REAL-TIME ENHANCEMENT OF IMAGE AND VIDEO SALIENCY USING SEMANTIC DEPTH OF FIELD

Zhaolin Su and Shigeo Takahashi
Graduate School of Frontier Sciences, The University of Tokyo, Japan

Keywords: Visual attention, Attention guidance, Saliency maps, Importance maps, Semantic depth of field.

Abstract: In this paper, we propose a method for automatically directing viewers' visual attention to important regions of images and videos in low-level vision. Inspired by the modern model of visual attention, the importance map of an input scene is automatically calculated by the combination of low-level features such as intensity and color, which are extracted using spatial filters in different spatial frequencies, together with a set of temporal features extracted using a temporal filter in case of dynamic scenes. A variable-kernel-convolution based on the importance map is then performed on the input scene, in order to make semantic depth of field effects in a way that important regions remain focused while others are blurred. The pipeline of our method is efficient enough to be executed in real time on modern low-end machines, and the associated experiment demonstrates that the proposed system can be complementary to the human visual system.

1 INTRODUCTION

We live in an information world that our sense organs receive tremendous information from surroundings, where 80% of which is visual information and can be up to 10^8 bits per second at the optic nerve. Although human visual system cannot fully process all of these information (Tsotsos, 1990), it can selectively allocates its hardware resources to focus on important regions, which made it possible to process the visual information efficiently by discontinuous fixations (Itti and Koch, 2001). Thus, the problem of how to direct viewers' attention has become a big issue when generating visual materials such as photos and videos.

For direct the human visual attention, the depth of field (DOF) effect is often employed in the art of film and photograph production. This type of effect is introduced by adjusting the properties of camera lens, so that objects with specified-distance from the camera can be displayed sharply while others are blurred intentionally. This technique is effective and can be accepted naturally by audience, because the DOF is processed by an intrinsic part of the human visual system. Previous study (Kosara et al., 2001) explored this type of effect for the use in computer visualization, named Semantic Depth of Field (SDOF), which blurs the images of the objects depend on the user-specific relevance value rather than the distance from the camera. Traditional methods for guiding human visual attention assume specific regions to be emphasized. Under some circumstances, however, we do not know where is important in the given images, and even we cannot predict it, for example, as seen in general-purpose remote monitoring systems. In this paper, we present a method for enhancing the salient image regions that are different from surroundings in intensity or colour in real time, by fully taking advantage of the SDOF formulation. Here, the salient regions correspond to the conspicuous spatial features in images and spatiotemporal features in videos, which will be extracted through our low-level vision. Our run time algorithm has been accomplished by introducing the definition of *importance map*, whose scalar value distribution topographically represents the perceptual importance of each pixel of the input visual scene. Based on this map, the essential regions of the input scene will remain focused while others will be blurred to generate semantic depth of field effects, as shown in Figure 1. Our results through eye-tracking experiments suggest that the method can reduce reaction times and increase fixation times on important regions in practice, and can help us overcome some congenital shortages of the human visual system such as change blindness phenomenon (see details in Section 4). Our implementation shows that the overall pipeline is computationally efficient enough to run in real time, thus being particularly suitable for applications that need run-time processing.

Su Z. and Takahashi S. (2010).
REAL-TIME ENHANCEMENT OF IMAGE AND VIDEO SALIENCY USING SEMANTIC DEPTH OF FIELD.
In *Proceedings of the International Conference on Computer Vision Theory and Applications*, pages 370-375
DOI: 10.5220/0002825703700375

(a) (b) (c)

Figure 1: Results of our method. (a) Input scene through a web camera. (b) Semantic depth of field effects generated by our algorithm. (c) The corresponding importance maps. The top row corresponds to a case of a static scene, where no specific object are not visually focused. The bottom row corresponds to a case a dynamic scene where the person is about to open the door. Note here the moving objects have been intentionally focus to direct our visual attention using the depth of field effects.

The remainder of this paper is organized as follows: First, Section 2 provides a brief survey on related work. Section 3 then shows the overall pipeline of our method. Section 4 introduces the implementation results and psychophysical validations by experiments. Finally, Section 5 concludes this paper.

2 RELATED WORK

In this section, we introduce related work on the computational models of visual attention to explain how the human visual selective mechanism works, and compare the conventional models of *saliency maps* and the *importance map* we will introduce. Previous studies on the SDOF will also be discussed.

2.1 Visual Attention Modelling

Although the performance of our visual system will be affected in a top-down manner by circumstances (e.g., task-depend), the low-level vision still plays an crucial role in our visual information processing and its corresponding bottom-up architecture has been proposed in (Koch and Ullman, 1985), inspired by the biological model of human visual system. This bottom-up architecture allows us to explain why some specific objects, for example, red apple among green ones, will pop-out from a scene, and further extended to the computational model of visual attention called *saliency maps* formulated by Itti et al. (Itti et al., 1998), where the centre-surround mechanism is used to extract low-level features such as colour, intensity, and orientation that are different from surrounding areas.

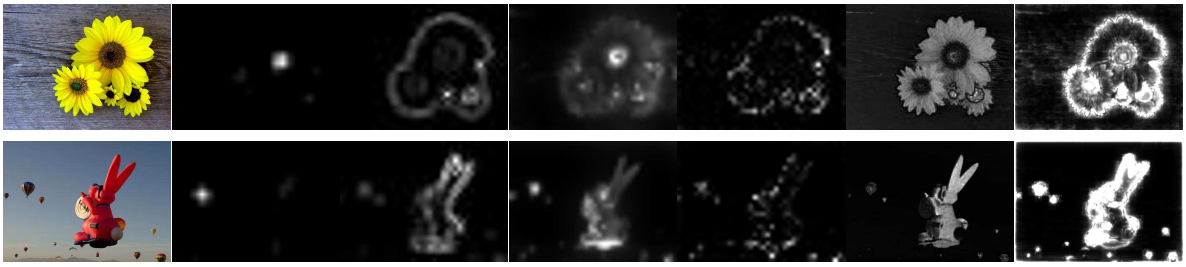
The saliency map is a one-channel scalar image that topographically represents the visual saliency of a corresponding visual scene. Based on this saliency map, a selection process deploys the gaze sequence on the visual scene by accessing to the corresponding 2D topographic scalar field using a *winner-takes-all* competition mechanism. In order to analyze video materials, Itti et al. extended their model to dynamic scenes using the Bayesian theory (Itti and Baldi, 2009).

In recent years, several purely computational models of saliency map, which are no longer based on biological principles, were also proposed. In these models, image processing methods were employed to estimate the saliency of each pixel in a given image. The basic ideas behind these methods are to calculate the centre-surround features (Ma and Zhang, 2003; Achanta et al., 2009), to maximize the mutual information of centre-surround features (Bruce and Tsotsos, 2007; Gao and Vasconcelos, 2007; Zhang et al., 2008), and to analyze frequency domain of the input image (Hou and Zhang, 2007).

2.1.1 Problems with Saliency Maps

Previous researches explored applications of the saliency map in image region segmentation, object detection, and robot vision simulation. However, the low-resolution output and heavy computational complexity of the saliency map still remain as crucial limitations for a wider utilization of the saliency map.

The reason why the most models of the saliency map produce a low-resolution output is that the



(a) (b) (c) (d) (e) (f) (g)

Figure 2: Comparison of the saliency maps in previous methods and the importance maps using our method. (a) The original image. Saliency maps using the method of (b) Itti et al. (Itti et al., 1998), (c) Ma and Zhang (Ma and Zhang, 2003), (d) Harel et al. (Harel et al., 2006), (e) Hou and Zhang (Hou and Zhang, 2007), and (f) Achanta et al. (Achanta et al., 2009). (g) The importance map using our method. The images (b)-(f) are provided courtesy of Achanta et al. (Achanta et al., 2009).

saliency map can be considered as a low-resolution abstract of the input scene (Koch and Ullman, 1985). To extract centre-surround features in different frequency domains, downsampling approach has been employed. The reason for the second issue, i.e., heavy computational complexity, is that some of the models employed a non-linear normalization process based on a biological approach (Itti et al., 1998), a graph-based approach (Harel et al., 2006), etc.

2.1.2 Saliency Maps vs. Importance Maps

In this paper, we propose the definition of an *importance map* in a hope to solve these two problems. The importance map is a scalar value image that topographically represents the perceptual importance of a visual scene. Because the ultimate goal of the importance map is to model the visual property of a given visual scene, we argue that the resolution of the importance map should be the same as that of the input image, and this can be accomplished by changing the filter size rather than the input image size (see details in Section 3.1). To reduce computational complexity, we simplified the non-linear normalization process to a linear one (see details in Section 3.2).

Another property of the importance map is that, its scalar value distribution of the importance map should be more flexible than that of the saliency map. This is because even objects that do not pop-out in a visual scene may also hold a high importance value (see details in Section 3.2). Figure 2 shows the implementation results of the saliency maps in previous methods and the importance map obtained using our method.

2.2 Semantic Depth of Field

Semantic depth of field (SDOF) proposed by Kosara et al. (Kosara et al., 2001) is a kind of Focus and Context (F+C) information visualization technique that allows us to investigate a specific local feature of the given data while identifying its position with respect to the global overview. The basic idea to assign a relevance value to every object in a visual scene first, and then blur each of them by different levels of Gaussian filters according on the corresponding relevance value. Furthermore, Kosara et al. studied the properties of the SDOF (Kosara et al., 2002b), explored its use in applications (Kosara et al., 2002a), and designed user-interactive experiments (Kosara et al., 2002b) to find that the SDOF can quickly and effectively guide viewer's attention.

3 PROPOSED METHOD

In this section, we describe our proposed method, which mainly consists of three steps. First, we extract low-level features in image intensity and colour from the input scene, together with motion feature in case of the dynamic scene. These features are then linearly normalized and combined into a single importance map. In addition, since regions that do not pop-out still may be important in perception, the histogram of importance map is adjusted by a sigmoid function to give more dynamic range according to the importance values in the map. Finally, a convolution with variable Gaussian kernels, whose coefficients change depending on the importance of the corresponding pixel, is performed on the input image.

3.1 Low-level Feature Extraction

By referring to the RGB colour components of each pixel (i.e., red (r), green (g), and blue (b)), we calculate the intensity contrast (I), red-green (RG) and blue-yellow (BY) double opponent channels as:

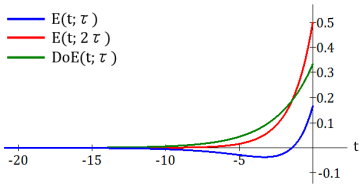


Figure 3: Plots of the DoE function with $\tau = 0.5$.

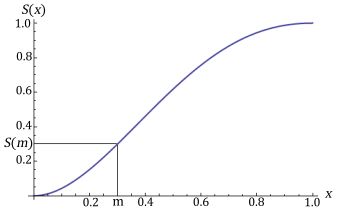


Figure 4: Plot of $S(x)$ with $m = 0.3$.

$$I = \frac{1}{3}(r+g+b), \quad (1)$$

$$RG = r-g, \quad \text{and} \quad (2)$$

$$BY = b-\frac{1}{2}((r+g)-|r-g|) \quad (3)$$

To extract the centre-surround features, we use a Difference of Gaussians (DoG) filter, which models the response of neurons in Lateral Geniculate Nucleus (LGN) at the early stages of the human visual system. Its kernel has the following form:

$$\mathrm{DoG}(x,y;\sigma) = \frac{\exp\left(-\frac{x^2+y^2}{2\sigma^2}\right)}{\sigma^2} - \frac{\exp\left(-\frac{x^2+y^2}{2(2\sigma)^2}\right)}{(2\sigma)^2} \quad (4)$$

To extract features of different spatial frequencies, we adjust the window size of the DoG filter without downsampling the input image, thus the filtered results retain the same resolution as the input image. We choose $\sigma \in \{2^1, 2^2, 2^3, 2^4, ..., 2^i\}$(pixels) for the window size, and the number of i can be changed depending on the input image resolution, because a higher resolution image may need a larger window size. Empirically, we choose $i = 5$ in our implementation.

Let $\mathrm{Input}(c)$ denote one of the three channels ($c \in \{I, RG, BY\}$), i denote the scale of the DoG filter, and $*$ denote the convolution operator. We derive $3 \times i$ spatial feature maps $\mathcal{F}_s(c,i)$:

$$\mathcal{F}_s(c,i) = \mathrm{Input}(c) * \mathrm{DoG}(\sigma_i). \quad (5)$$

For videos, we use a difference of Exponential-like temporal filter (Zhang et al., 2009) to extract motion features, which is given by (Figure 3)

$$\mathrm{DoE}(t;\tau) = E(t;2\tau) - E(t;\tau), \quad (6)$$

where

$$E(t;\tau) = \frac{\tau}{1+\tau} \cdot (1+\tau)^t. \quad (7)$$

Here, $t \in (-\infty, 0]$ is the frame number relative to the current frame (i.e., $t = 0$ corresponds to the current frame), and τ is the shape parameter. A bigger value τ means that the extracted motion features will receive more influences from the immediate neighbor frame, but not the distant past, while the condition $\tau \in [0.1, 1.0]$ works well in our implementation. Applying this filter to $\mathcal{F}_s(c,i)$ in Eq. (5), we derive another set of $3 \times i$ temporal feature maps $\mathcal{F}_t(c,i)$ for the current frame as:

$$\mathcal{F}_t(c,i) = \mathcal{F}_s(c,i) * \mathrm{DoE}(0;\tau). \quad (8)$$

3.2 Importance Maps

As we described above, the purpose of computing importance maps is to evaluate the perceptual importance of every pixel in the input scene. Since distinctive features should be more important than common ones (Zhang et al., 2008), we normalize the aforementioned feature maps by its self-average, and then linearly combined the results into one single map. For static images, the output I' is obtained as:

$$I' = \sum \frac{\mathcal{F}_s(c,i)}{avg(\mathcal{F}_s(c,i))}. \quad (9)$$

For video inputs, with extra motion feature responses, I' is reformulated as:

$$I' = \sum \frac{\mathcal{F}_s(c,i)}{avg(\mathcal{F}_s(c,i))} + k \cdot \frac{\mathcal{F}_t(c,i)}{avg(\mathcal{F}_t(c,i))}, \quad (10)$$

where k is a weight value that determines the proportion of the motion features in I'. This value can be adjusted by circumstances.

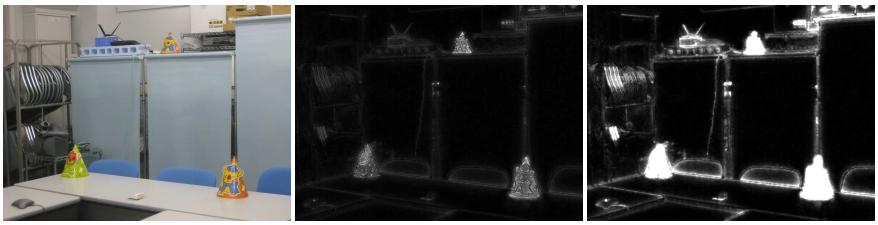
We then normalize I' into the range $[0,1]$, and adjust the histogram of I' using a fourth-order sigmoid interpolation polynomial (Figure 4) to remap every pixel x in I' to a new value $S(x)$ as follows:

$$S(x) = \frac{1}{d}(ax^2 + bx^3 + cx^4), \quad (11)$$

where

$$a = -1+4m^2-3m^3, b = 2-4m+2m^3,$$
$$c = -1+3m-2m^2, \text{ and } d = -1+2m^2-m^3. \quad (12)$$

Here, $x \in [0,1]$, $S(x) \in [0,1]$. m is the threshold value that satisfies $S(m) = m$, from which we enhance the dynamic range of the importance values. Empirically



(a) (b) (c)

Figure 5: Adjustment by the function $S(x)$. (a) Input image. (b) Importance map before the adjustment. (c) Importance map after the adjustment. Note that several objects (such as a blue basket on the upper left) gain high importance values while other objects that already pop out are further enhanced.

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

Figure 6: Change in the distribution of gaze fixation times. (a) An original image and (b) its corresponding distribution of gaze fixation times. (c) The enhanced image and (d) its corresponding distribution of gaze fixation times. Warning signs direct more visual attention due to the depth of field effects.

we choose half the average of I' as the threshold in our implementation. This transformation let us increase the importance values above m while suppressing the values below m, which exaggerates the contrast of I' and results in a more clearly enhanced output. Figure 5 shows the final importance map I obtained by applying this adjustment.

3.3 Applying Depth of Field Effects

The last step of the method is to reconstruct every pixel of the input scene using a variable-kernel-convolution to accomplish the final DOF effects. The kernel we used is a Gaussian function with $\sigma_{x,y}$:

$$G(x,y;\sigma_{x,y}) = \frac{1}{2\pi\sigma_{x,y}^2} \cdot \exp\left(-\frac{x^2+y^2}{2\sigma_{x,y}^2}\right) \qquad (13)$$

where $\sigma_{x,y}$ is proportional to the inverse of the local importance as:

$$\sigma_{x,y} = \eta \cdot \frac{1}{I_{x,y}}. \qquad (14)$$

Here, $I_{x,y}$ stands for the scalar value of the pixel (x,y) of I, and η is a weight coefficient that controls the degree of blur effects. In our implementation, we chose $\eta = 5$. After applying this variable-kernel-convolution, we obtain the final output as:

$$\text{Output} = (\text{Input} * G)(x,y). \qquad (15)$$

4 RESULTS

We implemented our method in C++ and tested it on a notebook with a 2.4GHz Intel CPU and 2GB memory. The results show that our system can comfortably process videos at 160×120 resolution in real time (about 40fps), and videos at 320×240 resolution in semi real time (about 16fps). The experiments also suggest that the computation time is roughly proportional to the number of pixels in the input image. Figure 1 shows the results of static and dynamic scenes, which were captured with an ordinary web camera.

Since the DOF effect is advantageous in that it influences each colour channel independently (Kosara et al., 2002a), our method can be applied to video clips that have only one intensity channel, for example, such as greyscale videos obtained through video surveillance systems. We also tested our system by applying it to videos that were used as stimuli in change blindness experiments. We found that our system can emphasize gradual changes such as objects slowing changing in its position, colour, and shape, while we often fail to detect these changes without such visual enhancements (Simons, 2000).

To evaluate the effectiveness of our method, we

also conducted eye-tracking experiments, in order to investigate how our method can aggressively direct the visual attention of the viewers. The original and visually enhanced versions of a natural scene were displayed in a random sequence with the resolution of 800x600 pixels. Subjects (8 males and 2 females) were asked to freely look at each image for 10 seconds while their gaze movements were recorded with a Tobii X120 non-intrusive eye tracker. Figure 6 shows the comparison of the results, which indicates the fact that we pay more visual attention to the warning signs on the top right in the enhanced version of the scene. Careful analysis of the recorded gaze movements also suggests that the first gaze fixation on the warning sizes has been reduced from 6.15s to 4.95s in average due to the DOF effects, while we count the time as 10.0s in the case that the subject did not notice the warning signs.

5 CONCLUSIONS

In this paper, we have first introduced the definition of the importance map, which represents the perceptual importance of a visual scene by extraction and combination of low-level features. Based on this importance map, we enhance the salient features in the input scene by applying semantic depth of field effects to naturally guide the visual attention of the viewers. The whole pipeline can be executed in real time without any user-intervention, and the experiment results suggest that our method can actually assist people in rapidly finding significant features in the scene.

ACKNOWLEDGEMENTS

We would like to thank Radhakrishna Achanta et al. for sharing their implementation results of previous researches, and anonymous reviewers for their valuable comments. This work has been partially supported by Japan Society of the Promotion of Science under Grants-in-Aid for Scientific Research (B) No. 20300033 and No. 21300033.

REFERENCES

Achanta, R., Hemami, S., Estrada, F., and Susstrunk, S. (2009). Frequency-tuned Salient Region Detection. In *Proc. IEEE International Conf. Computer Vision and Pattern Recognition (CVPR2009)*, pages 1597–1604.

Bruce, N. and Tsotsos, J. (2007). Attention based on information maximization. *Journal of Vision*, 7(9):950.

Gao, D. and Vasconcelos, N. (2007). Bottom-up saliency is a discriminant process. In *Proc. IEEE International Conf. Computer Vision (ICCV2007)*, pages 1–6.

Harel, J., Koch, C., and Perona, P. (2006). Graph-based visual saliency. In *Proc. Neural Information Processing Systems (NIPS2006)*, pages 570–577.

Hou, X. and Zhang, L. (2007). Saliency detection: A spectral residual approach. In *Proc. IEEE International Conf. Computer Vision and Pattern Recognition (CVPR2007)*, pages 1–8.

Itti, L. and Baldi, P. F. (2009). Bayesian surprise attracts human attention. *Vision Research*, 49(10):1295–1306.

Itti, L. and Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–203.

Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 20(11):1254–1259.

Koch, C. and Ullman, S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*, 4(4):219–227.

Kosara, R., Miksch, S., and Hauser, H. (2001). Semantic depth of field. In *Proc. IEEE Symp. Information Visualization 2001 (INFOVIS2001)*, pages 97–104.

Kosara, R., Miksch, S., and Hauser, H. (2002a). Focus+Context taken literally. *IEEE Computer Graphics and Applications*, 22(1):22–29.

Kosara, R., Miksch, S., Hauser, H., Schrammel, J., Giller, V., and Tscheligi, M. (2002b). Useful properties of semantic depth of field for better F+C visualization. In *Proc. Symp. Data Visualisation 2002 (VISSYM2002)*, pages 205–210.

Ma, Y.-F. and Zhang, H.-J. (2003). Contrast-based image attention analysis by using fuzzy growing. In *Proc. 11th ACM International Conf. Multimedia (MULTIMEDIA2003)*, pages 374–381.

Simons, D. J. (2000). Current approaches to change blindness. *Visual Cognition*, 7:1–15.

Tsotsos, J. (1990). Analyzing vision at the complexity level. *Behavioral and Brain Sciences*, 13(3):423–445.

Zhang, L., Tong, M. H., and Cottrell, G. W. (2009). Sunday: Saliency using natural statistics for dynamic analysis of scenes. In *Proc. 31st Annual Cognitive Science Society Conf. (CogSci2009)*, pages 2944–2949.

Zhang, L., Tong, M. H., Marks, T. K., Shan, H., and Cottrell, G. W. (2008). SUN: A Bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7):1–20.