

# A computational model of saliency depletion/recovery phenomena for the salient region extraction of videos

Clement LEUNG<sup>†,††</sup>, Akisato KIMURA<sup>††</sup>, Tatsuto TAKEUCHI<sup>††</sup>, and Kunio KASHINO<sup>††</sup>

<sup>††</sup> NTT Communication Science Laboratories, NTT Corporation Morinosato Wakamiya 3-1, Atsugi,  
Kanagawa, 243-0198 Japan

<sup>†</sup> Department of Electrical and Computer Engineering, University of British Columbia 2329 West Mall  
Vancouver, British Columbia, V6T1Z4 Canada

**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. The proposed algorithm has been evaluated with an eye tracking device to see how well it fits the human visual system. The results show that the proposed algorithm substantially outperformed previous algorithms when only gradual depletion was incorporated, and instantaneous depletion improved the performance in some cases.

**Key words** video processing, visual attention, early human visual system, saliency depletion, inhibition of return, neural adaptation

## 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. Advantages of human vision over artificial vision systems such as robustness, flexibility and efficiency may partly be due to this mechanism. The use of attention when selecting relevant data has several advantages. First, the amount of data to be processed is reduced, resulting in lower computational costs. More resources are available for the prominent inputs. Second, distracting information can be suppressed so that only the relevant data influence the activities of the system. From this standpoint, simulating visual attention constitutes a central part of the system because it selects the information on which the system activities are based.

Several 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 used to approximate the attention span of humans viewing video content. The architecture of the model uses object motion, camera motion, audio, and textures. 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 to 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 early 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 other attempts [3], [4] have been made to adjust and improve the Koch-Ullman model. Later, the model was extended to videos by adding flicker and motion features [5], [6]. 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. Considering that sensitivity to saliency depends strongly on the temporal dynamics of the early visual system, such temporal characteristics should be introduced to realize further improvements.

This report proposes a new algorithm for extracting the saliency of videos based on the above considerations. The

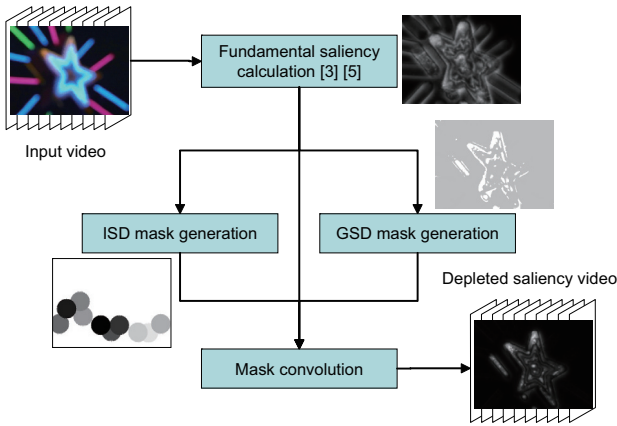


Figure 1 Basic structure of proposed method

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. [7], [8] that facilitates rapid and efficient understanding of visual scenes. Owing to this effect, humans tend to get delayed to realize salient events happened around the previously focusing region after attention is diverted away from the region. 2) Gradual saliency depletion with instantaneous recovery, which is derived from the “*Neural Adaptation*” theorem [9], [10]. 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.

It should be noted that the above-mentioned properties of the human visual system are also contrastive in terms of the situations in which the corresponding properties occur. The “inhibition of return” phenomenon often occurs in “*active viewing*” situations, which means that one wants or needs to capture as much information as possible from one’s view as quickly as possible. In contrast, the “neural adaptation” phenomenon usually occurs in “*passive viewing*” situations, where one does not need to obtain any information or has already captured almost all relevant information from one’s view.

Each of the above properties can be modeled as a combination of two masks, one of which simulates saliency depletion, the other of which mimics saliency recovery. An explicit construction of such masks is described in the following sections, which is the main contribution of this paper.

## 2. Basic algorithm structure

Fig.1 shows the basic structure of our algorithm, which is composed of 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

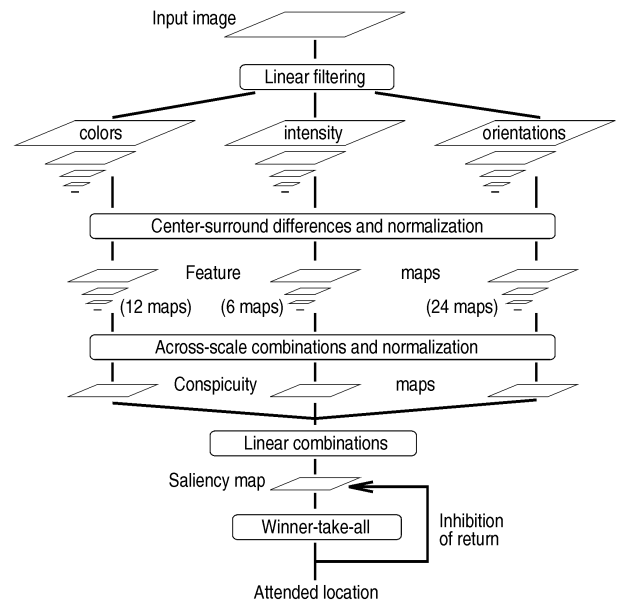


Figure 2 Fundamental saliency maps (quoted from [3])

GSD masks, and e) lastly extracting depleted saliency videos.

Fundamental saliency maps  $S(i)$  are extracted from each frame  $i$  of a video based on the previous algorithms, which will be detailed in Section 3. ISD masks  $ID(i)$  and GSD masks  $GD(i)$  are generated from fundamental saliency maps. The construction of these masks will be detailed in Sections 4. and 5.. 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:

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

where  $\omega_I \in [0, 1]$  and  $\omega_G \in [0, 1]$  are weighting constants for the ISD and GSD masks, respectively, and  $S(i)_{(x,y)}$  is the value on the position  $(x, y)$  of the image  $S(i)$ .

## 3. Fundamental saliency maps

Fundamental saliency maps are extracted based on the previous algorithm proposed by Itti et al. [3], [5]. Fig. 2 illustrates the algorithm. In the following, we will review how to obtain fundamental saliency maps.

First, intensity, color, orientation, flicker, and motion feature images are extracted for each frame  $i$  of a video. With  $r(i)$ ,  $g(i)$  and  $b(i)$  being the red, green and blue channels of the frame  $i$  of the input video, an intensity image  $I(i)$  is obtained as  $I(i) = (r(i) + g(i) + b(i))/3$ . Four broadly-tuned color images are also created:  $R(i) = r(i) - (g(i) + b(i))/2$  for red,  $G(i) = g(i) - (r(i) + b(i))/2$  for green,  $B(i) = b(i) - (r(i) + g(i))/2$  for blue,  $Y(i) = (r(i) + g(i))/2 - |r(i) - g(i)|/2$  for yellow. Four orientation images  $O_\phi(i)$  are obtained from the intensity image using oriented Gabor filters  $g_\phi$ , where  $\phi = 0, \pi/4, \pi/2, 3\pi/4$  is the preferred orientation. A flicker image  $F(i)$  is also obtained from the intensity images as

$F(i) = I(i) - I(i - 1)$ . Two motion images  $M_x(i)$  and  $M_y(i)$  are obtained using the Lukas-Kanade method [12], where each subscript  $x$  and  $y$  corresponds to horizontal and vertical component, respectively.

Each feature image is used to create an Gaussian pyramid, e.g.  $I(i; l)$  for the intensity feature image, where  $l = 0, 1, \dots, 8$  indicates the scale and the scale 0 is the original scale. Center-surround difference (namely interpolation to the finer scale and point-by-point subtraction) between a ‘‘center’’ fine scale  $c \in \{2, 3, 4\}$  and a ‘‘surround’’ coarser scale  $s = c + d$ ,  $d \in \{3, 4\}$  yield the feature maps, obtained as follows:

$$RS_I(i; c, s) = |I(i; c) - I(i; s)|$$

$$RS_{RG}(i; c, s) = |(R(i; c) - G(i; c)) - (R(i; s) - G(i; s))|$$

$$RS_{BY}(i; c, s) = |(B(i; c) - Y(i; c)) - (B(i; s) - Y(i; s))|$$

$$RS_O(i; \phi; c, s) = |O_\phi(i; c) - O_\phi(i; s)|$$

$$RS_F(i; c, s) = |F(i; c) - F(i; s)|$$

$$RS_{M_k}(i; c, s) = |M_k(i; c) - M_k(i; s)| \quad (k = x, y)$$

Each feature map is normalized by the following function  $N(\cdot)$ :

$$N(FM(i)) = \frac{(m^* - \bar{m})^2}{m^*} FM(i),$$

where  $FM(i)$  is a feature map,  $m^*$  is the global maximum of the map, and  $\bar{m}$  is the local maxima average. Normalized feature maps of each feature are summed up to form a conspicuity map obtained as follows:

$$CM_I(i) = \sum_{c=2}^4 \sum_{s=c+3}^{c+4} N(RS_I(i; c, s))$$

$$CM_C(i) = \sum_{c=2}^4 \sum_{s=c+3}^{c+4} \{N(RS_{RG}(i; c, s)) + N(RS_{BY}(i; c, s))\}$$

$$CM_O(i) = \sum_{k=0}^{n_\phi-1} N \left( \sum_{c=2}^4 \sum_{s=c+3}^{c+4} N(RS_O(i; k\pi/n_\phi; c, s)) \right)$$

$$CM_F(i) = \sum_{c=2}^4 \sum_{s=c+3}^{c+4} N(RS_F(i; c, s))$$

$$CM_M(i) = \sum_{k=x, y} N \left( \sum_{c=2}^4 \sum_{s=c+3}^{c+4} N(RS_{M_k}(i; c, s)) \right)$$

The conspicuity maps are normalized again, and summed up into a fundamental saliency map  $S(i)$ . The sequence of fundamental saliency maps  $S(i)$  comprises a fundamental saliency video.

$$S(i) = \sum_j \frac{w_j(i)}{\sum_j w_j(i)} N(CM_j(i)).$$

## 4. Instantaneous saliency depletion

### 4.1 Overview

Instantaneous saliency depletion along with gradual recovery simulates ‘‘Inhibition of return’’ effect. Due to this effect,

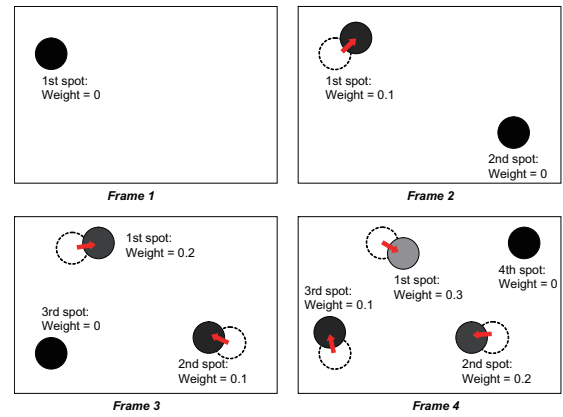


Figure 3 Example of ISD mask

after attention is diverted away from the previously focusing location, humans tend to fail to notice interest events at the peripheral region of the original focusing location. There is then a delay (approximately 500-900 ms) before attention returns to the original location of interest. Instantaneous saliency depletion has been implemented in the previously reported algorithm only 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 MSRs of the saliency video. Simultaneously, the recovery mask gradually retains saliency in the saliency video. The two masks are combined to form an overall ISD mask  $ID(i)$ .

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

### 4.2 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 in the MSR of saliency video frame  $S(i - 1)$  are set at 0.

Every MSR moves in accordance with the average optical flow values corresponding to the previous position of the MSR as shown in Fig.3. The position of a MSR extracted from the  $k$ -th frame when it reaches the  $i$ -th frame,  $MSR(i; k)$  ( $i \geq k$ ), is obtained as follows:

$$MSR(i; k) = \{(x + \Delta M_x(i; k), y + \Delta M_y(i; k)) : (x, y) \in MSR(i - 1; k)\},$$

$$\Delta M_x(i; k) = (\pi r^2)^{-1} \sum_{(x', y') \in MSR(i-1; k)} M_x(i-1)_{(x', y')},$$

$$\Delta M_y(i; k) = (\pi r^2)^{-1} \sum_{(x', y') \in MSR(i-1; k)} M_y(i-1)_{(x', y')},$$

where  $M_x(i)$  and  $M_y(i)$  are horizontal and vertical optical flows of the  $i$ -th frame,  $\Delta M_x(i; k)$  and  $\Delta M_y(i; k)$  are averaged optical flow values that correspond to the depletion circle  $MSR(i-1; k)$ , and  $MSR(i; i) = MSR(i)$ . Every optical flow image is calculated by the Lukas-Kanade method [12].

The above procedures to generate the MSR depletion masks are summarized as follows:

$$ID_1(i)_{(x,y)} = \begin{cases} 0 & \text{if } (x, y) \in MSR(i-1) \\ ID(i-1)_{(x', y')} & \text{else if } (x, y) \in MSR(i; k) \\ 1 & \text{else if } (x, y) \in MSR(i-1; k) \\ ID(i-1)_{(x,y)} & \text{otherwise} \end{cases}$$

$$(x', y') = (x - \Delta M_x(i; k), y - \Delta M_y(i; k))$$

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 MSRs to regain their value. The MSRs eventually reach the original mask value of 1 after  $1/\alpha = 10$  frames (= 667 ms when the frame rate of the fundamental saliency video is 15 frame/sec).

$$ID_2(i)_{(x,y)} = \alpha \cdot \forall i.$$

Fig.3 shows an example of ISD masks created by the above procedures. At frame 1, the MSR of the previous frame is depleted with the MSR depletion mask. At 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 for the subsequent frames.

## 5. Gradual saliency depletion

### 5.1 Overview

According to the “*Neural Adaptation*” theorem, in situations where there are no surprising events, e.g. unchanging backgrounds or constant flickering, the human visual system exhibits a gradual decrease in sensitivity to saliency over time. The sensitivity is recovered or retained only in cases where surprising events occur, such as an increase in the velocity of an object, or a change of scenery.

Gradual saliency depletion with instantaneous recovery is developed by creating two masks for each frame of the saliency video: the *whole-region depletion mask*  $GD_1(i)$  and the *surprise mask*  $GD_2(i)$ . The whole-region depletion mask creates the gradual depletion of saliency in the

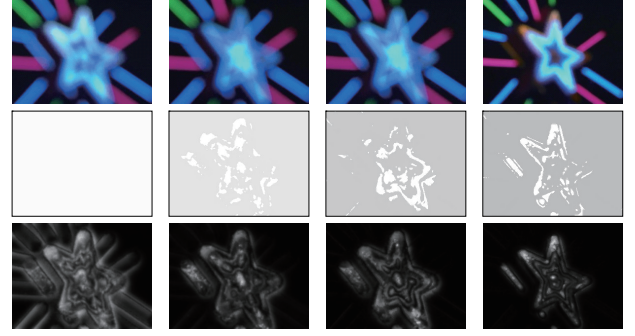


Figure 4 Example GSD masks. From the left, frames 0, 29, 67 and 89. Top row: original frames, mid row: GSD masks, bottom row: saliency video multiplied by GSD masks.

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)  $GD(i)$ .

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

### 5.2 Mask construction

The whole-region depletion mask  $GD_1(i)$  is a gray-scale image with all the pixel values initially at 1. At each frame a constant value  $\beta = 0.0025$  is subtracted from each pixel in the whole-region depletion mask unless a scene change does not occur at the frame. If a scene change is detected at the frame, the whole-region depletion mask regains values.

$$GD_1(i)_{(x,y)} = \max\{0, GD_1(i-1)_{(x,y)} - \beta\}.$$

The surprise mask  $GD_2(i)$  indicates where surprising events occur. Previous work [6] used a probabilistic approach to compute surprise. However, that approach needs 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 Gaussian pyramids used for extracting feature 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:

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

$$G_{(c,s)}(k) = G_{c/2}(k) - G_{s/2}(k),$$

where  $j$  corresponds to a feature type, and  $G_\sigma$  is a Gaussian distribution with mean 0 and variance  $\sigma^2$ . The temporal response maps are summed up and binarized with the threshold  $\theta (= 0.25 \max_{(x,y)} SP(i)_{(x,y)})$  to form the surprise mask  $GD_2(i)$ , as follows:

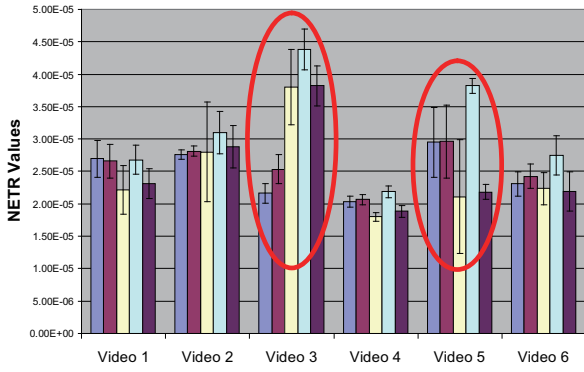


Figure 5 Experimental results: Average NETR value for each test video. Bars from the left to the right for each video: Still image algorithm, moving algorithm, our algorithm case 1 (ISD mask), case 2 (GSD mask) and case 3 (ISD+GSD mask)

$$GD_2(i)_{(x,y)} = \begin{cases} 1 & SP(i)_{(x,y)} \geq \theta \\ 0 & \text{otherwise} \end{cases}$$

Fig.4 shows example GSD masks created by the above-mentioned procedures. The GSD masks become darker owing to the whole-region depletion masks, while the surprise masks retain saliency values in surprising regions.

## 6. Experiments

### 6.1 Conditions

To evaluate the performance of our algorithm relative to the previous algorithms [3], [5], we chose five subjects to view six different sample video clips. The eye movement of each subject was tracked using an eye tracking device and the subjects viewed each video twice. We gave the subjects few instructions when they are viewing the sample videos. This implies that the subjects viewed the sample videos in “passive viewing” situations, where the “neural adaptation” phenomenon usually occurs. The eye tracking data was compared with the saliency videos produced by each algorithm to see 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.

### 6.2 Comparing eye tracking data with saliency videos

The eye movement may contain some amount of noises because of small eye movements of fixations [10], [13]. Thus, instead of the eye tracking data itself, we use the *eye tracking region ETR*  $ETR(i; k, n)$  defined as follows:

$$ETR(i; k, n)$$

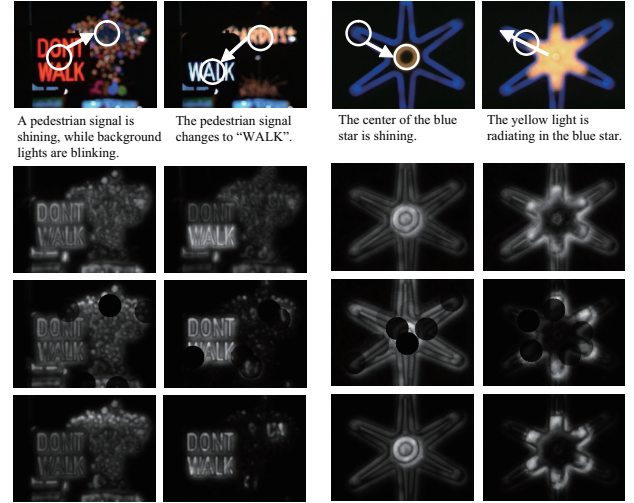


Figure 6 Saliency videos produced from Video 3 (left half) and Video 5 (right half). Top row: original video with circles and arrows that shows typical focusing points and eye movements, respectively, second row: moving algorithm, third row: our algorithm case 1, bottom row: our algorithm case 2. (Videos will be seen in the following URL: <http://www.brl.ntt.co.jp/people/akisato/research.html>)

$$= \{(x, y) : (x - x(i; k, n))^2 + (y - y(i; k, n))^2 \leq \delta^2\},$$

where  $i$ ,  $k$  and  $n$  correspond to a frame number, a video, and a subject, respectively, and  $(x(i; k, n), y(i; k, n))$  is the position on which subject  $n$ 's eye focused when viewing frame  $i$  of sample video  $k$ .

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*  $V(i; k, n)$ ).

$$V(i; k, n) = \frac{\sum_{(x', y') \in ETR(i; k, n)} S_D(i)_{(x', y')}}{\sum_{(x', y')} S_D(i)_{(x', y')}} \quad (1)$$

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.5 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 compared with the previous algorithms, while instantaneous 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 with the previous algorithms. Fig.6

shows saliency videos in such cases.

In all cases, our algorithm outperformed the previous algorithms for Video 3. Fig.6 (left half) shows that all lighting regions in the sample video are enhanced when the previous algorithm was used (see the second row), while for all the cases 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.6 (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.6.

## 7. 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.

The discussion in Section 6.2 implies that instantaneous depletion can have only limited advantages in “passive viewing” situations. Note that rapid eye movements derived from the “visual search” effect could not be observed in the eye tracking test. However, instantaneous depletion is expected to work well under “active viewing” situations, where the “visual search” phenomenon and the subsequent “Inhibition of Return” effect frequently occur. The advantages of instantaneous depletion and the disadvantages of gradual depletion in such situations will be investigated and clarified in future work.

## Acknowledgements

The authors would like 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. The authors also thank Dr. Takehiko Ohno of NTT for providing an eye tracking device [14]. Lastly, the authors thank Dr. Yoshinobu Tonomura, Dr. Hiromi Nakaiwa and Dr. Shoji Makino for their help.

## Bibliography

- [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 circuitry,” *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, N. Dhavale, and F. Pighin, “Realistic avatar eye and head animation using a neurobiological model of visual attention,” *Proc. SPIE International Symposium on Optical Science and Technology*, Vol.5200, pp.64–78, August 2003.
- [6] L. Itti and P. Baldi, “A principled approach to detecting surprising events in video,” *Proc. Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.631–637, June 2005.
- [7] M.I. Posner and Y. Cohen, “Components of visual orienting,” in *Attention and Performance*, ed. H. Bouma and D.G. Bouwhuis, pp.531–556, Erlbaum, 1984.
- [8] R.M. Klein, “Inhibition of return,” *Trends in Cognitive Sciences*, Vol.4, No.4, pp.138–147, April 2000.
- [9] 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.
- [10] S. Martinez-Conde, S.L. Macknik, and D.H. Hubel, “The role of fixational eye movements in visual perception,” *Nature Reviews*, Vol.5, pp.229–240, March 2004.
- [11] V. Navalpakkam and L. Itti, “Optimal cue selection strategy,” *Advances in Neural Information Processing Systems (NIPS)*, pp.987–994, December 2005.
- [12] B. Lucas and T. Kanade, “An iterative image registration technique with an application to stereo vision,” *Proc. of International Joint Conference on Artificial Intelligence*, pp.674–679, 1982.
- [13] R.W. Ditchburn and B.L. Ginsborg, “Vision with a stabilized retinal image,” *Nature*, Vol.170, pp.36–37, July 1952.
- [14] T. Ohno, N. Mukawa, and A. Yoshikawa, “FreeGaze: A gaze tracking system for everyday gaze interaction,” *Proc. Symposium on ETRA*, pp.125–132, 2002.