A COMPUTATIONAL MODEL OF SALIENCY DEPLETION/RECOVERY PHENOMENA FOR THE SALIENT REGION EXTRACTION OF VIDEOS

Clement Leung[*†], *Akisato Kimura*[†], *Tatsuto Takeuchi*[†] *and Kunio Kashino*[†]

[†] NTT Communication Science Laboratories, NTT Corporation, Japan
[*] Department of Electrical and Computer Engineering, University of British Columbia, Canada
URL: http://www.brl.ntt.co.jp/people/akisato/

ABSTRACT

This report proposes a new algorithm for extracting salient regions of videos by introducing two important properties of the early human visual system: 1.) *Instantaneous* saliency depletion with *gradual* recovery, whereby saliency is instantaneously suppressed and gradually recovered in previously attended regions. 2.) *Gradual* saliency depletion with *instantaneous* recovery, whereby saliency is gradually decreased over time in non-surprising regions and at the same time recovered in surprising locations. With the introduction of these properties, redundant information in videos can be suppressed and important information is eventually enhanced.

1. INTRODUCTION

Visual attention in computer vision aims to mimic the ability of the human visual system to select just the relevant aspects from a broad visual input. The advantages of human vision over computer vision systems, which include robustness, flexibility and efficiency, may partly be to the result of this mechanism. The use of attention when selecting relevant data has several benefits. First, the amount of data to be processed is reduced, resulting in lower computational costs. Second, distracting information can be suppressed so that only relevant data influence the activities of the system. From this standpoint, simulating visual attention constitutes a central part of computer vision systems because they select the information on which the system activities are based.

Previous studies have employed this approach based on the characteristics of human visual attention. For example, Ma et al. [1] developed a user attention model that is employed to approximate the attention span of people viewing video content. However, this model is mainly based on high-level human attributes, which depend strongly on an individual's knowledge, experiences and preferences. Thus the model may not be sufficiently versatile for use in a wide range of applications. In contrast, the fundamental properties of the early human visual system provide a bottom-up approach capable of retrieving important information from various kinds of video content. An early computational model for explaining the human attention system was proposed by Koch and Ullman [2]. This model analyzes still images to produce fundamental features, such as intensity, color and orientation, which are combined to form a *saliency map* that represents the relevance of visual attention. Many attempts [3, 4, 5] have been made to improve the Koch-Ullman model. However, visual attention models for videos have not been fully investigated. It is well known that human sensitivity to such visual features varies with time. Given that sensitivity to saliency depends strongly on the temporal dynamics of the early visual system, such temporal characteristics should be introduced if we are to realize further improvements.

This report proposes a new algorithm for extracting the saliency of videos based on the above considerations. The proposed algorithm models two contrastive properties of the temporal dynamics of the early human visual system: 1) Instantaneous saliency depletion with gradual recovery, which simulates the *"Inhibition of Return"* effect. [6] This effect tends to result in a delay (approximately 500-900 ms) when humans notice salient events occuring around a previously focusing region after attention has been diverted away from that region. 2) Gradual saliency depletion with instantaneous recovery, which is derived from the *"Neural Adaptation"* theorem [7]. Based on this theorem, sensitivity to saliency gradually decreases over time when no surprising events occur in a video, and it is only retained in surprising locations in the video. Each of the above properties can be modeled as a combination of two masks, where one simulates saliency depletion, and the other mimics saliency recovery. The explicit construction of such masks is described in the following sections, and constitutes the main contribution of this report.

2. BASIC ALGORITHM STRUCTURE

Our algorithm has five parts: a) extracting fundamental saliency maps, b) generating *instantaneous saliency depletion (ISD) masks*, c) generating *gradual saliency depletion (GSD) masks*, d) masking fundamental saliency maps with ISD and GSD masks, and e) lastly extracting depleted saliency videos.

The first author contributed to this work during his internship at NTT.

Fundamental saliency maps are extracted based on Itti's strategy [3, 8], which creates Gaussian pyramids for fundamental features such as intensity, color, orientation, flicker, and motion to create a conspicuity map $CM_j(i)$, where i corresponds to a frame and j corresponds to a feature type, e.g. intensity. The conspicuity maps are normalized, and summed up into a fundamental saliency map $S(i)$. The sequence $\{S(i)\}_i$ of fundamental saliency maps comprises a fundamental saliency video. ISD masks $\{ID(i)\}_i$ and GSD masks $\{GD(i)\}_i$ are generated from fundamental features, conspicuity maps and saliency maps. The construction of these masks will be detailed in the subsequent sections. Every fundamental saliency map $S(i)$ is multiplied with the corresponding ISD mask $ID(i)$ and GSD mask $GD(i)$ to form a depleted saliency map $S_D(i)$, as follows:

$$S_D(i) \quad = \quad ID(i)^{\omega_I} \cdot GD(i)^{\omega_G} \cdot S(i),$$

where $\omega_I \in [0, 1]$ and $\omega_G \in [0, 1]$ are weighting constants for the ISD and GSD masks, respectively.

3. INSTANTANEOUS SALIENCY DEPLETION

3.1. Overview

Instantaneous saliency depletion along with gradual recovery simulates the "Inhibition of return" effect. Instantaneous saliency depletion has been implemented in a previously reported algorithm solely for still images [3]. However, the previous implementation did not consider any transitions of saliency depletion. We have extended the idea of instantaneous saliency depletion to videos considering the temporal dynamics of instantaneous depletion. Instantaneous saliency depletion with gradual recovery is implemented by creating two masks for each frame of the saliency video: the *MSR depletion mask* $ID_1(i)$ and the *recovery mask* $ID_2(i)$. The MSR depletion mask creates the instantaneous depletion of saliency at the most salient regions (*MSRs*) of the fundamental saliency video. Simultaneously, the recovery mask gradually retains saliency in the fundamental saliency video. The two masks are combined to form an overall ISD mask $ID(i)$.

$$ID(i)_{(x,y)} \quad = \quad \min\{1, ID_1(i)_{(x,y)} + ID_2(i)_{(x,y)}\},$$

where $ID(i)_{(x,y)}$ is the pixel value at the position (x, y) in the image $ID(i)$.

3.2. Mask construction

The MSR depletion mask $ID_1(i)$ is a gray-scale image created with all pixel values initially at 1. For every frame i, the MSR depletion mask $ID_1(i)$ is generated from the ISD mask $ID(i-1)$ of the previous frame such that all the pixel values within the MSR of the fundamental saliency video frame $S(i-1)$ are set at 0. The circular area is defined as a *depletion circle* $MSR(i-1)$ with a radius of $r = 25$ pixels. Every depletion circle $MSR(i)$ moves in accordance with the average optical flow values. Every optical flow value is calculated by the Lukas-Kanade method [9].

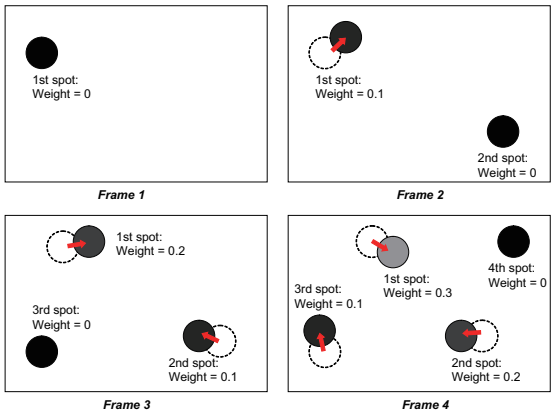

Fig. 1. *Example of ISD mask*

The recovery mask $ID_2(i)$ is a gray-scale image with all the pixels at a constant value of $\alpha = 1/10$. The recovery mask is added to the ISD mask $ID(i)$ for every consecutive frame, causing depletion circles to regain their value. The depletion circles eventually reach the original mask value of 1 after $1/\alpha = 10$ frames (= 667 msec when the frame rate of the fundamental saliency video is 15 frames/sec).

Fig.1 shows an example of ISD masks created by the above procedures. In frame 1, the MSR of the previous frame is depleted with the MSR depletion mask. In the next frame, the MSR of the previous frame is depleted, while the first depletion circle moves according to average optical flows and its value is restored by the recovery mask. Such a pattern is repeated in the subsequent frames.

4. GRADUAL SALIENCY DEPLETION

4.1. Overview

Gradual saliency depletion with instantaneous recovery simulates the "neural adaptation" theorem, and it is developed by creating two masks for each frame of the fundamental saliency video: the *whole-region depletion mask* $GD_1(i)$ and the *surprise mask* $GD_2(i)$. The whole-region depletion mask induces the gradual depletion of saliency in the saliency videos. Simultaneously, the surprise mask retains full saliency for surprising areas. The two masks are combined to form a gradual saliency depletion mask (GSD mask) $GD(i)$.

$$GD(i)_{(x,y)} \quad = \quad \min\{1, GD_1(i)_{(x,y)} + GD_2(i)_{(x,y)}\}.$$

4.2. Mask construction

The whole-region depletion mask $GD_1(i)$ is a gray-scale image with all the pixel values initially at 1. In each frame a

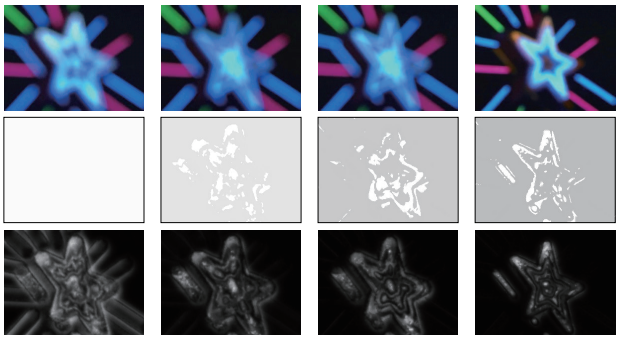

Fig. 2. *Example of GSD masks. Top row: original frames, mid row: GSD masks, bottom row: saliency video with GSD masks.*

constant value $\beta = 0.0025$ is subtracted from each pixel in the whole-region depletion mask unless no scene change occurs in the frame. If a scene change is detected in the frame, the whole-region depletion mask regains its values.

The surprise mask $GD_2(i)$ indicates where surprising events occur. Previous work [5] used a probabilistic approach to compute surprise. However, this technique requires a large amount of calculation. By contrast, our approach involves calculating the temporal response for each conspicuity map. This strategy is inspired by the center-surround subtraction of the Gaussian pyramids used for extracting conspicuity maps in Itti's strategy [3]. Difference-of-Gaussian (DOG) filters $G_{(c,s)}$ are convoluted with each conspicuity map sequence to form a temporal response map $T_j(i)$, as follows:

$$\begin{aligned} T_j(i) &= 2\sum_{c=2}^{4}\sum_{s=c+3}^{c+4}\sum_{k=0}^{8} G_{(c,s)}(k)\cdot CM_j(i-k), \\ G_{(c,s)}(k) &= G_{c/2}(k) - G_{s/2}(k), \end{aligned}$$

where j corresponds to a feature type, and G_σ is a Gaussian distribution with mean 0 and variance σ^2. The temporal response maps are summed up and binarized with the threshold θ to form the surprise mask $GD_2(i)$.

Fig.2 shows example GSD masks created by using the above procedures. The GSD masks become darker owing to the whole-region depletion masks, while the surprise masks retain saliency values in surprising regions.

5. EXPERIMENTS

5.1. Conditions

To evaluate the performance of our algorithm relative to previous algorithms [3, 8], we chose five subjects and had them view six different sample video clips. The eye movement of each subject was tracked using an eye tracking device [10] and the subjects viewed each video twice. We gave the subjects few instructions when they were viewing the sample videos. The eye tracking data were compared with the saliency

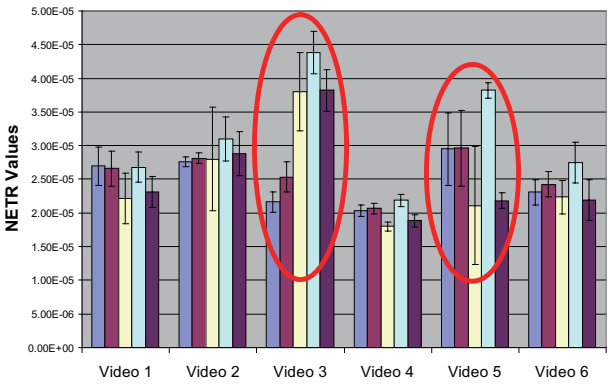

Fig. 3. *Experimental results: Bars from left to right for each video: Still image algorithm, moving algorithm, our algorithm case 1(ISD mask), case 2 (GSD mask) and case 3 (ISD+GSD mask)*

videos produced by each algorithm to determine which algorithm best mimicked the human visual system. We tried three sets of parameters: 1.) $(\omega_I, \omega_G) = (1, 0)$, which corresponds to the saliency video with the ISD mask, 2.) $(\omega_I, \omega_G) = (0, 1)$, which corresponds to the saliency video with the GSD mask, 3.) $(\omega_I, \omega_G) = (1, 1)$, namely both the ISD and GSD masks were included.

5.2. Comparing eye tracking data with saliency videos

The eye movement may contain some noise because of "small eye movements of fixations" [11]. Thus, instead of the eye tracking data, we use the *eye tracking region* $ETR(i;k,n)$ defined as a circular area whose center is the position on which each subject's eye is focused and whose radius is δ, where i, k and n correspond to frame number, video, and subject, respectively.

The average pixel value in the ETR (*ETR value*) was calculated for each frame of a saliency video. The ETR values were normalized by dividing by the sum of the frame's pixel values for each saliency video, to obtain normalized ETR values (*NETR values* $V(i;k,n)$).

$$V(i;k,n) = \frac{\sum_{(x',y')\in ETR(i;k,n)} S_D(i)_{(x',y')}}{\sum_{(x',y')} S_D(i)_{(x',y')}},$$

Larger NETR values indicate that the eye is being directed towards locations with high values in the saliency video. Thus, we defined the best performance as that derived from the algorithm providing the largest NETR value.

Fig.3 shows the experimental results, which include NETR values for each video averaged over the subjects, along with standard errors. This figure implies that the average NETR values combined for each video were almost the same or substantially better in the proposed algorithm with gradual depletion than with the previous algorithms, while instantaneous

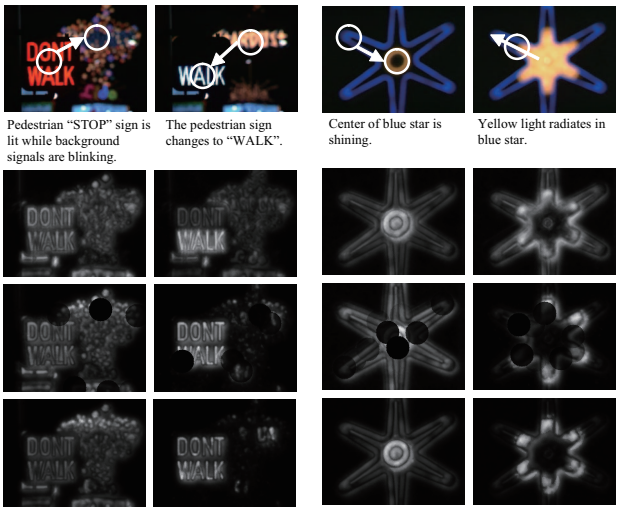

Fig. 4. *Saliency videos produced from Video 3 (left half) and Video 5 (right half). Top row: original video with circles and arrows that show typical focusing points and eye movements, respectively, second row: moving algorithm, third row: our algorithm case 1, bottom row: our algorithm case 2.*

depletion sometimes degraded the performance of the proposed algorithm.

In the following, we discuss the operation of the proposed algorithm in detail for videos where the performance differed significantly from that obtained with the previous algorithms. Fig.4 shows saliency videos in such cases.

In all cases, our algorithm outperformed the previous algorithms for Video 3. Fig.4 (left half) shows that all the lighting regions in the sample video are enhanced when the previous algorithm was used (see the second row), while in every case with the proposed algorithm, non-surprising regions are suppressed (see the third and bottom row). When utilizing gradual depletion, saliency depletion in non-surprising regions came directly from the whole-region depletion mask $GD_1(i)$. Similar effects also occurred in the algorithm with instantaneous depletion due to the MSR depletion mask $ID_1(i)$.

For Video 5, however, the performance of the proposed algorithm depended on whether or not the ISD mask was included. As shown in the third row of Fig.4 (right half), salient regions in the previous algorithms are fully covered with MSR depletion masks $ID_1(i)$ in the algorithm with instantaneous depletion. In contrast, gradual saliency depletion worked well, namely non-surprising regions are depleted by whole-region depletion masks $GD_1(i)$ in the algorithm with gradual depletion, as shown in the bottom row of the Fig.4.

6. CONCLUDING REMARKS

We have proposed a new computational model of visual attention by incorporating two contrastive properties of the early human visual system, namely instantaneous saliency depletion with gradual recovery, and gradual saliency depletion with instantaneous recovery. We evaluated the proposed algorithm by comparison with previously reported algorithms and found that on average the proposed algorithm with gradual depletion performed better than the previous algorithms in mimicking the human visual system. The proposed algorithm with instantaneous depletion improved the performance for some videos.

Acknowledgements

The authors wish to thank, Dr. Shin'ya Nishida, Dr. Isamu Motoyoshi, and Dr. Junji Yamato of NTT Communication Science Laboratories for their valuable comments and helpful discussions, which helped improve this work.

7. REFERENCES

[1] Y. Ma, X. Hua, L. Lu, and H. Zhang, "A generic framework of user attention model and its application in video summarization," *IEEE Trans. Multimedia*, vol. 7, no. 5, pp. 907–919, October 2005.

[2] C. Koch and S. Ullman, "Shifts in selective visual attention: Towards the underlying neural circuity," *Human Neurobiology*, vol. 4, pp. 219–227, 1985.

[3] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, November 1998.

[4] C. M. Privitera and L. W. Stark, "Algorithms for defining visual regions-of-interest: Comparison with eye fixations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 9, pp. 970–982, 2000.

[5] L. Itti and P. Baldi, "A principled approach to detecting surprising events in video," in *Proc. Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2005, pp. 631–637.

[6] M. I. Posner and Y. Cohen, "Components of visual orienting," in *Attention and Performance*, H. Bouma and D. G. Bouwhiuis, Eds., vol. 10, pp. 531–556. Erlbaum, 1984.

[7] H. K. Hartline, "The nerve messages in the fibers of the visual pathway," *Journal of Optics Society of America*, vol. 30, pp. 239–247, 1940.

[8] L. Itti, N. Dhavale, and F. Pighin, "Realistic avatar eye and head animation using a neurobiological model of visual attention," in *Proc. SPIE International Symposium on Optical Science and Technology*, August 2003, vol. 5200, pp. 64–78.

[9] B. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. of International Joint Conference on Artificial Intelligence*, 1982, pp. 674–679.

[10] T. Ohno, N. Mukawa, and A. Yoshikawa, "FreeGaze: A gaze tracking system for everyday gaze interaction," in *Proc. Symposium on ETRA*, 2002, pp. 125–132.

[11] R. W. Ditchburn and B. L. Ginsborg, "Vision with a stabilized retinal image," *Nature*, vol. 170, pp. 36–37, July 1952.