose a lower threshold in order to capture more fixations, we will get a much more varied output than VI because of the steepness of the slope. Accordingly we will not provide comparison of the VI compression using the bottom-up models since we risk missing a large proportion of eye fixations.

The ex-ROI blur method is chosen for the final validation of the segmentation algorithm in combination with JPEG. This is the natural algorithm to choose since it is the one commonly integrated into the capture process for single-shot colour imagers and because the input to the algorithm is related to the output[2]. Background blur will also identify segmentation failures: if the segmentation fails to detect the salient detail, the blur will render the detail impossible to process by an observer. In contrast, due to the block nature of the JPEG algorithm, the blocks lying fully within the core are kept at *exactly* the same fidelity as if the JPEG algorithm were to be applied to the un-blurred image. The boundary region blocks will be an average of the blurred and high-fidelity information lying in the block.

We use the circuit in Figure 10 to compress our images. The threshold in VI is fixed at 0.825. The other parameter is σ, effectively the degree of Gaussian blur to be applied to the image. Clearly a higher degree of blur $\sigma+1$ will naturally be more efficient to encode than degree σ since there is less

[2] There is a region of interest (ROI) capability in JPEG2000, but the parameters for the JPEG2000 algorithm are not related to the output quality. In contrast, the JPEG algorithm was designed using data from observer tests: a Q value of 50 is expected to produce good visual quality for photo-real imagery.

statistical "information" within the image. We choose $\sigma = 8$ as a value that should smooth out the high frequencies but allow for the background context to be interpreted if an observer attends the blurred regions.

2.13 Practical Implementation

We apply "Visual Interest" blindly to *unknown* data from *different* sensors and test observer performance. We further utilise object recognition from interest points to contextualises the energy in the attention map to the horizontal region about the recognised objects, such as in the approach of [29].

The blind test data (that is, not previously analysed by the algorithm) consisted of 18 images collected from three different cameras. Six images of Edinburgh at street level were collected from each of three sensors, a Canon "EOS400D", a Sony "Cybershot" and a Nikkon "Coolpix5000" each on their standard general settings. In each image were alphanumeric objects for observers to read. The alphanumerics in the image were scaled down to a range of cross-letter cycle values so as to be across the limit of resolution, such as indicated in [19]. The values examined were 2.5, 3.0, 3.5, 4.0 cycles across the letter width. The VI segmentation is computed using the procedure and parameters defined above (see Section 2.11. Note that we apply Gaussian blur $\sigma = 8$ to the parts of the image outwith the VI region and maintain the above-threshold parts at the original resolution.

A further category was constructed based around the object-matching task contextualisation. The technique used here is object recognition from previously-seen objects, based upon the distribution of matched surf points between test and reference images. This object recognition technique is shown in [15]. The SURF-point descriptors were checked against a database of SURF point descriptors pre-recorded over the objects of interest in the images (i.e. the fronts of cars around the alphanumerics in the car numberplates). This is far from being the most robust object recognition technique, but we used manually extracted, head-on objects from the test images and the object identification error rate was zero over the image set. This object recognition technique suits the validation application of this work, but more general object recognition schemes (such as those described by Viola and Jones [31] or Fergus [5]) could be applied to seek more general object classes.

In the event of multiple point matching for a given image, the centroid of the matched points was taken as the centre point of the horizontal context for the image class. A horizontal bar is then blended with the VI attention probability surface to focus probability to the horizontal region about the recognised object. This works because ground-level images usually have like objects (or object classes) distributed horizontally across the image. Where an object is detected, an object context "task" map is constructed that consists of a bar the width of the image and 1/3 height centred around the matched object point centroid, tailing off at the vertical edges away from the bar. For the blending of the data, we use the same approach as Torralba et al. in [29].

The Visual Interest surface was generated as previously but the values of the VI map and the object contextualisation bar are re-scaled to the intervals [0.32 0.95] and [0.93 0.95], respectively, to allow for a good blending. These maps are then blended together to give a VI+Object attention map using equation 6, with $\alpha = 0.03$.

$$VI + Object = VI^{\alpha} + ObjectBar \tag{6}$$

A comparison of the general VI with the VI+OBJ segmentation is shown in Figure 11 using cars as the reference objects stored as an array of descriptors. Note the reduction in salient area when object contextualisation is exploited. This figure shows the trade off involved in applying object contextualisation *from a previously-seen object*. This scheme is similar to that described in [15] and in [18], involving an assessment of point to point matching along with a check of the overlap of point distribution (e.g. through Homography). While in the object recognition case shown in Figure 11 some detail is destroyed in the image, this does not matter if the purpose of the image is restricted to observer analysis of cars or numberplates. The trade off is filesize. In this example the Original is stored at 0.66 *bpp*, (j) at 0.62 *bpp* and (k) at 0.55 *bpp*. The alphanumerics on the car numberplate are scaled to 3 cycles across the letter width. (c.f. Table 2, listing achieved *bpp* for given resolution scaling)

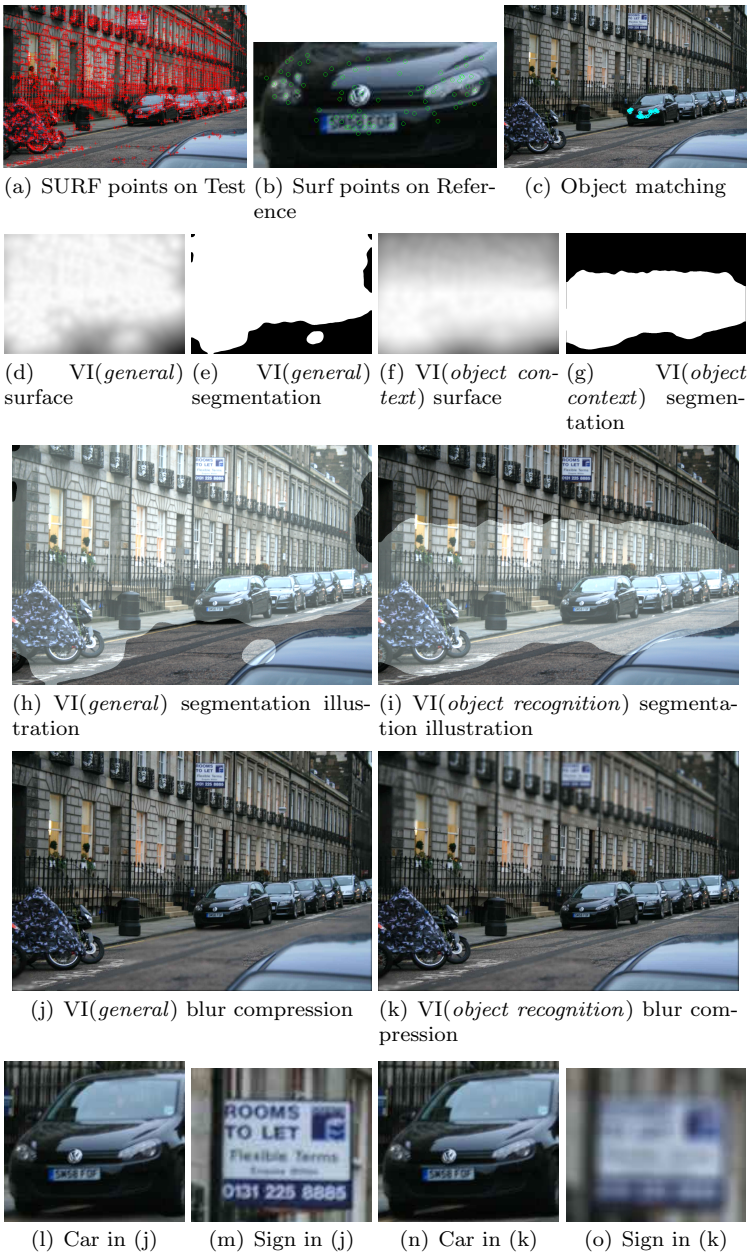
The whole set of 18 images was processed as described above, over four different resolution scales: number of cycles across the letters was set to 2.5, 3.0, 3.5 and 4.0. The compression *bpp* statistics were collected and are entabulated in tables 2 and 3.

Examination of the tables shows that applying the VI blur circuit results in a saving in image filesize of approximately 15% relative to global JPEGQ40[3] and approximately 15% relative to global JPEGQ50. Applying the VI+OBJ blur circuit results in approximately 25% relative to global JPEGQ40 and approximately 25% relative to global JPEGQ50.

2.14 Observer Performance

Seven observers were asked to read the alphanumeric image data (numberplates, streetname signs) within the image set. The observers were shown the data at NCycles = 2.5, 3.0, 3.5, & 4.0. The observer responses from the Global_JPEG(Q40,Q50) images were then recorded as were the observer responses to VI_JPEG(Q40,Q50) and VI+OBJ_JPEG(Q40,Q50). The images were presented in a random order, disallowing for sequential repetition of the same image at different compression or resolution treatment. The threshold for minimum reliable readability is usually 3.0 cycles for ortho-rectified data. We aimed for high readability at N = 3.5 cycles, as shown to be possible above, given the offset nature of our dataset. At that level of resolution and above, the reading performance of observers viewing the global schemes (JPEG(Q40,

[3] Read as: *JPEG with quality level, Q = 40.*



(a) SURF points on Test (b) Surf points on Reference (c) Object matching

(d) VI(*general*) surface (e) VI(*general*) segmentation (f) VI(*object context*) surface (g) VI(*object context*) segmentation

(h) VI(*general*) segmentation illustration (i) VI(*object recognition*) segmentation illustration

(j) VI(*general*) blur compression (k) VI(*object recognition*) blur compression

(l) Car in (j) (m) Sign in (j) (n) Car in (k) (o) Sign in (k)

Fig. 11 Selective content preservation through compression: fixed compression algorithm parameters. (a): a Test image with SURF points. (b): a Reference image stored in a database. (c): matched Test and Reference descriptors showing object recognition. (d), (e) and (f) and (g) show the VI(*General*) and VI(*Object recognition*) surface and segmentation respectively. The horizontal context around the detected object is prioritised. (h) and (i) illustrate the segmentation. (j) and (k) show a gaussian preblur with $\sigma = 8$ applied to the background regions prior to JPEG Q40. Image pairs (l) & (m) and (n) & (o) show the preservation of different objects in the scene: information in the sign above the car is preserved by the general VI but destroyed as a part of the blur process in the object case.

Q50)) against the two visually salient schemes was exactly the same. This results because the alphanumeric information in the images in these cases was lying above threshold and therefore the difference between the task-relevant sections in the contextualised and the global images is statistically (and thus also perceptually) zero. Note that each observer had an *individual* task performance level with some hesitation between similar alphanumerics, but that there was no fall off in performance for the same observer between performing the task on the global or segment-blurred images. The means that each observer achieved the same task performance on the images, even though the treated images were of lower filesize. Below N = 3.5 cycles, the levels of performance were maintained, provided that the alphanumerics remained above threshold. For the "VI-only" binary compressed images this was the case for the clear majority of images: for 2.5 and 3.0 cycles only 1/18 numberplates in the images was illegible due to the blur. For the object contextualisation add-on, this problem was resolved and all of the alphanumerics were above threshold with observer performance levels accordingly equal to that achieved on the globally compressed images.

While noting that the performance of observers *below the theoretical resolution limit* has some minor errors, the important factor was that the performance of the observers on the regionally-treated imagery was *identical* to that on global imagery on alphanumeric objects above the resolution limit. This allows for substantial advantage to be gained at a small risk of performance cost. As alphanumeric objects are scaled above the task resolution limit, the chance of their inclusion above threshold becomes considerably higher, due to the larger number of likely features surrounding high frequency regions such as high contrast alphanumeric data.

NCycles	Original *bpp*	JPEGQ40 *bpp*	JPEGQ40(VI) *bpp*	JPEGQ40(VI+OBJ) *bpp*
2.5	1.6015 ± 0.1332	0.7432 ± 0.1888	0.6440 ± 0.1368	0.5688 ± 0.1224
3.0	1.5587 ± 0.1317	0.6976 ± 0.1768	0.6072 ± 0.1264	0.5400 ± 0.1168
3.5	1.5211 ± 0.1316	0.6600 ± 0.1672	0.5776 ± 0.1176	0.5144 ± 0.1072
4.0	1.4867 ± 0.1328	0.6280 ± 0.1592	0.5576 ± 0.1184	0.4952 ± 0.1032

Table 2 Table of bpp values obtained using JPEGQ40 and the Visual Interest techniques with VI($\sigma = 8, t = 82.5$). NCycles refers to the number of cycles across the alphanumerics in the test images.

NCycles	Original *bpp*	JPEGQ50 *bpp*	JPEGQ50(VI) *bpp*	JPEGQ50(VI+OBJ) *bpp*
2.5	1.6015 ± 0.1332	0.8536 ± 0.2112	0.7336 ± 0.1528	0.6480 ± 0.1376
3.0	1.5587 ± 0.1317	0.8176 ± 0.1984	0.6912 ± 0.1408	0.6152 ± 0.1312
3.5	1.5211 ± 0.1316	0.7592 ± 0.1880	0.6584 ± 0.1312	0.5864 ± 0.1200
4.0	1.4867 ± 0.1328	0.7232 ± 0.1792	0.6360 ± 0.1328	0.5648 ± 0.1168

Table 3 Table of bpp values obtained using JPEGQ50 and the Visual Interest techniques with VI($\sigma = 8, t = 82.5$). NCycles refers to the number of cycles across the alphanumerics in the test images.

2.15 Validation Conclusions

In conclusion, a method to segment images which discriminates between salient and contextual information has been presented. It operates blindly as a single pass algorithm and successfully predicts eye-fixations and objects in real world scenes under different task conditions. It is based on feature points and the descriptors from the feature points can be used for database matching with stored object representations to refine the salient area in the case of object class search based on a-priori knowledge of task based search mechanisms. This segmentation has been used successfully to save *bpp* in compression through using it to control a pre-processing contextual blur filter with *no* observer performance tail-off for alphanumerics above the limit of resolution. The recorded filesizes achieved were on average 15% less than the global case using the interest point only technique and 25% less than the global case when the technique was combined with object contextualisation.

3 Application of the Visual Interest segmentation algorithm to JPEG2000

Here, the intention is to use the validated Visual Interest segmentation in its "general" mode to compress known computer vision datasets, with selective preservation of the VI-selected region of interest to improve image quality over the visually salient area relevant to multiple tasks. The JPEG2000 algorithm with its built in ROI encoder is a natural candidate for combination with the Visual Interest segmentation algorithm.

This is in fact not as straightforward a procedure as might be imagined. JPEG2000 is a highly parameterised algorithm with a huge number of inputs. Rather than the straightforward "quality level" input of JPEG, in JPEG2000 the main input is a "rate" parameter which seeks a specified number of *bpp*. The algorithm can also be built up in a specified number of code "layers" of wavelet-encoded information in frequency sub-bands. The other major parameter is that of the resolution "levels" of the data.

Additionally, there are two methods for implementing the region of interest in the JPEG2000 algorithm. The "maxshift" method offers a strong region definition and potentially higher gains. However, the disadvantage of this method is that it relies on a scaling factor to reduce the foreground quantisation step sizes relative to the background. This factor must be very large (circa $2^{12} = 4096$ and for much real data, some compliant encoders may not recover any background at all for the method. The alternative method is to weight the cost function code-block contributions in accordance with the ROI. This does allow for data transmission for regions lying outwith the ROI but the regions definition is notably poorer. It is advised to used code-blocks of only 32 by 32 wavelet coefficients, rather than the standard 64 by 64 while using ROI compression, since adjustments can only be made on a code-block by code-block basis [26].

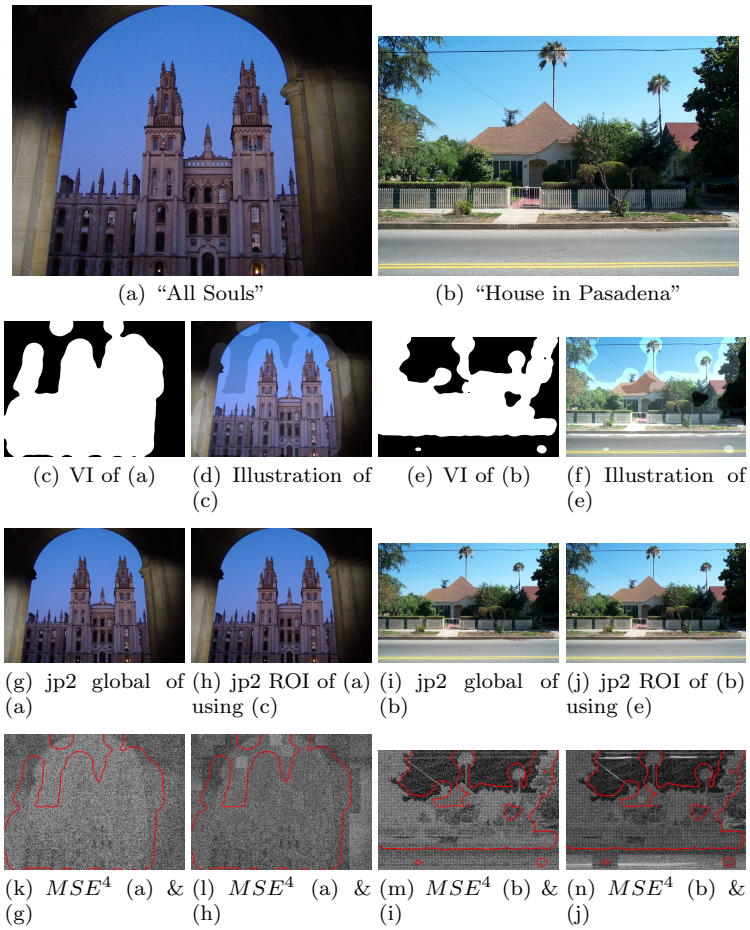
Here, the proposal is to feed the visual interest segmentation into the JPEG 2000 architecture to improve compression performance using the second method of ROI encoding. This method will not keep rigorously to the input ROI, but it can be expected to use the ROI as a strong guide to the overall compression process. This is in fact an advantage over the blur process used previously in certain respects: *both* the Visual Interest algorithm and the JPEG2000 algorithm react to the image content and in the event that the Visual Interest algorithm fails to detect some salient detail, the JPEG2000 algorithm running in this cost-function weighting ROI mode may still retain it at suitable definition, in contrast to our prior validation process in which misclassification destroys sub-resolution alphanumerics through the blurring process. We choose a suitable number of layers and levels to be shared in common between the global and ROI implementations and measure two quality measures, Peak Signal to Noise Ratio (PSNR [30]) and Structural SIMilarity (SSIM [32]) between the original image and both the global and ROI implementations while reducing the "rate" parameter.

The experiment here uses two publicly-available datasets: a set of houses in Pasadena, California provided by the Vision group at California Institute of Technology [1] and pictures of the architecture of All Souls college, Oxford, provided by the Visual Geometry Group of Oxford University [23]. All of the images are saved as JPEG files and do not contain any information regarding the collection parameters. The Pasadena dataset is composed of 241 colour images of dimensions 1168×1760 pixels with an average *bpp* storage size value of 0.3027 ± 0.0588 and the All Souls set comprises 129 colour images of dimensions 768×1024 with an average *bpp* value of 0.5182 ± 0.2001.

An illustrative example of the JPEG2000 images set to a rate of 0.75 is presented in Figure 12. Note the high quality of the images produced and note the difference images between the global- and ROI-processed images: the regions treated are "squared" off due to the poor regions definition available by the code block weighting ROI, but selective preservation of the salient regions has occurred. (See also Figure 13.)

3.1 Experimental procedure

The JPEG2000 ROI performance data is collected using the following procedure for both the Pasadena and All Souls datasets. The rates used (relative to the original JPEG image filesizes) are: 1, 0.9, 0.75, 0.5, 0.25, 0.125. (We find that PSNR and SSIM do not improve at rates lower than 0.25.) For each image in the dataset we (automatically) create a JPEG2000 file from the original image with the chosen "rate", determined by the factors above. We then compute the Visual Interest segmentation for the Original image and apply the JPEG2000 ROI algorithm using the Visual Interest segmentation. Then we measure the PSNR and the SSIM between the Original image and the *global* and the Original image and the *ROI-compressed* images over the pixels in the ROI.



(a) "All Souls" (b) "House in Pasadena"

(c) VI of (a) (d) Illustration of (c) (e) VI of (b) (f) Illustration of (e)

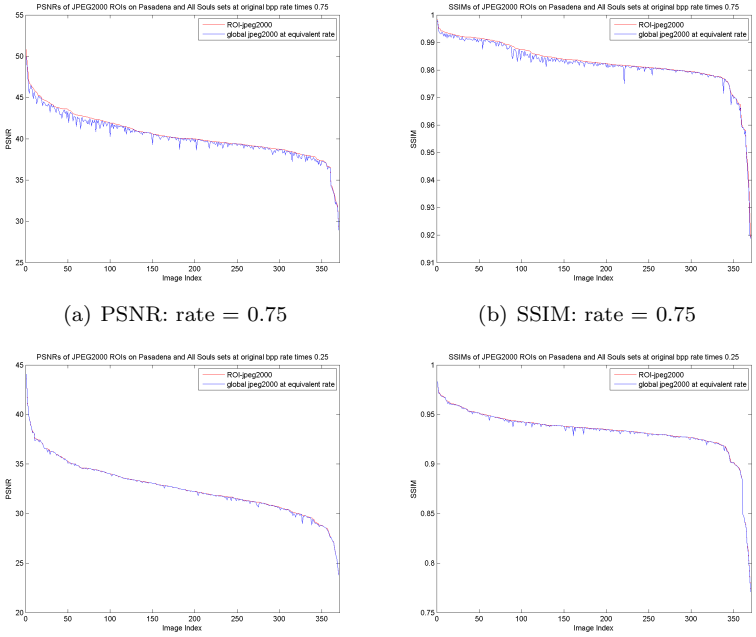
(g) jp2 global of (a) (h) jp2 ROI of (a) using (c) (i) jp2 global of (b) (j) jp2 ROI of (b) using (e)

(k) MSE^4 (a) & (g) (l) MSE^4 (a) & (h) (m) MSE^4 (b) & (i) (n) MSE^4 (b) & (j)

Fig. 12 Figure showing the VI segmentation in JPEG2000 (Example shown 0.75 times initial filesize). While the visual difference between the global and ROI files is hard to perceive (comparing (g) (h) and (i) (j)) there is a strong statistical difference as shown in (k), (l), (m) and (n). In the bottom row, brighter values indicate higher errors. It can be seen that the JPEG2000 ROI has less bright values within the core and that code blocking effects are present away from the ROI. The code block constraint on ROI processing is shown clearly in (l) and (n) where the regions away from the ROI are plainly block-segmented for encoding.

The algorithm used is the Kakadu implementation of the standard[4]. A search was undertaken over the Torralba, Aberdeen and Validation datasets to find parameters which provided a benefit in terms of the PSNR and SSIM measurements over the ROI [26,27]. (The parameters are $Clayers = 21$, $Cblk = 32, 32$, $Creversible = no$ $Rweight = 1500000$, $Rlevels = 21$.)

[4] based on the $kdu_{compress}$ examples of the Kakadu Software Company.



(a) PSNR: rate = 0.75

(b) SSIM: rate = 0.75

(c) PSNR: rate = 0.25

(d) SSIM: rate = 0.25

Fig. 13 Upper row ((a) and (b)): All images saved at 0.75 times their *bpp rate*. Lower row ((c) to (d)): All images saved at 0.25 times their *bpp rate*. The statistics in each chart are sorted by ROI JPEG2000 output in descending order over both image sets. Note that a rate of 0.75 (the upper row) approximately matches JPEG2000 to the average filesize obtained when VI + object is used in a standard JPEG scheme with pre-blur, described in Section 2.12. The rate 0.25 is the lowest rate observed to produce a statistical advantage (e.g. PNSR, SSIM) within the ROI.

3.2 Results

The quality measures shown in Figure 13 over the VI-selected ROI, show that a rate of $0.25 \times (initial)$ rate delivers PSNR and SSIM advantage over all of the images in both the Pasadena and All Souls sets (In this Figure, the data from both image sets is combined and sorted in order of declining *bpp*.) At the next smallest rate (0.125, not shown) some images achieve better quality measurement statistics over the ROI, but most of the images achieve the same quality. However, if a filesize reduction of 25% is sought (such as we found in our JPEG-based validation) (case, $rate = 0.75\times$ Original), the quality over the original ROI is increased relative to the global case. We conclude that we can use the visual interest segmentation in combination with JPEG2000 ROI to reduce the filesize to approximately 25% of that of the original while *gaining* in quality over the salient regions, relative to the global application of JPEG2000. If we choose a more modest outcome comparable to our validation experiments, there is a more notable advantage in conversion using the ROI.

4 Conclusion

In this paper we have presented an image segmentation algorithm to segment image areas which are visually salient to observers performing multiple tasks. In contrast to bottom-up saliency alone, and combined with specific task-search models our technique finds image areas relevant to the performance of multiple objective tasks and without the need for prior learning. The general mode acts on eye-level imagery with parameters chosen from careful experimentation and requires no machine learning stage. The technique is built upon feature points and the descriptors of these feature points can be compared to database representations of stored objects to narrow the focus of the attention prediction map for object class search, all in one algorithmic iteration.

In summary, the goal was to find a way of segmenting salient content without recourse to the limitations of biologically-inspired modelling. We wanted to segment image content that is salient to multiple viewing conditions without reverting to narrow constraints: to select image regions that are likely to be attended by observers performing multiple tasks on the same image. Furthermore we wanted to have a segmentation threshold method that would allow for a varying image area to be marked as salient between images depending on the image content. Finally, we wanted to avoid a learning element as a basis (for reliability reasons), but include the possibility to use it if it could be of benefit.

In contrast to the biological models looking for the most salient points in an image, the approach was to take an unknown image and state a priori that all of the pixels could be potentially salient prior to content analysis and to exclude those regions where there is little content worthy of inspection.

Utilising the fact that feature points are a good measure of potential visual saliency, we can accordingly view features as primitive detectors of visually-salient material. By setting an appropriate feature threshold related to the density of features, we can construct a map based on these features. By rigorous analysis of eye fixation count vs. map segmentation thresholds it was possible to select a threshold that captures reliably 70-75% of cluttered indoor scene fixations and considerably higher than that for outdoor scenes, 85-95%, even while the task is varied. (See Figure 5.)

Thus rather than claiming to have generated a vague, biologically-inspired probability map that we can arbitrarily threshold as we wish, we offer a segmented binary-level probability map where the expectation is that circa 85% of all task-directed fixations (from many tasks) are in the segmented area. Additionally, the area selected varies based on the image content. If the image has few areas with dense features, the segmented area is lower than if the image has many areas with high feature density. (See Figure 2.)

This fulfilled the initial goal: with no a priori information regarding the image or the observer's task, we have an algorithm that, by expectation, will segment image regions relevant to both bottom up and multiple tasks without requiring any machine learning. This was validated on new images though a

rigorous set of observer experiments using the parameterised Visual Interest system.

Regarding the applications of this technique, as a part of the validation process, we have demonstrated that it is possible to save store images with 15% less filesize relative to the equivalent global implementation of JPEG using the general VI algorithm based on image content only. This mode of operation retains visually salient information *relevant to many tasks* at a high quality level. Using object recognition contextualisation via object matching we can store images for 25% less than the equivalent global JPEG implementation filesize.

Finally, using the validated Visual Interest algorithm as a ROI in the JPEG2000 compression algorithm we achieve reduction to 25% of the original filesize while also gaining improved statistics over the ROI compared to the equivalent global implementation of JPEG2000. The final results are based on the general mode of Visual Interest operation, acting on the image content only, as opposed to the object contextualisation introduced in Section 2.

4.1 Future Work

The largest advantage in terms of compression gain is likely to come from implementation at the sensor head, where these high percentages would be acting on higher initial numbers (e.g. RAW images) for most captured images. Of course, the application to compression is just one possible use for a segmentation algorithm based on visually salient information. It could be used as an efficiency filter for automatic target recognition algorithms or as a task-oriented filtering algorithm to build saliency-marked images to aid image analysts, based on a more sophisticated object detection algorithm.

In future, utilising object recognition to narrow the salient area could result in improved JPEG2000 ROI quality even relative to these results.

References

1. Pasadena dataset web address: (accessed 08/2011. Http://www.vision.caltech.edu/html-files/archive.html
2. Bay, H., TuyteFlaars, T., Gool, L.V.: Surf: Speeded up robust features. Computer Vision and Image Understanding (CVIU) **110**(4), 346–359 (2006)
3. Brockmole, J.R., Castelhano, M.S., Henderson, J.M.: Contextual cueing in naturalistic scenes: Global and local contexts. Journal of Experimental Psychology: Learning, Memory, and Cognition **32**(4), 699–706 (2006)
4. deCampos, T.E., Csurka, G., Perronnin, F.: Images as sets of locally weighted features. Tech. Rep. VSSP-TR-1/2010, FEPS, University of Surrey, Guildford, UK (2010)
5. Fergus, R.: Visual object category recognition. Ph.D. thesis, University of Oxford (2005)
6. Hansen, B.C., Essock, E.A.: A horizontal bias in human visual processing of orientation and its correspondence to the structural components of natural scenes. J. Vis. **4**(12), 1044–1060 (2004)
7. Harding, P., Robertson, N.M.: A comparison of feature detectors with passive and task-based visual saliency. LNCS **5575**, 716–725 (2009)

8. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In: Advances in Neural Information Processing Systems 19, pp. 545–552 (2007)
9. Hopfinger, J.B., Buonocore, M.H., R.Mangun, G.: The neural mechanisms of top-down attentional control. Nature Neuroscience **3**, 284–291 (2000)
10. Itti, L.: Automatic foveation for video compression using a neurobiological model of visual attention. IEEE Transactions on Image Processing **13**(10), 1304–1318 (2004)
11. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis (1998)
12. Kadir, T., Brady, M.: Saliency, scale and image description. Int Journ. Comp. Vision **45**(2), 83–105 (2001)
13. L. Itti, C.K.: Computational modelling of visual attention. Nature Reviews Neuroscience. **2**(3), 194–203 (2001)
14. Lindeberg, T.: Scale-space theory: A basic tool for analysing structures at different scales. Journal of Applied Statistics **21**(2), 224–270 (1994)
15. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60**, 91–110 (2004)
16. Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust wide baseline stereo from maximally stable extremal regions. In: Proc. of British Machine Vision Conference, pp. 384–393 (2002)
17. Mikolajczyk, K., Schmid, C.: An affine invariant interest point detector. In: 7th European Conference on Computer Vision., vol. 1, pp. 128–142 (2002)
18. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Transactions on Pattern Analysis & Machine Intelligence **27**(10), 1615–1630 (2005)
19. Miller, J.L., Wiltse, J.M.: Resolution requirements for alphanumeric readability. Optical Engineering **42**(3), 846–852 (2003)
20. Navalpakkam, V., Itti, L.: Modeling the influence of task on attention. Vision Res **45**(2), 205–231 (2005)
21. Navalpakkam, V., Itti, L.: Search goal tunes visual features optimally. Neuron **53**(4), 605–617 (2007)
22. Peters, R., Itti, L.: Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–8 (2007)
23. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2007). URL http://www.robots.ox.ac.uk/ vgg/data/oxbuildings/index.html
24. Rosten, E., Drummond, T.: Fusing points and lines for high performance tracking. In: 10th IEEE International Conference on Computer Vision, vol. 2, pp. 1508–1511 (2005)
25. Rosten, E., Drummond, T.: Machine learning for high-speed corner detection. In: Proc 9th European Conference on Computer Vision, vol. 1, pp. 430–443 (2006)
26. Taubman, D.: Kakadu v5.0 survey document (2001)
27. Taubman, D.S., Marcellin, M.W.: JPEG 2000: Image Compression Fundamentals, Standards and Practice. Kluwer Academic Publishers, Norwell, MA, USA (2001)
28. Torralba, A.: Contextual priming for object detection. International Journal of Computer Vision **53**(2), 169–191 (2003)
29. Torralba, A., Oliva, A., Castelhano, M., Henderson, J.: Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. Psychological Review **113**(4), 766–786 (2006). URL http://people.csail.mit.edu/torralba/GlobalFeaturesAndAttention/
30. Union, I.T.: Reference algorithm for computing peak signal to noise ratio (psnr) of a video sequence with a constant delay. ITUT Standard (2009)
31. Viola, P., Jones, M.: Robust real-time object detection. International Journal of Computer Vision **57**, 137 – 154 (2001)
32. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing **13**, 600–612 (2004)
33. Wolfe, J.: Visual attention. In: De Valois KK, editor. Seeing. 2nd ed. San Diego, CA. Academic Press. (2000)

34. Wolfe, J.M., Horowitz, T.S., Kenner, N., Hyle, M., Vasan, N.: How fast can you change your mind? the speed of top-down guidance in visual search. Vision Research. **44**, 1411–1426 (2004)
35. Yu, S., Lisin, D.: Image compression based on visual saliency at individual scales. In: Advances in Visual Computing, *LNCS*, vol. 5875, pp. 157–166. Springer (2009)
36. Zhai, Y., Shah, M.: Visual attention detection in video sequences using spatiotemporal cues. In: ACM Multimedia, pp. 815–824 (2006)