A Parametric Spectral Model for Texture-Based Salience

Kasim Terzić, Sai Krishna and J.M.H. du Buf
{kterzic,dubuf}@ualg.pt

Vision Laboratory/LARSys, University of the Algarve

Abstract. We present a novel saliency mechanism based on texture. Local texture at each pixel is characterised by the 2D spectrum obtained from oriented Gabor filters. We then apply a parametric model and describe the texture at each pixel by a combination of two 1D Gaussian approximations. This results in a simple model which consists of only four parameters. These four parameters are then used as feature channels and standard Difference-of-Gaussian blob detection is applied in order to detect salient areas in the image, similar to the Itti and Koch model. Finally, a diffusion process is used to sharpen the resulting regions. Evaluation on a large saliency dataset shows a significant improvement of our method over the baseline Itti and Koch model.

1 Introduction

Texture is known to be a powerful cue in early vision [32, 20] and has consequently received much attention from the Computer Vision and Neuroscience communities. The seminal work on saliency maps by Itti and Koch included an orientation component, calculated by a bank of Gabor filters [16], and there has been much work on texture segmentation. However, texture remains one of the hardest feature channels to model, and most recent work on saliency focuses on colour, contrast and local region descriptors.

In this paper, we return to the problem of texture in saliency models by extending the Itti and Koch model. By interpreting oriented Gabor filter responses as a local power spectrum of the image, we define a simple parametric model in order to characterise local texture in terms of orientation, anisotropy, scale and complexity. The model parameters are then used as features and processed by a set of centre-surround cells, as in [16] and followed by a simple diffusion process to obtain preliminary results. Evaluation on a standard saliency dataset shows that our texture-based saliency model outperforms other texture-based models. It is competitive with the original Itti and Koch model, despite only using texture. A combination of texture and colour outperforms the baseline Itti and Koch model and achieves promising results.

Our texture model is built on top of responses of complex cells in V1 which can be efficiently computed [30]. Consequently, it not only adds a powerful feature to saliency estimation methods, but could also serve as a plausible texture model for early vision.

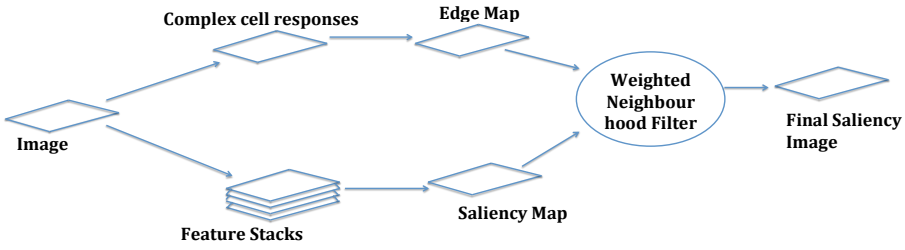

Fig. 1. Overview of our texture-based saliency method. First, the input image is processed using a bank of Gabor filters. The responses are used to obtain complex cell responses and the edge map (top), and to obtain a stack of texture features which are processed by a set of centre-surround cells to obtain a saliency map (bottom). The saliency map is combined with the edge map in a diffusion filtering step to provide the final texture-based saliency map.

2 Related Work

Much work on visual saliency is motivated by the early processing in the visual cortex. One of the first biological models was created by Itti, Koch and Niebur [18, 16], where intensity, colour and orientation maps are processed by a bank of centre-surround filters. This influential model shaped much of later work on saliency and attention. Related work includes weighting of different feature maps after identifying useful features [15] and exploring the role of saliency in overt attention [25]. It has been noted that the original Itti and Koch model, designed for eye movement simulation, is not well-suited for object-based salience, and an extended model was shown to reach state-of-the-art results [8]. Similarly, eye fixation maps were combined with traditional segmentation methods in [22].

In recent years, there has been a shift towards detecting complete salient objects in scenes, with a large region covering most of an object. Often, an image is segmented, and regions are labelled according to colour and luminance [1], region-based contrast [6, 5] or dissimilarity between image patches [7]. One approach attempts to learn a correct foreground object segmentation from training images [23]. Object-based saliency is important for interfacing with scene-understanding systems from AI [28, 24] or for cognitive robotics [29], where sequential scene processing is common.

Other approaches from Computer Vision include image regions which represent the scene in terms of visual perception [10], graph-based visual saliency [13], and object-based saliency features [12]. There have also been attempts to model saliency as a discriminant process [9], a regression problem [19], or using a Bayesian surprise criterion [17]. It has been shown that hierarchical, multi-scale processing can improve saliency on small-scale, high-contrast patterns [33].

Very few saliency methods explicitly use spatial frequency or texture. In addition to the approaches related to the Itti and Koch model, which use orientation

as one of the feature channels, there have been several approaches using the frequency spectrum. Achanta et al. [2] used bandpass filtering to obtain uniform regions with sharp boundaries, but their features were still based on colour. Two approaches extracted saliency from the frequency spectrum of the image. Hou and Zhang introduced a method based on the global Fourier transform [14]. By subtracting the average log-spectrum of many images from the log-spectrum of the individual image, they obtain a spectral residual which, when transformed back into the spatial domain, indicates salient regions which potentially correspond to objects. Guo et al. [11] built on this concept, but argued that the phase, not amplitude, of the spectrum is key to finding salient regions. They extended this concept to the Quaternion Fourier Transform which can represent intensity, colour and motion of each pixel. Neither of these methods is biologically plausible, or based on texture. We are not aware of any recent work on saliency which attempts to explicitly model and compare texture.

In the rest of this paper we present a new and more biological interpretation of the local Gabor filter responses. We describe the local texture using a parametric model. The parameters of this model represent new features, which are then processed using centre-surround filters.

3 Method

Our method attempts to find consistent regions which are different from their surroundings, using centre-surround blob detection. To this end, we characterise local texture at each pixel using a parametric model, where the four parameters correspond to orientation selectivity (isotropic-anisotropic), dominant orientation, scale selectivity, and dominant scale (from coarse to fine). Figure 1 shows an overview of our method. We calculate the edge map based on the responses of complex cells. In parallel we extract four feature maps based on texture and calculate a salience map by performing blob detection. The salience map is combined with the edge map in a weighted-filtering step.

3.1 V1 Model

Our method begins by extracting responses of oriented Gabor filters at multiple orientations and scales. Gabor filters are commonly used as a model of so-called simple cells in the early visual cortex. In our implementation, we rely on the fast V1 model from [31], applying default parameters: 8 orientations and 7 logarithmically spaced scales. Complex Gabor filters are modelled by

$$G_{\lambda,\sigma,\theta}(x,y) = \exp\left(-\frac{\tilde{x}^2+\gamma\tilde{y}^2}{2\sigma^2}\right)\exp\left(i\frac{2\pi\tilde{x}}{\lambda}\right) \ , \tag{1}$$

where

$$\tilde{x} = x\cos\theta + y\sin\theta \tag{2}$$

$$\tilde{y} = y\cos\theta - x\sin\theta \ , \tag{3}$$

λ is the wavelength in pixels, and σ the receptive field size in pixels. We apply default parameters from [31]: $\sigma/\lambda = 0.56$, $\gamma = 0.5$ and θ assumes 8 values, equally spaced on $[0, \pi)$.

Responses of simple cells are obtained by convolving the image with the complex Gabor filters:

$$S_{\lambda,\theta} = I * G_{\lambda,\theta} \ . \tag{4}$$

The moduli of simple cell responses are used to model complex cortical cells:

$$C_{\lambda,\theta}(x, y) = |S_{\lambda,\theta}(x, y)| \ . \tag{5}$$

3.2 Local Texture Model

Since Gabor filters are bandpass filters, it is evident that the responses of all complex cells computed at a particular position represent the frequency spectrum of the local region. Each filter response represents a sample in this power spectrum. Although Gabor filtering is typically expensive, it is the first step of any biological model, and there are optimised and GPU-accelerated solutions [31].

Like the corresponding Gabor filters, the spectrum has two dimensions: orientation (corresponding to filter orientation) and frequency (corresponding to filter wavelength), which effectively yields a 2D matrix. This matrix is cyclic in the orientation, i.e., a cylinder. In our model, we assume that the power spectrum can be approximated by a 2D-separable Gaussian function. This is obviously a very rough approximation, but we are not interested in reconstructing the texture, only in measuring whether there is a noticeable difference between the textures at neighbouring positions. In practice, we found that approximating the marginals by two 1D Gaussians is simpler and also produces good results.

The processing of each 2D matrix is very simple and fast. First, the noisy spectrum is smoothed by applying a 3x3 lowpass block filter. Then, the 2D array is projected (summed) into two 1D arrays: the scale array S_i and the (cyclic) orientation array O_i. In both arrays, the local maximum is detected, yielding the "means" μ_s and μ_o, after which the standard deviations σ_s and σ_o are computed, taking into account the periodicity of O_i. Experimental results revealed no significant differences between using the maxima as means and using the real means as computed by moments. Figure 2 illustrates this process.

As described above, the local power spectrum is modelled by four parameters: the means and standard deviations in the orientation and frequency dimensions. The mean orientation of the Gaussian μ_o thus encodes the dominant orientation of the texture, and the standard deviation σ_o is a measure of isotropy: small values of σ_o indicate a strong preference for a particular direction, while large values mean that many different orientations are present. In terms of frequency, μ_s encodes the characteristic scale of the texture, coarse vs. fine, while σ_s tells us whether there is one characteristic scale or a mixture of coarse and fine scales. Figure 3 shows the four texture features extracted from a real image.

This model is obviously not very discriminative: it does not deal with multimodal and non-Gaussian spectra. However, it is considerably more powerful

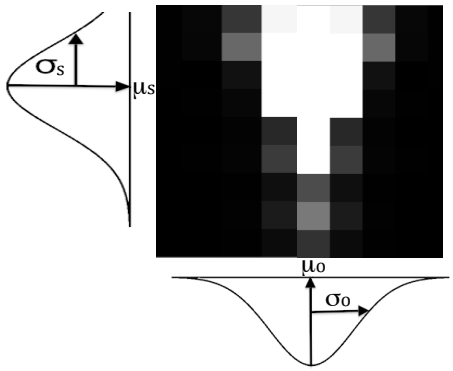

Fig. 2. Our texture model. A local power spectrum is a 2D matrix where the dimensions represent orientation (horizontal axis) and frequency (vertical). The spectrum resembles a 2D Gaussian function. We can fit two 1D Gaussians to the 1D marginals of the spectrum to obtain the means and standard deviations of orientation and frequency, which we use as features.

Fig. 3. Our texture features extracted from a real image. Blob detection on these feature maps is used to produce a saliency map.

than just using the dominant local orientation, and the additional complexity is minimal – Gabor filtering is far more expensive than curve fitting. The four parameters described above are then used with blob-detection kernels to extract salient textured regions. Then blob extraction is applied as shown in Fig. 4.

3.3 Blob Detection

The blob detection step is the same as in [16]. The four parameters $\mu_o, \sigma_o, \mu_s, \sigma_s$ are calculated for each pixel of the image and stored in four maps with the same dimensions as the image: M^o_μ, M^s_μ, M^o_σ and M^s_σ. The algorithm works on full-sized images, but subsampling is possible for improving the speed of the filtering operations.

The four maps are then processed by a bank of centre-surround Difference-of-Gaussian filters with different sizes, as is common. In our implementation, we use three sizes and the filters are typical "mexican hat" kernels which combine a positive Gaussian and a negative one with a larger standard deviation. For the

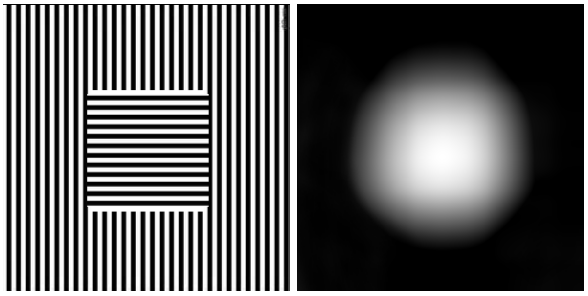

Fig. 4. An example of texture saliency. The image on the left has a salient region identified only by texture (average intensity and colour are the same). Blob detection based on colour fails in this case, but blob detection based on texture features as described in this paper detects a salient blob (right image).

three different filter sizes, the standard deviations of the positive Gaussians are 45, 90 and 180 pixels, and those of the negative Gaussians are 90, 180 and 360, respectively. The filters need to be large in order to capture large salient objects, but this presents a problem with smaller images. We therefore apply extensive border-replicating padding of the feature images to avoid this problem. The resulting 12 saliency maps are summed and normalised to 0-255 to obtain the pre-final saliency map.

3.4 Region Sharpening

Blob detection is good at identifying the centres of salient regions, but blob boundaries are poorly defined. It may be useful for overt attention models, but less useful for localising and segmenting salient objects. In order to sharpen region boundaries and to create more homogeneous regions which better correspond to complete objects, we apply a non-linear diffusion step. Although the idea of diffusion in early vision is not without controversy, it has been suggested that colouring and surface interpolation mechanisms take place in V1 [21], especially as a result of feedback from higher areas V2 and V4 [27].

We begin by taking the sum of all complex cell responses extracted at the finest scale:

$$C_{\text{edge}}(x,y) = \sum_{\theta} C_{\lambda,\theta}(x,y) \ , \tag{6}$$

where λ corresponds to the finest scale applied in the previous step. The combined map C_{edge} resembles an edge map, where large values correspond to narrow bars or sharp transitions between different intensity values. This map is normalised to the range 0–1.

We then apply a weighted neighbourhood filter to each point in the saliency map S, based on the values of its neighbours:

$$s(x,y) = \frac{1}{8} \sum_{i \in 1}^{8} w_i S_i \ , \tag{7}$$

w1	w2	w3
w8		w4
w7	w6	w5

Fig. 5. Weights used in the diffusion filtering step. We simulate the diffusion process by repeated weighted average filtering.

where s_i are the 8 neighbours of the central pixel $s(x, y)$: $S_1 = s(x-1, y-1)$, $S_2 = s(x, y-1)$, etc. (see Fig. 5 for an illustration). The weights w_i depend on the strength of the edge map C_{edge} at that pixel:

$$w_i = (1 - C_i) \,, \tag{8}$$

where C_i is the value of the edge map C_{edge} at relative position i.

The result of this filtering is a strong influence of neighbouring pixels not lying on an edge, and no influence of pixels located on edges. This can be seen as a dynamical diffusion process in the early visual cortex, where neighbouring cells (representing saliency) excite each other, but the connections are inhibited by complex cells. In our model, we repeat the filtering process a set number of times to approximate the equilibrium solution. A further improvement can be obtained by extracting closed contours from the image before filtering.

The filtering process ensures that closed regions become more uniformly salient, while outside regions become less salient. Figure 6 shows an example of this process on a real image. It can be seen that the shape of the blob is acceptable although it is too big because of the sizes of the Difference-of-Gaussian filters. It is also stronger close to the edges, and some parts of the object have low salience. The diffusion filtering on the basis of responses of complex cells at the finest scale is able to correct the size and, because responses outside the blob are suppressed, thresholding can be applied to obtain a binary mask. Below, the threshold value will be used as a free parameter in quantitative evaluation.

4 Evaluation

We evaluated the texture-based saliency method on the standard saliency dataset developed by Achanta et al. [2] The dataset consists of 1000 images, each containing a single salient object, plus hand-annotated ground-truth masks.

Figure 7 shows the results of our algorithm on some of the images from the dataset. It can be seen that our algorithm consistently highlights the salient regions in the images. The diffusion step results in well-defined region boundaries which correspond to entire objects.

Figure 8 (left) shows a comparison of our texture-only algorithm against similar algorithms: the Itti and Koch baseline model on this dataset and two

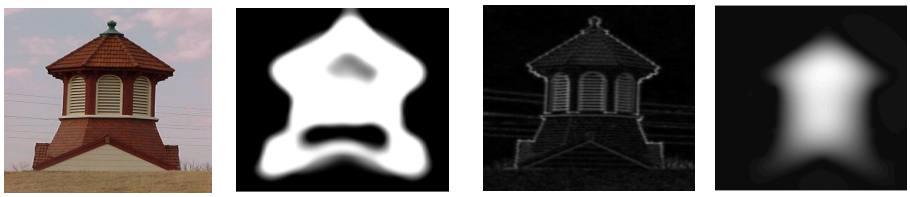

Fig. 6. Left to right: input image, the result of blob detection, edges obtained from the responses of complex cells, and saliency corrected by diffusion filtering. Texture saliency responds strongly to areas where texture is different from its surrounding, but it does not uniformly cover the entire object and, due to large blob detection kernels, it also responds outside object boundaries. Combining saliency with image edges during the diffusion filtering step results in smoother, object-based saliency.

approaches based on texture or frequency. We plot the precision-recall curves obtained by varying the threshold used to binarise the saliency images. Precision and recall are computed by comparing each pixel in the saliency map with the hand-annotated ground truth map, counting all true and false positives and negatives in all 1000 images. Our algorithm outperforms the other methods, significantly improving the state of the art in terms of texture-based saliency. It can be seen that our texture saliency alone can slightly outperform the classic Itti and Koch model. This is most likely due to the selection of kernel sizes for blob detection, since their model was designed before this dataset became popular, and was optimised for modelling sequential saccadic eye movements. Figure 8 (right) shows a comparison with two state-of-the-art methods. We added three colour features to our model and averaged the salience maps for this experiment.

5 Discussion

The texture parameters applied in the salience model are based on a more complex model [3], but extremely simplified in order to be applicable in real-time applications. Nevertheless, the very good results in terms of salience suggest that further refinements may not be required if texture is going to be combined with colour, motion and stereo disparity. More advanced texture models exist, for example based on models of cortical grating cells on top of which the texture symmetry order could be detected (linear, rectangular, hexagonal, etc.) [4], but this information may not be very accurate in real-world applications where almost no textures show perfect symmetries.

There are several recent methods which obtain better results than our method on this dataset. We stress that our work was aimed at creating novel texture-based features and that results presented here are preliminary. Integration of our features with state-of-the-art methods and a wider selection of features is expected to make our model more competitive. In this paper, we concentrated

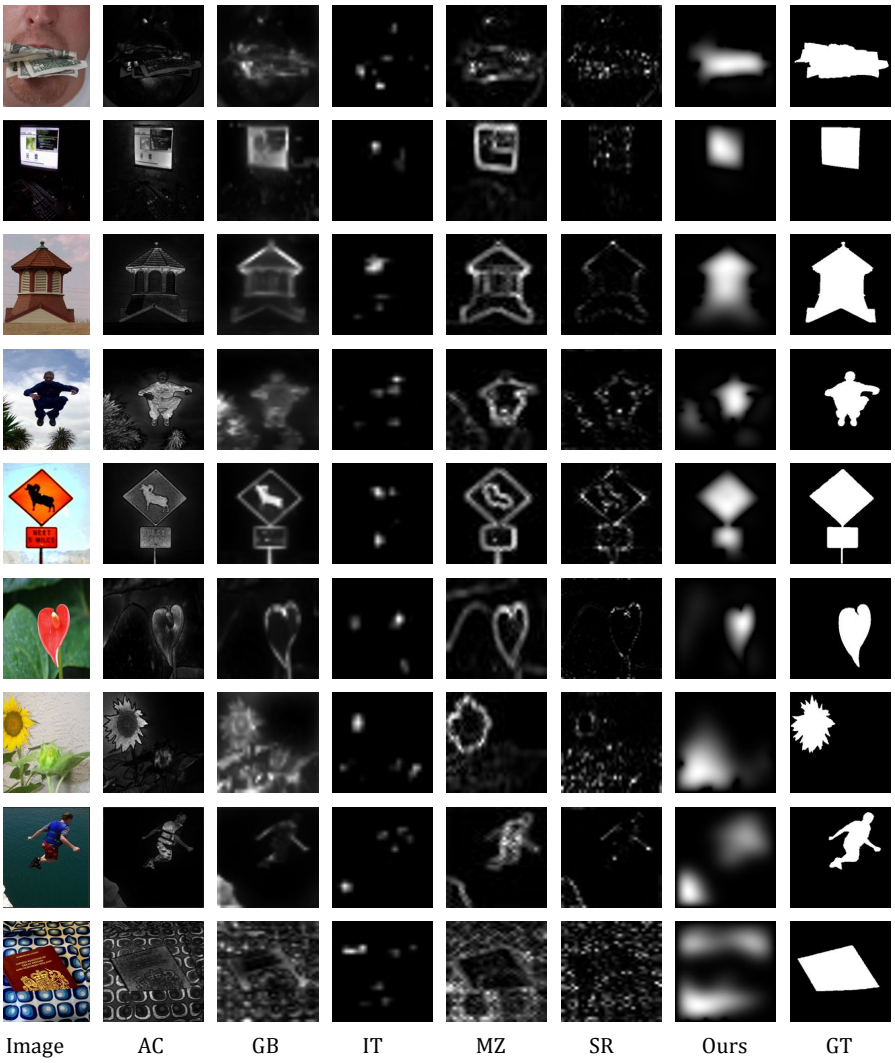

Fig. 7. Visual comparison of results on the saliency dataset. The input images are shown in the left column. The ground truth annotations are shown in the right column. The remaining columns, from left to right, show the results of AC [2], GB [13], IT [16], MZ [11], SR [14], and our algorithm, before thresholding. The bottom three rows show some difficult examples. Our algorithm responds strongly to the alternating textures of the leaves and the wall in the bottom left corner of the sunflower image (third from below), and fails completely with the passport image (bottom row). In the second row from below, we also detect the rock, which is salient but not annotated.

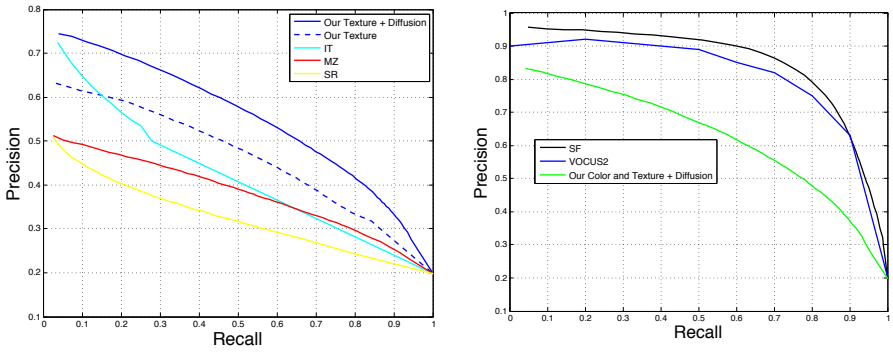

Fig. 8. Comparison against some state-of-the-art models. The left graph shows a comparison against methods which incorporate texture or frequency: the original Itti and Koch model (IT) [16], and the spectrum-based models of Hou and Zhang (SR) [14] and Guo et al. [11]. The right graph shows our model extended with colour against two state-of-the-art models: Perazzi et al. (SF) [26] and the improved Itti and Koch model VOCUS2 [8].

on the improvement in texture-based saliency, which is a much overlooked part of saliency models.

6 Conclusion

Although texture is considered an important cue for attention, segmentation and object detection, only few saliency models currently exploit texture. In this contribution, we have presented a novel texture-based method which extends the Itti and Koch model and shows that texture can be a very useful cue for advancing saliency models. We are not aware of any texture-based work achieving significant results on standardised saliency datasets, so showing results using only texture is an interesting achievement.

Evaluation on the standard dataset shows that our saliency model alone outperforms the baseline Itti and Koch model, and that a combination of texture and colour adds an additional boost. Unlike many popular methods which are based on region segmentation and local descriptors, our method is biologically motivated and could help to explain the role of texture in early saliency processing, and how it can drive saccadic eye movements to objects.

Ongoing work focuses on integrating further cues such as motion and disparity, and applying the saliency model on a real-time robot.

Acknowledgements This work was supported by the EU under the FP-7 grant ICT-2009.2.1-270247 *NeuralDynamics* and by the FCT under the grants LarSYS UID/EEA/50009/2013 and SparseCoding EXPL/EEI-SII/1982/2013.

References

1. Achanta, R., Estrada, F.J., Wils, P., Süsstrunk, S.: Salient region detection and segmentation. In: ICVS. pp. 66–75 (2008)
2. Achanta, R., Hemami, S.S., Estrada, F.J., Süsstrunk, S.: Frequency-tuned salient region detection. In: CVPR. pp. 1597–1604 (2009)
3. du Buf, J.: Abstract processes in texture discrimination. Spatial Vision 6, 221–242 (1992)
4. du Buf, J.: Improved grating and bar cell models in cortical area V1 and texture coding. Image and Vision Computing 25, 873–882 (2007)
5. Cheng, M.M., Mitra, N.J., Huang, X., Torr, P.H.S., Hu, S.M.: Global contrast based salient region detection. IEEE T-PAMI 37(3), 569–582 (Mar 2015)
6. Cheng, M., Zhang, G., Mitra, N.J., Huang, X., Hu, S.: Global contrast based salient region detection. In: CVPR. pp. 409–416 (2011)
7. Duan, L., Wu, C., Miao, J., Qing, L., Fu, Y.: Visual saliency detection by spatially weighted dissimilarity. In: CVPR. pp. 473–480 (2011)
8. Frintrop, S., Werner, T., Martin-Garcia, G.: Traditional saliency reloaded: A good old model in new shape. In: CVPR (2015)
9. Gao, D., Vasconcelos, N.: Bottom-up saliency is a discriminant process. In: ICCV. pp. 1–6 (2007)
10. Goferman, S., Zelnik-Manor, L., Tal, A.: Context-aware saliency detection. In: CVPR. pp. 2376–2383 (2010)
11. Guo, C., Ma, Q., Zhang, L.: Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. In: CVPR (2008)
12. Han, J., Ngan, K.N., Li, M., Zhang, H.: Unsupervised extraction of visual attention objects in color images. IEEE Trans. Circuits Syst. Video Techn. 16(1), 141–145 (2006)
13. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In: NIPS. pp. 545–552 (2006)
14. Hou, X., Zhang, L.: Saliency detection: A spectral residual approach. In: CVPR (2007)
15. Hu, Y., Xie, X., Ma, W., Chia, L., Rajan, D.: Salient region detection using weighted feature maps based on the human visual attention model. In: PCM. pp. 993–1000 (2004)
16. Itti, L., Koch, C.: A saliency-based search mechanism for overt and covert shifts of visual attention. Vision Research 40(10-12), 1489–1506 (May 2000)
17. Itti, L., Baldi, P.: Bayesian surprise attracts human attention. In: NIPS. pp. 547–554 (2005)
18. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. 20(11), 1254–1259 (1998)
19. Jiang, H., Wang, J., Yuan, Z., Wu, Y., Zheng, N.: Salient object detection: a discriminative regional feature integration approach. In: CVPR (2013)
20. Julesz, B.: Textons, the elements of texture perception, and their interactions. Nature 290(5802), 91–97 (Mar 1981)
21. Lee, T.S., Mumford, D., Romero, R., Lamme, V.F.: The role of the primary visual cortex in higher level vision. Vision Research 38, 2429–2454 (1998)
22. Li, Y., Hou, X., Koch, C., Rehg, J.M., Yuille, A.L.: The secrets of salient object segmentation. In: CVPR. pp. 280–287 (2014)
23. Liu, T., Sun, J., Zheng, N., Tang, X., Shum, H.: Learning to detect a salient object. In: CVPR (2007)

24. Neumann, B., Terzić, K.: Context-based probabilistic scene interpretation. In: IFIP AI. pp. 155–164 (Sep 2010)
25. Parkhurst, D., Law, K., Niebur, E.: Modeling the role of salience in the allocation of overt visual attention. Vision Res. 42(1), 107–123 (Jan 2002)
26. Perazzi, F., Krähenbühl, P., Pritch, Y., Hornung, A.: Saliency filters: Contrast based filtering for salient region detection. In: IEEE CVPR. pp. 733–740 (2012)
27. Self, M.W., van Kerkoerle, T., Super, H., Roelfsema, P.R.: Distinct roles of the cortical layers of area V1 in figure-ground segregation. Current Biology 23, 2121–2129 (2013)
28. Terzić, K., Hotz, L., Šochman, J.: Interpreting structures in man-made scenes: Combining low-level and high-level structure sources. In: International Conference on Agents and Artificial Intelligence. Valencia, Spain (Jan 2010)
29. Terzić, K., Lobato, D., Saleiro, M., Martins, J., Farrajota, M., Rodrigues, J., du Buf, J.: Biological models for active vision: Towards a unified architecture. In: ICVS 2013, LNCS. vol. 7963, pp. 113–122 (Jul 2013)
30. Terzić, K., Rodrigues, J., du Buf, J.: Fast cortical keypoints for real-time object recognition. In: ICIP. pp. 3372–3376. Melbourne (Sep 2013)
31. Terzić, K., Rodrigues, J., du Buf, J.: BIMP: A real-time biological model of multi-scale keypoint detection in V1. Neurocomputing 150, 227–237 (2015)
32. Wolfe, J.M., Horowitz, T.S.: What attributes guide the deployment of visual attention and how do they do it? Nature Reviews Neuroscience 5, 495–501 (Jun 2004)
33. Yan, Q., Xu, L., Shi, J., Jia, J.: Hierarchical saliency detection. In: CVPR (2013)