Saliency Detection Based on Scale Selectivity of Human Visual System

Fang Fang[1,3], Laiyun Qing[2,3], Jun Miao[3], Xilin Chen[3], and Wen Gao[1,4]

[1] School of Computer Science and Technology,
Harbin Institute of Technology, Harbin 150001, China
[2] Graduate University of Chinese Academy of Science, Beijing 100049, China
[3] Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
[4] Institute of Digital Media, Peking University, Beijing 100871, China
{ffang,lyqing,jmiao,xlchen,wgao}@jdl.ac.cn

Abstract. It is well known that visual attention and saliency mechanisms play an important role in human visual perception. This paper proposes a novel bottom-up saliency mechanism through scale space analysis. The research on human perception had shown that our ability to perceive a visual scene with different scales is described with the Contrast-Sensitivity Function (CSF). Motivated by this observation, we model the saliency as weighted average of the multi-scale analysis of the visual scene, where the weights of the middle spatial frequency bands are larger than others, following the CSF. This method is tested on natural images. The experimental results show that this approach is able to quickly extract salient regions which are consistent with human visual perception, both qualitatively and quantitatively.

Keywords: Saliency Detection, Human Visual System, Scale Selectivity.

1 Introduction

It is well known that the visual attention and saliency mechanism play an important role in human visual perception. In recent years, there have been increasing efforts to introduce computational models to explain the fundamental properties of biological visual saliency. It is generally agreed that visual attention of Human Vision System (HVS) is an interaction between bottom-up and top-down mechanisms. In this paper, we will focus on the bottom-up saliency detection.

Bottom-up saliency drives attention only by the properties of the stimuli in a visual scene and is independent of any high level visual tasks. Inspired by the early visual pathway in biological vision, the features used in saliency models include low-level simple visual attributes, such as intensity, color, orientation and motion. In one of the most popular models for bottom-up saliency [1], saliency is measured as the absolute difference between feature responses at a location and those in its neighborhood, in a center-surround fashion. This model has been shown to successfully replicate many

B.-L. Lu, L. Zhang, and J. Kwok (Eds.): ICONIP 2011, Part I, LNCS 7062, pp. 172–181, 2011.

observations from psychophysics [2]. Gao and Mahadevan [3] implemented the center-surround mechanisms as a discriminate process based on the distributions of local features centering and surrounding at a given point. In a recent proposal, Kienzle et al. [4] employed machine learning techniques to build a saliency model from human eye fixations on natural images, and showed that a center-surround receptive field emerged from the learned classifier. Some other recent works model the saliency in information theory and deriving saliency mechanisms as optimal implementations of generic computational principles, such as the maximization of self-information [5], local entropy [6] or 'surprise' [7]. Other methods are purely computational and are not based on biological vision principles [8, 9].

This paper proposes a saliency detection method based on the scale selectivity of HVS. In natural scenes, objects and patterns can appear at a wide variety of scales. In other words, a natural image includes signals of multiple scales or frequencies. However, the HVS is able to quickly select the frequency band that conveyed the most information to solve a given task and interpret the image, not conscious of the information in the other scales [10]. Researches on human perception had suggested that image understanding is based on a multi-scale, global to local analysis of the visual input [11, 12]. The global precedence hypothesis of image analysis implies that the low spatial frequency components dominate early visual processing. Physiological research has also shown that our ability to perceive the details of a visual scene is determined by the relative size and contrast of the detail present. The threshold contrast necessary for perception of the signal is found to be a function of its spatial frequency, described by the CSF [13, 14], in which contrast sensitivities are higher in the middle frequency bands than the other bands. These inspire us to detect saliency in the visual by using the low-middle frequency components and ignoring the other ones.

The rest of the paper is organized as follows. Some related works on scale selectivity of human vision system will be given in Section 2. The analysis of the eye tracking data and the details of the proposed saliency detection methods are described in Section 3. Experimental results are shown in Section 4. Finally the conclusions are given in Sections 5.

2 Scale Selectivity of Human Vision System

Scale selectivity is a visual processing property which passes different spatial frequencies. This behavior is characterized by a modulation-transfer function (MTF) which assigns an amplitude scale factor to different spatial frequency [15]. The amplitude scale factor ranges from 1.0 for spatial frequencies that are completely passed by the filter to 0.0 for spatial frequencies that are completely blocked. In certain situations, the MTF can be described by the CSF [13] which is the reciprocal of contrast sensitivity as a function of spatial frequency. This function describes the sensitivity of the human eye to sine-wave gratings at various frequencies.

The CSF tells us how sensitive we perceive different frequencies of visual stimuli. As illustrated in Fig.1 [15], the CSF curve is band pass in which spatial frequencies around 3-6 cycles/degree are best represented, while both lower and higher are poorly, and it is meaningless for that above 60 cycles/degree. That is to say, if the frequency of visual stimuli is too high, we will not be able to recognize the stimuli pattern any more. For example, the stripes in an image consisting of vertical black and white stripes are thin enough (i.e. a few thousand per millimeter), then we will not be able to see the individual stripes.

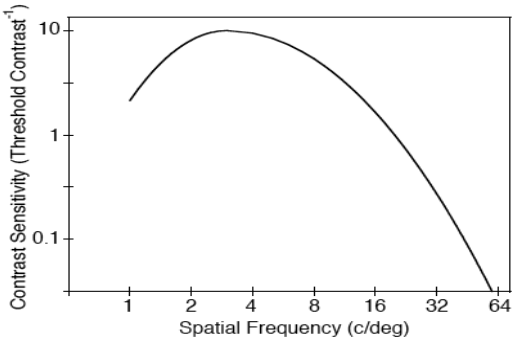


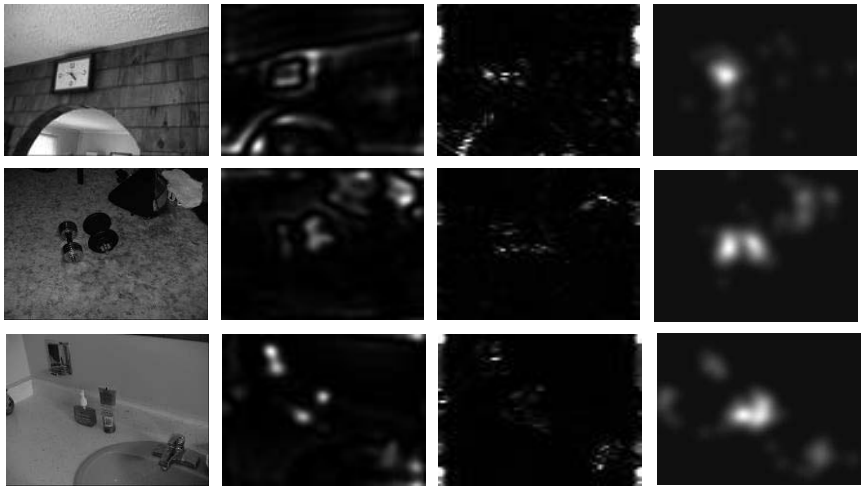
Fig. 1. Contrast-Sensitivity Function [15]

3 Saliency Detection Based on Scale Selectivity

Motivated by the observation in scale selectivity of human vision system, we propose a novel algorithm in this section. Although the form of CSF is known, the measurements of CSF only utilize sine-wave gratings, which is a simple stimulus [13]. Nevertheless, its application on complex stimuli, such as nature images, is dubious. Therefore we firstly analyze the sensitivity of the human eye at each spatial frequency according to eye tracking data over images.

3.1 Analysis of Eye Tracking Data

It is well known that any image can be represented equivalently in image space and frequency space. One can move back and forth between image space and frequency space via Fourier transformation and inverse Fourier transformation [17]. Fig.2 shows some natural images, their corresponding pictures in different frequency band and human saliency map, respectively. It is obvious that low-middle frequency components of image are closest to the human's perception when compare with the saliency map from the eye tracking data, which is consistent with previously described CSF.

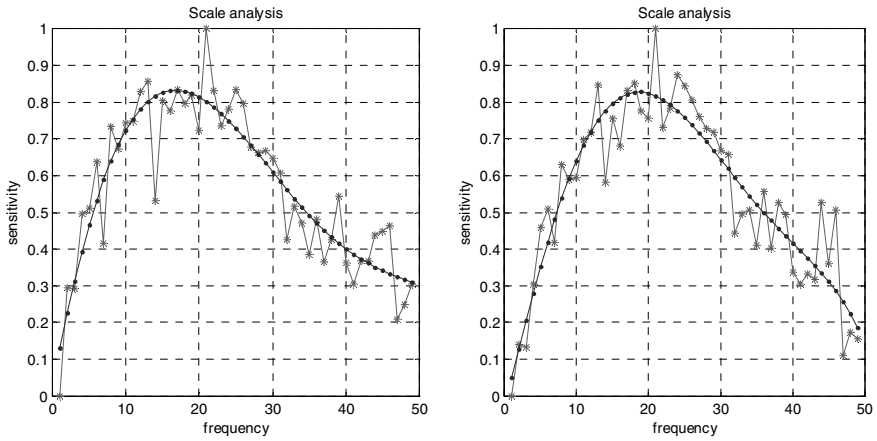


(a) (b) (c) (d)

Fig. 2. Analysis of spatial frequencies. In each group, we present (a) original image, (b) decomposition of image into low-middle frequency components, (c) decomposition of image into high frequency bands, (d) saliency map from the eye tracking data.

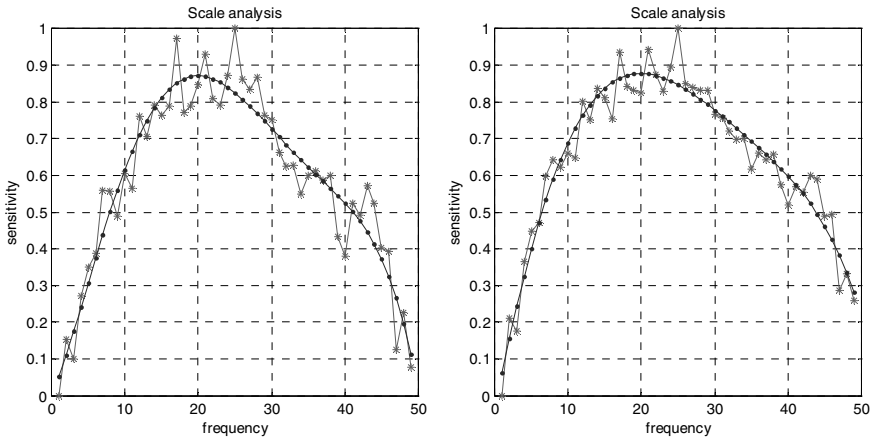
As stated before, different frequencies which human perceive are quite distinct, namely, human have intense perception in some frequency but in others have a little or scarcely sensitive. We analyze the sensitivity of the human eye at each spatial frequency on two pubic datasets proposed by Bruce et al. [5] and Judd et al. [16]. The dataset of Bruce et al. [5] contains eye fixation records from 20 subjects on 120 images of size 511×681. The dataset of Judd et al. [16] contains 1003 natural images covering a wide range of situations, and the eye fixation data is collected from 15 subjects. All of images are down-sampled to the size of 64× 86.

We use the ROC metric for quantitative analysis human sensitive on each frequency. Under the criterion, we make comparisons between every frequency map from an image and eye fixation records from humans [5, 16]. By varying the threshold, we draw the ROC curve, and the area under the curve indicates how well the frequency map of any frequency can predict the ground truth fixations. More specifically, the larger area, the more sensitive human perceive, and vice versa. According to the value of the area, we obtain a set of coefficients about human sensitivity in every frequency. Fig.3(a)(b) and Fig.4(a)(b) are the sensitivity curves showing the changes in scales which are obtained from both the whole and part of the dataset, respectively. The red curves in these figures represent the sensitivity statistic of human fixation in different scales, while the blue ones are the 5-order polynomial fitting with respect to the red. As shown in these figures, we find that the sensitivity curve of different scenarios share similar trends.



(a) (b)

Fig. 3. Sensitivity curves on different frequency (a) over 60 images and (b) the whole images in Bruce et al. [5], where red curves represent the sensitivity statistic of human fixation in different scales, and the blue ones are the polynomial fitting with respect to the red ones



(a) (b)

Fig. 4. Sensitivity curves on different frequency (a) over 300 images in Judd et al. and (b) the whole images in Judd et al. [16], where red curves represent the sensitivity statistic of human fixation in different scales, and the blue ones are the polynomial fitting with respect to the red ones

Table 1. Test results of frequency sensitivity by cross-validation

	Bruce dataset [5]	Judd dataset [16]
Bruce dataset [5]	0.71259	0.68612
Judd dataset [16]	0.71590	0.69438

According to the sensitivity coefficients from Fig.3 and Fig.4, we conduct four tests. In the first two tests, we utilize the sensitivity coefficients from part of the images as the weighted value of amplitude, and test on the other part of same dataset. In the later two, we carry out in a cross-dataset validation way, where training and testing are on different datasets. More specifically, we use each sensitivity coefficient from the whole dataset of Judd as the weighted value of amplitude, and test on Bruce dataset. Similarly, we conducted on the whole datasets of Bruce and test on Judd. Table 1 lists the four test results of cross-validation. For example, the value 0.71259 indicates the result training and testing both on Bruce dataset. As shown in Table 1, the best test result comes from training on Judd and testing on Bruce; while the worst result from training on Bruce and testing on Judd. This may due to the fact that the images in Judd dataset include more semantic objects which attract human eyes such as faces, cars, which can not explain by low level signal features only. However, difference between training results are little, which indicate that we can mimic the scale selectivity of HVS by designing a band-pass filter which passes the low-middle frequency components while suppresses the others.

3.2 Saliency Based on Scale Selection

Based on the scale selection properties of human vision system, we decompose the input image into the multi-scale bands and detect the saliency of the input scene by designing a band-pass filter which mimic the scale selectivity of human vision system.

We use Fourier transform to get a multi-scale representation of the input. The 2D formulation of Fourier function is:

$$F(u,v) = \sum_{(x,y)} I(x,y) e^{2\pi j(ux+vy)} \quad , \tag{1}$$

where (x, y) is the coordinate of the current pixel and $I(x,y)$ is the intensity function of input image. The variable u and v represent spatial frequency coordinate of natural image in horizontal and vertical directivity, respectively. The amplitude in each frequency band is represented as:

$$A(u,v) = \left| F(u,v) \right| . \tag{2}$$

Weighted value of each scale in image is equal to the corresponding coefficients. As a result, the amplitude $B(\omega)$ in each scale after weighting can be represented as:

$$B(\omega) = A(\omega) \cdot H(\omega) . \tag{3}$$

The weighted amplitude map in different frequencies bands is shown in Fig.5(c). It can be seen that the center of the amplitude map is lighter, which further demonstrate the amplitude is more intense at low-middle frequency.

Then we can acquire saliency image in each spatial frequency by Inverse Fourier Transform. The value of the saliency map is obtained by Eq.4:

$$S(x,y) = g(x,y) * F^{-1}\left[B(\omega) \cdot \exp(i \cdot P(w)) \right]^2 , \tag{4}$$

whereF^{-1} denotes Inverse Fourier Transform, $P(w)$ is the phase spectrum of the image, which is preserved during the process, and $g(x, y)$ indicates a 2D Gaussian filter to smooth the saliency map. An example of the saliency image is shown in Fig.5(d).

As stated before, human vision system are more sensitive to the low-middle frequency components. Based on the observation, we find that simply using the low-middle frequency components of image as the saliency map produces excellent result.

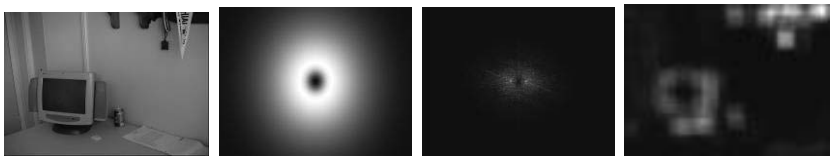
The low-middle frequency component is extracted using the following band-pass filter:

$$H(\omega) = \exp(-\omega^2 / 2\sigma_1^2) - \exp(-\omega^2 / 2\sigma_2^2), \tag{5}$$

where w is the frequency, it is represented as: $w = \sqrt{u^2 + v^2}$, σ_1, σ_2 are the variances of the Gaussian function. The relationship between σ and cut-off frequency ω_0 is described as:

$$\sigma = \omega_0 \frac{1}{\sqrt{2 \ln 2}}. \tag{6}$$

According to perception ability that HVS on different spatial frequency visual signal, we indicate that the variances of σ are 15.2 and 90.5, respectively. The band-pass filter that we defined is shown in Fig.5(b). It preserves the low-middle frequency of the image for the detection of saliency.



(a) (b) (c) (d)

Fig. 5. (a) is the original image. The corresponding weighted amplitude map (c) is computed using the band-pass filter (b). And the saliency map is shown in (d).

4 Experimental Results

In this section, we evaluate the proposed method on natural images to demonstrate its effectiveness. In the experiments, we use the image dataset and its corresponding eye fixations collected by Bruce et al. [5] as the benchmark for comparison. We down-sample the images to the size of 64× 86 pixels. The results of our model are compared with two state-of-the-art methods: information maximization approach [5] and spectral residual approach [9], as shown in Fig.6 and Fig.7.

For qualitative analysis, we show two challenging saliency detection cases. The first case (Fig.6) includes images with a large amount of textured regions in the background. These textured regions are usually neglected by the human beings, whose saliency will be obviously inhibited. For such images, we expect that only the object's pixels will be identified as salient. In Bruce et al.'s method [5], the pixels on

the objects are salient, but other pixels which are on the background are partially salient as well. As a consequence, Bruce et al.'s method is quite sensitive to textured regions. In Hou et al.'s method [9], which is somewhat better in this respect, however, many pixels on the salient objects are not detected as salient, e.g., the clock. Our method detects the pixels on the salient objects and is much less sensitive to background texture.

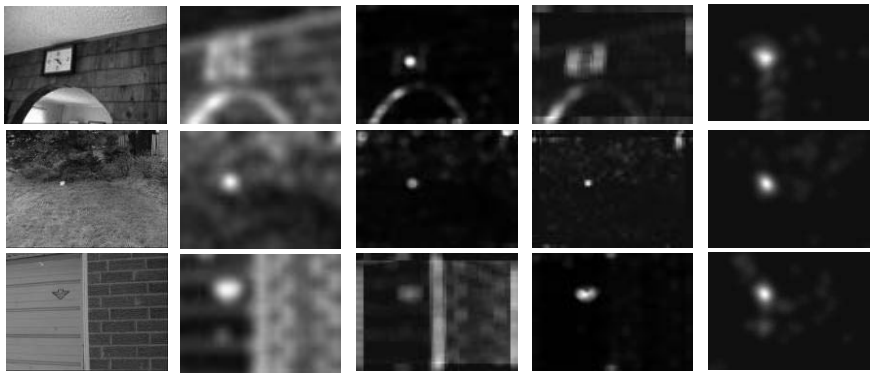


Fig. 6. Comparative saliency results on images with a large amount of textured regions in the background. The first image in each row is the original image (a), the rest saliency maps from left to right are produced by Bruce et al.[5] (b), Hou et al.[9] (c), our method (d) and human fixations (e), respectively.

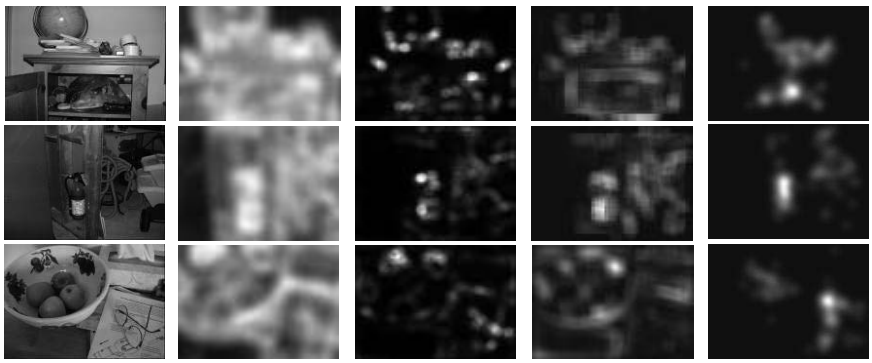


Fig. 7. Comparative saliency results on images of complex scenes. The first image in each line is the original image (a), the rest saliency maps from left to right are produced by Bruce et al.[5] (b), Hou et al.[9] (c), our method (d) and human fixations and (e), respectively.

The second case includes images of complex scenes. Fig.7 shows images of messy scene indoor. In this situation, the core objects in cluttered scene are expected as salient. It can be observed that our approach capture salient parts. For example both

the globe and the table are detected in the first scene, and the hydrant is detected in the second scene. Taking the advantage of the property of middle spatial frequency, the proposed method achieves the best visual-evaluated performances among all comparative studies.

For quantitative evaluation, we exploit Receiver Operating Characteristic (ROC) curve. The fixation data collected by Bruce et al. [5] is compared as the benchmark. From Fig.8, we could see that our algorithm outperforms other methods. In addition, we computed the Area Under ROC Curve (AUC). The average values of ROC areas are calculated over all 120 test images. And the results are reported in Table 2, which further demonstrates the superiority of the proposed method.

Table 2. Performances on static image saliency

Method	Bruce et al.[5]	Hou et al.[9]	Our method
AUC	0.6919	0.7217	0.7265

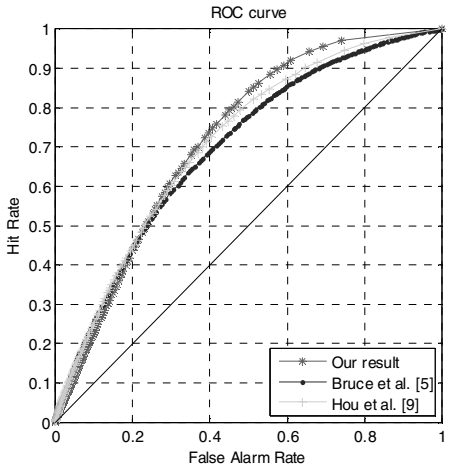


Fig. 8. ROC curves for different methods

5 Conclusion

In this paper, we propose a novel saliency detection method through scale space analysis. The saliency is based on the principle which is observed in the psychological literature: our ability to perceive a visual scene on different scales. This inspires us to detect saliency by using the low-middle frequency components and ignoring the others. Experiments on real world datasets demonstrated that our method achieves a high degree of accuracy and the computing cost is less. We would like to learn the benefits of our method in applications, such as image classification and image quality assessment in the future.

Acknowledgements. This research is partially sponsored by National Basic Research Program of China (No.2009CB320902), Beijing Natural Science Foundation (No. 4102013), Natural Science Foundation of China (Nos.60970087, 61070116 and 61070149) and President Fund of Graduate University of Chinese Academy of Sciences(No.085102HN00).

References

1. Itti, L., Koch, C., Niebur, E.: A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. IEEE Trans. on Pattern Analysis and Machine Intelligence 20(11), 1254–1259 (1998)
2. Frintrop, S., Klodt, M., Rome, E.: A real-time visual attention system using integral images. In: Proceedings of the 5th International Conference on Computer Vision Systems, Biefeled (2007)
3. Gao, D.S., Mahadevan, V., Vasconcelos, N.: The Discriminant Center-Surround Hypothesis for Bottom-Up Saliency. In: Proceedings of Neural Information Processing Systems, pp. 497–504. MIT Press, Cambridge (2008)
4. Kienzle, W., Wichmann, F.A., Schölkopf, B., Franz, M.O.: A Nonparametric Approach to bottom-Up Visual Saliency. In: Proceedings of Neural Information Processing Systems, Canada, pp. 689–696 (2006)
5. Bruce, N., Tsotsos, J.: Saliency Based on Information Maximization. In: Proceedings of Neural Information Processing Systems, Vancouver, pp. 155–162 (2006)
6. Kadir, T., Brady, M.: Saliency, Scale and Image Description. Journal of Computer Vision 45(2), 83–105 (2001)
7. Itti, L., Baldi, P.F.: Bayesian Surprise Attracts Human Attention. In: Proceedings of Neural Information Processing Systems, pp. 547–554. MIT Press, Cambridge (2005)
8. Achanta, R., Estrada, F., Wils, P., Süsstrunk, S.: Salient Region Detection and Segmentation. In: Gasteratos, A., Vincze, M., Tsotsos, J.K. (eds.) ICVS 2008. LNCS, vol. 5008, pp. 66–75. Springer, Heidelberg (2008)
9. Hou, X.D., Zhang, L.Q.: Saliency detection: A Spectral Residual Approach. In: Proceedings of Computer Vision and Pattern Recognition, Minnesota, pp. 1–8 (2007)
10. Campbell, F.: The Transmission of Spatial Information Through the Visual System. In: Schmitt, F.O., Wonden, F.G. (eds.) The Neurosciences Third Study Program. M.I.T. Press, Cambridge (1974)
11. Schyns, P., Oliva, A.: From Blobs to Boundary Edges: Evidence for Time and Spatial Scale Dependent Scene Recognition. Journal of Psychological Science 5, 195–200 (1994)
12. Navon, D.: Forest before trees: The precedence of global features in visual perception. Journal of Cognitive Psychology 9, 353–383 (1977)
13. Van Nes, R.L., Bouman, M.A.: Spatial Modulation Transfer in the Human Eye. Journal of the Optical Society of America 57, 401–406 (1967)
14. Campbell, F., Robson, J.: Application of Fourier Analysis to the Visibility of Gratings. Journal of Physiology 197, 551–566 (1968)
15. Loftus, G.R., Harley, E.M.: Why Is It Easier to Identify Someone Close Than Far Away. Journal of Psychonomic 12, 43–65 (2005)
16. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: Proceedings of International Conference on Computer Vision, Kyoto, pp. 1–8 (2009)
17. Bracewell, R.N.: The Fourier Transform and its Applications, 2nd edn. McGraw Hill, New York (1986)