Finding Visually Interesting Regions using SURF Points

Patrick J.N. Harding
Heriot-Watt University
Edinburgh, United Kingdom
pjh3@hw.ac.uk

Neil M. Robertson
Heriot-Watt University
Edinburgh, United Kingdom
n.m.robertson@hw.ac.uk

Abstract

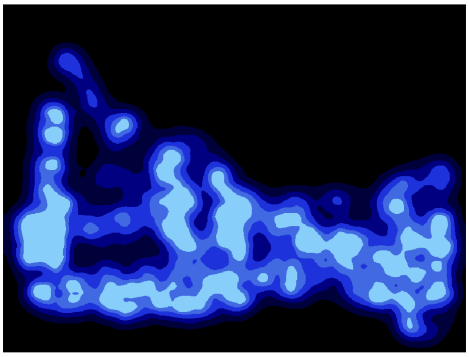
Recent work has shown that there is a varying coincidence of common interest-points towards the regions of an image that are visually salient. In this paper we compute a new saliency map which is derived from SURF interest points only. SURF points have been shown to be naturally distributed towards the visually salient regions in an image [4]. A probability map is computed which is then thresholded into the top 10 to 50% most salient pixels. We then compare the results with comprehensive eye-tracker data taken from human observers showing that up to 90% of the points attended by the observers can be recovered by our method. We then use this saliency map to perform more efficient image compression by extending the JPEG scheme to re-weight the image blocks by Q=50 or Q=5 depending on whether that region lies within a visually salient region of the image or not. We show the compression ratio is significantly higher *and* the more visually interesting regions are retained at higher resolution using our method.

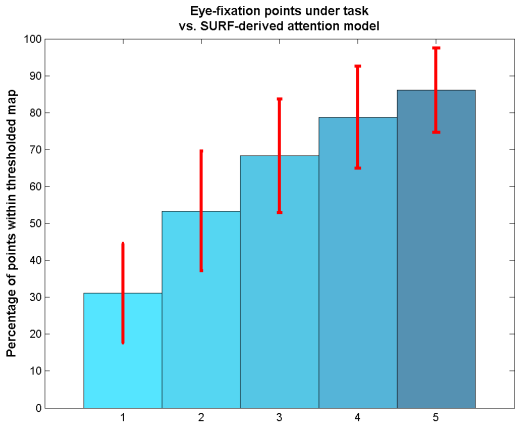
(a) (b)

Figure 1: Transforming interest points into a surface is the first step in our method. Using robust interest points this surface can be expected to represent interesting regions even under different observer viewing angles and/or conditions. (a) Original image with SURF points superimposed, (b) interest surface computed from SURF points.

1 Introduction

There exist reliable models of visual saliency under passive viewing derived from bottom-up visual processes, such as described in [5]. Under observer tasking, there is a substantial shift in attention away from the passive case strongly driven by the nature of the task [4, 2]. There are models of attentive prediction under task (such as in [8, 6, 7], but they are not general models based on image content and involve prior learning of object categories and contexts. Given that certain interest-points have a high correspondence with the visually salient in both passive and task-based cases [4] we propose a method of construction for a general purpose "probability" map of what is visually interesting in an image based on the best-performing interest point scheme analysed in [4], Speeded





(a) (b)

Figure 2: Validating the proposed method by comparing with eye-tracker data: (a) shows the thresholded suface from Figure 1(b); chart (b) shows the overlap with the maps at different threshold levels for *all* eye-points, gathered over 8 experimental participants. The bar indices 1 to 5 correspond to the 10 to 50 surface percentage coverage of the masks. The bars indicate average overlap at each threshold (standard deviation in red).

Up Robust Features [1]. This has the advantage of highlighting a range of regions of interest, that *could* be attended under different viewing conditions and requires no prior learning. This is of value because we can use this interest-surface to apply image compresion based on image region importance such that if the viewing conditions by a subsequent analyst are changed, the key details of the image are preserved. In Section 2 we describe how the SURF points are used to compute an "interest" surface which can be thresholded to various degrees of visual saliency. This surface is then tested against a large set of eye-tracker data taken from human observers under strict experimental conditions. In Section 3 we introduce a new compression scheme based on JPEG where the salient regions are used to define the level of compression applied to each block depending on its coincidence with the saliency map.

2 Saliency computation from SURF points

An illustration of the SURF points on a test image is shown in Figure 1(a). Figure 1(b) shows the attentional surface derived from the distribution of the SURF points. The visual interest surface is built by calculating the Euclidean distance of each point in a blank template to each SURF point. The distances are then ranked in order and the probability map value for each pixel is assigned as the sum of the second to the tenth ranked distance elements. The map is then inverted and normalised to the interval [0.1 1]. This construction technique delivers an attention map that does not experience strong peaks at the points themselves and is also dependent upon the local density of SURF interest points. The lower bound further allows for the possibility of attention in non-SURF rich regions, which may be useful in any future combination with other attentional surfaces.

We tested our new visual attention probability map against eye-fixation data from observers under task. The eye-tracker data and image set from Torralba *et al.* has been used to validate the model [8]. The test image data set for this paper comprises 72 images and 108 search scenarios (3x36 tasks) performed by 8 observers performing a search-and-count task. (The tasks were count people in outdoor contexts and count paintings and cups in indoor contexts. Note that there are 36 images that have two different tasks applied to them). The visual-interest maps were constructed as described above for each image. The probability map was then thresholded to 10, 20, 30, 40 and 50% of the most salient pixels by image area, representing the supposedly more salient half of the image to finer

(a) (b) (c)

Figure 3: (a) The original image, (b) Standard Q50 JPEG compression, (c) Our new visually interesting region compression: Q50 in top 40% visually interesting regions, Q5 elsewhere. Note by inspection that the visually important regions in the image are generally preserved, while the contextual information outwith the core remains valuable, although highly degraded. The compression ratio is improved from 5.25 ± 0.5 to 6.0 ± 0.3 using our method without loss of detail at the salient locations (in this case the house).

degrees. Finally, the overlap % of the eye-tracker data was counted at each threshold level to assess the accuracy of the interest-point derived map at predicting human eye-fixations under task. The results are shown in Figure 2. The high coincidence of the eye-fixations with our attention based regions is a strong result. It remains consistent across different tasks, validating our assumption that the interest-points are a good way of assessing regions of visual interest under varying observation conditions.

3 JPEG encoding reweighted towards salient regions

We next use this result for a practical application. We choose to demonstrate a scheme for compression based on the JPEG algorithm which is designed for good visual quality in photo-real images. JPEG relies on quantisation of the Discrete Cosine Transform applied to 8 by 8 pixel blocks of an image. This reduces the relatively unimportant high frequency components in each block, allowing for efficient huffman or arithmetic coding. The quantisation is performed using a quantisation matrix derived from psychovisual tests and this matrix can be weighted to provide the required degree of compression in the block. The reverse process decodes the image [3, 9]. The heavier this quantisation, the larger the compression ratio achieved, however this is tempered by the fact that over-quantisation will produce blocking atifacts that significantly reduce image quantity and can damage real information within the image. In regular JPEG, the quantisation is fixed across the whole image. In our case, however, we have a reliable method of selecting regions of visual interest. In this example presented here, we choose to preserve the top 40% of the image by "visual interest" from SURF-point distribution, which we can expect in a probabilistic sense to attract 80% of eye fixations under task. We will compress the other 60% to a much higher degree. This information will not be lost altogether and will be available for contextual information.

We use a greyscale copy of the image and choose two quality factors to impose a high or low quality on the image region. The quality factor (Q) of 50 uses an unweighted matrix which is the original matrix derived from psychovisual experiments to give acceptable compression. The quantisation matrix we use is that specified in Annex K of the JPEG standard for the luminance component of images [3], appropriate for grayscale. We choose a low value of Q=5 for the outlying regions and weight the quantisation matrix according to the following relationship: $(50/Q) * Qmatrix$.

We set a threshold such that if the pixel in the image was in the most "visually interesting" 40% of the image it would have the Q50 weighting applied, Q5 otherwise. The quantised set of blocks were vectorised according to the jpeg zig-zag pattern [9] and the DC components were encoded dif-

ferentially according to the previous element. We appended each DCT block with one more element to give a block length of 65 - 0 if the block was for high compression and 1 if the block was for normal compression. Since the DCT-quantisation process generally results in large numbers of zeros at the highest of frequencies, the huffman encoding scheme that we used is generally efficient. We do add one more piece of information per 64 pixels, but we are able to discard more information with confidence than otherwise and there is a net gain from the efficiency in the huffman coding process.

We applied this visually-interesting region compression over all of our 72 images. As a comparison, we also performed a normal JPEG process at Q50. An illustration of the output is shown in figure 3. Over all of the 72 images the average compression ratio achieved by our Q50 JPEG compression was 5.25 ± 0.5 and for our visually interesting region based compression the achieved compression ratio was 6.0 ± 0.3. From the image set statistics above and from the illustration in figure 3, it is clear that there is an advantage in the method in terms of performance over regular JPEG as well as being capable of producing usable images where the core details of the scene survive the compression. e.g. in figure 3(b) the sky and bland textural details have suffered strong degradation, but the interesting content of the scene is largely preserved.

4 Conclusion and future work

In conclusion, we have used a robust interest-point detector (known to coincide with the visually salient parts of an image under different observer conditions) to construct a map of visually interesting regions of an image. We have validated this technique against observers acting under different tasks and the method is a good predictor of eye scan points under object count tasking. We have further demonstrated a compression scheme using the visual interest map as a guide that offers advantage in terms of filesize while preserving the core details of images. A strong advantage of the method is that it is simple and can be performed "live", being based on existing image content and not requiring any prior learning stages.

Avenues for future work include introducing a spectrum of Q values for the encoding of the JPEG compression rather than the binary approach taken here. It is likely that compression is not the only image processing technique which would benefit from intelligent application based on salient regions. We propose next to investigate segmenting an image using saliency then discriminately performing image enhancement on regions of interest.

References

[1] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. *Computer Vision and Image Understanding (CVIU)*, 110(4):346–359, 2006.

[2] M. S. Castelhano, M. L. Mack, and J. M. Henderson. Viewing task influences eye movement control during active scene perception. *Journal of Vision*, 9(3):1–15, 3 2009.

[3] JPEG Committee. Iso/iec 10918-1. ISO Standard, 1994.

[4] P. Harding and N. M. Robertson. A comparison of feature detectors with passive and task-based visual saliency. *LNCS*, 5575:716–725, 2009.

[5] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *Advances in Neural Information Processing Systems 19*, pages 545–552, 2007.

[6] V. Navalpakkam and L. Itti. Modeling the influence of task on attention. *Vision Res*, 45(2):205–231, January 2005.

[7] R.J. Peters and L. Itti. 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)*, Jun 2007.

[8] A. Torralba, A. Oliva, M.S. Castelhano, and J.M. Henderson. 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, October 2006.

[9] G.K. Wallace. The jpeg still picture compression standard. *Commun. ACM*, 34(4):30–44, 1991.