Proceedings of the 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems
pp. 5198-5203, October 2009

Stability and Sensitivity of Bottom-Up Visual Attention for Dynamic Scene Analysis

Yukie Nagai

***Abstract*— This paper presents an architecture extending bottom-up visual attention for dynamic scene analysis. In dynamic scenes, particularly when learning actions from demonstrations, robots have to stably focus on the relevant movement by disregarding surrounding noises, but still maintain sensitivity to a new relevant movement, which might occur in the surroundings. In order to meet the contradictory requirements of stability and sensitivity for attention, this paper introduces biologically-inspired mechanisms for retinal filtering and stochastic attention selection. The former reduces the complexity of peripheral signals by filtering an input image. It results in enhancing bottom-up saliency in the fovea as well as in detecting only prominent signals from the periphery. The latter allows robots to shift attention to a less but still salient location in the periphery, which is likely relevant to the demonstrated action. Integrating these mechanisms with computation for bottom-up saliency enables robots to extract important action sequences from task demonstrations. Experiments with a simulated and a natural scene show better performance of the proposed model than comparative models.**

I. Introduction

Robots mostly face a dynamic scene. When interacting with humans and learning tasks from them, robots have to appropriately gaze at the partners' movement so as to be better accepted as social agents as well as to succeed in task learning. The issue concerning the spatial aspect of visual attention is described as "where to attend," that is, robots have to select a location to attend to depending on the context. When a partner is demonstrating a task, robots' attention should be directed to his body movement so that they can learn the means of the task. If an object is involved in the task, robots have to pay attention also to the object so as to learn the goal of the task. The issue of "where to attend" is the central question in attentional control. In dynamic scenes, furthermore, "when to attend" becomes crucial (i.e., the temporal aspect of attention). When a task starts being demonstrated, robots must quickly respond to the partner's movement and stably follow it without being interrupted by a distraction. At the same time, robots have to also maintain sensitivity to a new relevant movement, which might occur in the surroundings. For example, if a demonstrator, who is handling an object with his right hand, starts moving his left hand, robots' attention must quickly be shifted to the new movement for learning the coordination of the actions. Thus, the issue of "when to attend" requires fulfilling the contradictory requirements: stability of attention to a certain target and its sensitivity to a new target.

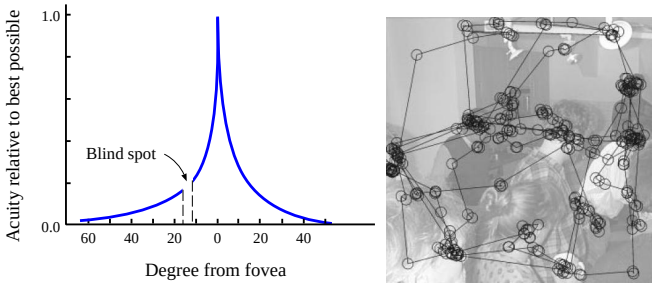
Y. Nagai is with the Research Institute for Cognition and Robotics, Bielefeld University, 33594 Bielefeld, Germany yukie@techfak.uni-bielefeld.de

Regarding the issue of "where to attend," architectures of bottom-up attention have widely been investigated. It has been shown that a model based on visual saliency (e.g., [1], [2]) is able to detect likely important locations in a scene. In the context of human-robot interaction, the model enables robots to detect human partners as well as salient objects [3]–[5]. Since a human face and hands have conspicuous features in terms of color, edge, and motion, they can be distinguished from the surroundings. Bottom-up attention might also extract robots' body parts from a scene [6], [7]. To associate the extracted features with the controllability of them allows robots to acquire their *body image*. Combining several modalities enriches bottom-up attention. Integration of auditory saliency with visual saliency facilitates the detection of relevant targets [8]. In certain situations, a top-down modulation can be added to a bottom-up system [9], [10]. If contextual knowledge is given, robots can utilize it to determine "where to attend."

Although there have been many studies employing bottom-up attention, the issue of "when to attend" has hardly been investigated. In the above studies, only the locations to attend to (i.e., "where") and the frequency of attending the locations have been evaluated. It has been reported that purely bottom-up architectures have a weakness in the temporal aspect of attention [11]. Bottom-up attention can easily be distracted by a disturbance [4]. In order to interact with humans and learn actions from them, robots have to overcome the "when" issue as well as the "where" by addressing the stability and sensitivity of attention.

This paper proposes an architecture extending bottom-up attention for dynamic scene analysis. To meet the contradictory requirements of stability and sensitivity, biologically-inspired mechanisms of retinal filtering and stochastic attention selection are introduced. The former contributes to stabilizing attention. Imitating human vision, the mechanism produces a retinal image, in which the fovea has high acuity while the peripheral area has low. It results in enhancing bottom-up saliency in the fovea as well as in extracting only highly-conspicuous signals from the periphery. The latter achieves sensitivity to new signals after the stabilization of attention. This mechanism allows robots to shift attention to a less but still conspicuous location in the periphery, which appears to be important in the context. Integrating these two mechanisms with computation for bottom-up saliency enables robots to detect meaningful action sequences from human task demonstrations.

The rest of this paper is organized as follows: Section II provides evidences about human vision and attention. It



(a) Visual acuity (adapted from [12])

(b) Typical scan paths for static image (from [13])

Fig. 1. Human vision and attention

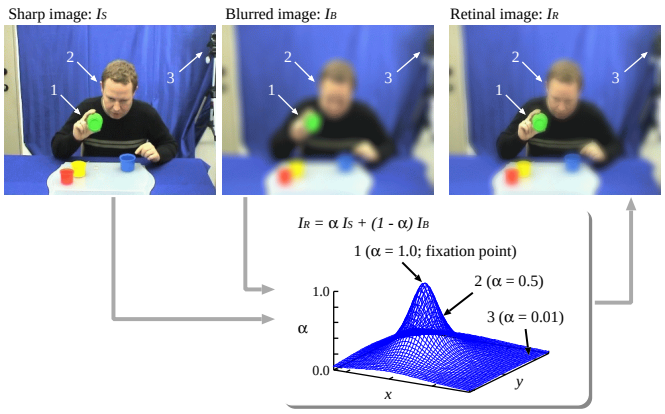


Fig. 2. Retinal filtering for enhancing focus on currently-attended location

describes biological mechanisms inspiring retinal filtering and stochastic process for attention selection. Section III explains a model integrating bottom-up attention with the above mechanisms. The validity of the model is examined using a simulated and a real scene in Sections IV and V. Section VI gives conclusion and discussions.

II. Human Vision and Attention

A. Visual Acuity

Human vision has different acuity depending on the retinal area (see Fig. 1 (a)) [12]. In the fovea, the acuity is 100 [%] and thus humans perceive a sharp image, whereas it rapidly drops to 10 [%] or less in the periphery. This is caused by differences in the density of photoreceptor cells and in their mechanism to process signals.

Cone cells, which are centralized in the fovea, allow high acuity for the foveal vision. The density of cone cells looks like the visual acuity shown in Fig. 1 (a): It is highest at the center of the fovea while rapidly decreasing in the periphery. Rod cells, by contrast, are decentralized in the peripheral region. They do not exist in the fovea, increases as cone cells are decreasing, and then gradually descends to the edge of the periphery. The low density of rod cells causes low acuity for the peripheral vision. In addition, the mechanism for rod cells to process signals promotes less acuity for the periphery. Signals perceived by multiple rod cells converge on a single inter-neuron in order to be amplified unlike those for cone cells. This mechanism, on the one hand, causes the loss of detailed information in the peripheral vision, but on the other hand, allows humans to focus more on the signals perceived in the fovea.

B. Stochastic Attention

Humans determine where to attend using the retinal image. The process for selecting the attentional point is stochastic rather than deterministic [13]. Even when people observe the same scene, their scan paths differ between trials. Fig. 1 (b) shows a typical path recorded when a person was looking at a static image. His fixation accumulated on salient targets like human face, hands, and convex objects. Such strongly-relevant targets are mostly selected as attentional locations over trials; however, the order for the fixation might change between trials. For example, a subject might first look at a human face and next hands in one trial, but vice versa in another trial. It is supposed that such a stochastic process prevents anchoring our attention to one salient location as well as to minimize the typical time needed to process a visual scene.

III. Bottom-Up Visual Attention for Dynamic Scene Analysis

Inspired by human vision, this paper introduces retinal filtering and stochastic attention selection. The former contributes to stabilizing attention to the current fixation point whereas the latter allows to maintain sensitivity to new signals observed in the periphery. Integrating these mechanisms with saliency computation enables robots to overcome the issue of "when to attend" as well as "where to attend."

A. Retinal Filtering

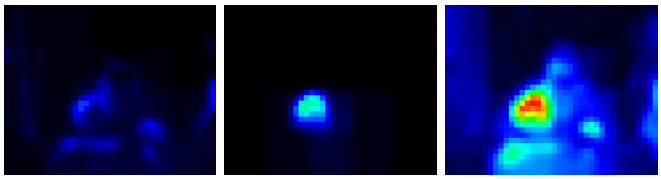
The proposed model first generates a retinal image like human vision. Fig. 2 illustrates the mechanism. The retinal image $\boldsymbol{I}_R$ is created by combining two different sharpness of input images: the sharp one $\boldsymbol{I}_S$, which is directly captured from a camera, and the blurred one $\boldsymbol{I}_B$, which is generated by globally smoothing $\boldsymbol{I}_S$ with a Gaussian filter. Let $\boldsymbol{x}_F(t-1)$ be the fixation point at time $t-1$. The image value $I_R(\boldsymbol{x}, t)$ at the location $\boldsymbol{x} = (x,\ y)$ is calculated by summing $I_S(\boldsymbol{x}, t)$ and $I_B(\boldsymbol{x}, t)$ using a weight with respect to the distance from $\boldsymbol{x}_F(t-1)$:

$$I_R(\boldsymbol{x}, t) = \alpha I_S(\boldsymbol{x}, t) + (1 - \alpha) I_B(\boldsymbol{x}, t) \tag{1}$$

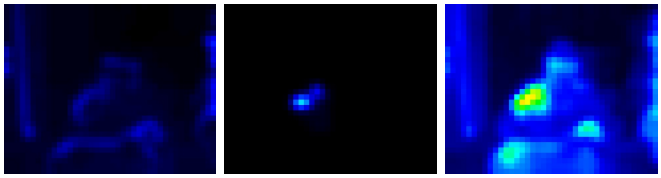
$$\text{where} \quad \alpha(\boldsymbol{x}, t) = \frac{D^2}{\|\boldsymbol{x} - \boldsymbol{x}_F(t-1)\|^2 + D^2}. \tag{2}$$

The weight $\alpha(\boldsymbol{x}, t)$ is a Cauchy distribution whose center is $\boldsymbol{x}_F(t-1)$, amplitude 1.0, and diameter D.

The resulting image is shown in the top right in Fig. 2. At location 1 (i.e., the fixation point), α equals 1.0 and thus $\boldsymbol{I}_R$ is as sharp as $\boldsymbol{I}_S$. The fingers of the person's right hand can clearly be recognized. From location 1 to 2 and then to 3, as α becomes smaller, $\boldsymbol{I}_R$ gets more blurred. At location 3, $\boldsymbol{I}_R$ is as blurred as $\boldsymbol{I}_B$.



(a) Saliency map derived from I_R



(b) Saliency map derived from I_S

Fig. 3. Saliency map with (a) and without (b) retinal filtering. From left to right, orientation map, motion map, and final saliency map are presented.

B. Saliency Computation

The effect of the retinal filtering can be observed in visual saliency. The model next computes the saliency for $\boldsymbol{I}_R$ employing the model proposed by Itti et al. [1], [2]. Saliency is calculated as the difference between a focused region and the surroundings. For example, a red cup can be detected as salient against blue background in terms of color. The model proposed here computes saliency with respect to five features: color, intensity, orientation, flicker (i.e., temporal change in the intensity), and motion (i.e., optical flow). The first three are responsible for static features whereas the last two are for dynamic. The generated conspicuity maps are then normalized within each feature, and are finally integrated into the saliency map with equal weights. Refer to [1], [2] for a more detailed explanation.

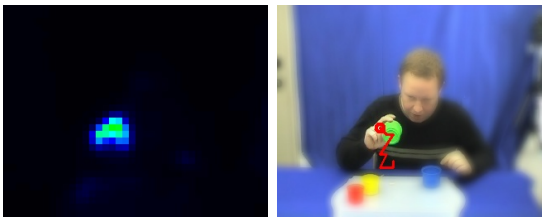
Fig. 3 shows the saliency map derived from $\boldsymbol{I}_R$ (a) and from $\boldsymbol{I}_S$ (b), which are corresponding to the scene shown in Fig. 2. From left to right, the orientation map, the motion map, and the final saliency map are presented. Comparing Fig. 3 (a) to (b), we can see that the retinal filtering has the effect of enhancing the saliency in the fovea (i.e., the region corresponding to the person's right hand) while suppressing it in the periphery, which leads to the stability of attention.

C. Stochastic Attention Selection

The model then selects an image location to attend to based on the saliency. In order for robots to quickly respond to a new conspicuous target, a stochastic algorithm used in [14], [15] is adopted. The model calculates $\phi(\boldsymbol{x}, t)$, which defines the transition probability for attention from the current fixation point $\boldsymbol{x}_F(t-1)$ to $\boldsymbol{x}$ at t:

$$\phi\left(\boldsymbol{x}, t\right) = \frac{\exp\left(-\beta\left(s\left(\boldsymbol{x}_F(t-1), t\right) - s(\boldsymbol{x}, t)\right)\right)}{\sum_{\boldsymbol{x}'} \exp\left(-\beta\left(s\left(\boldsymbol{x}_F(t-1), t\right) - s(\boldsymbol{x}', t)\right)\right)}, \quad (3)$$

where $s(\boldsymbol{x}, t)$ is the saliency for $\boldsymbol{x}$, and β defines the amplitude to enhance the difference in the saliency. Fig. 4 (a) shows the probability map derived from Fig. 3 (a). Only the locations with high saliency present high probability.



(a) Transition probability map (b) Transition of attention indicated by red line

Fig. 4. Stochastic attention selection

Note that this computation takes only the saliency value into account, but not the distance from the fovea. That is, if there are locations which are as salient as the fovea, they are given with the same probability.

The model then determines the next fixation point $\boldsymbol{x}_F(t)$ using a Metropolis algorithm. It selects a candidate location $\boldsymbol{x}_{F'}$ based on $\phi(\boldsymbol{x}, t)$, and accepts it if

$$\Delta s(\boldsymbol{x}_{F'}, t) = s(\boldsymbol{x}_{F'}, t) - s\left(\boldsymbol{x}_F(t-1), t\right) > 0. \quad (4)$$

Otherwise $\boldsymbol{x}_{F'}$ is accepted with a probability:

$$p(\boldsymbol{x}_{F'}, t) = \exp\left(\Delta s(\boldsymbol{x}_{F'}, t)/T\right), \quad (5)$$

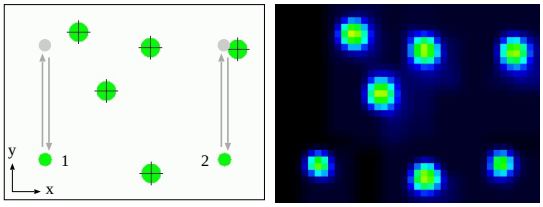
where T defines the randomness of the stochastic process. The higher T is, the more a location with less saliency is accepted. This process is repeated until a newly selected location satisfies the condition.

The transition of the attentional location is represented by a red line in Fig. 4 (b). Since the person's right hand was strongly salient due to the movement, it has continuously attracted the model's attention. The stochastic algorithm, however, allows robots to sensitively respond to a new conspicuous target, for example, the person's left hand pointing the blue cup.

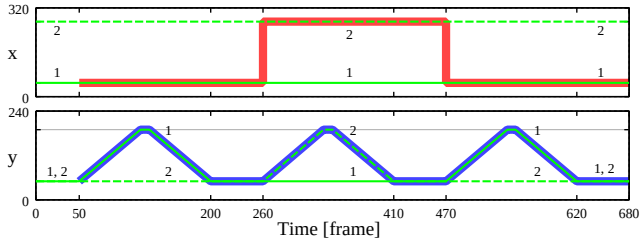
IV. Experiment I: Simulated Dynamic Scene

A. Setting

The first experiment evaluated the proposed model with a simulated scene. Fig. 5 (a) shows a snapshot of the video input (left) and the corresponding saliency map (right), which was computed from the sharp image. There are two dynamic targets, which are smaller green circles indicated by 1 and 2, among five static distractors, which are larger circles with a cross mark on it. When no target is moving, one of the distractors is the most salient due to the larger colored area as well as the cross feature. In Fig. 5 (a), where no target is moving, the upper left-hand distractor shows the highest saliency. However, once a target starts moving, it becomes more salient than the distractors due to the movement. The trajectories of the targets are plotted in Fig. 5 (b). The solid and the dashed green lines indicate the positions of Targets-1 and -2, respectively. First Target-1 moves up and down, next Target-2, and then again Target-1 while keeping their x positions in the image. There are intervals between the movements, during which neither target nor distractor moves.



(a) Input video (left) and saliency map (right) without retinal filtering. Two targets move along gray lines.



(b) Transition of two targets (Target-1: solid line, -2: dashed line). Thicker red and blue lines indicate the desired transition of attention.

Fig. 5. Test video and transition of two targets

The red and the blue lines in Fig. 5 (b) indicate the desired transition of attention. It shows that attention should follow the movement while one of the targets is moving, and stay at the same target during the following interval. For example, Target-1 attracts attention from 50 to 200 [frame] because of the movement and from 200 to 260 [frame] too despite no movement. Similarly, Target-2 draws continuous attention from 260 to 470 [frame] until Target-1 restarts moving. These movements simulate a scene where a human demonstrator moves his left and right hands sequentially. While presenting a task, the demonstrator might involuntary take pauses between movements in order to change the motion direction and/or to ease the action segmentation. During such short intervals, robots' attention should be kept to the same target (e.g., one of the demonstrator's hands) instead of being shifted to others (e.g., his face or the other hand) because the demonstrated action likely continues with the same target. At the same time, robots have to also quickly respond to a new target (e.g., movement produced by the other hand), which might occur in the surroundings. This experiment evaluates how stably the proposed model gazes at a moving and a pre-moving target, and how sensitively it responds to a new target.

B. Comparative models

Performance of the proposed model was compared with three other models. The proposed model is hereafter called

- *STC-R model*, which indicates the *ST*o*C*hastic algorithm with the *R*etinal filtering.

The three comparative models are:

- *WTA model*, which employs the *W*inner-*T*ake-*A*ll algorithm for the attention selection, where the saliency is computed from the sharp input image,
- *WTA-R model*, which additionally adopts the *R*etinal filtering with WTA model, and
- *STC-C model*, which employs the same *ST*o*C*hastic algorithm as STC-R model without the retinal filtering, but instead filters the transition probability with a *C*auchy distribution.

The winner-take-all is an algorithm to select the most salient location as the attentional point. It has been adopted in most of the current studies and showed reasonable performance in static environments (e.g., [1], [8]) . In dynamic scenes, however, it suffers from a drawback of instability of attention as well as of strong anchoring to the most salient location.[1] Comparing WTA model with WTA-R allows to investigate the effect of the retinal filtering on stabilizing attention.

STC-C model is an architecture inspired by a behavioral evidence about human attention [14], [15]. As described in Section II, human attention is stochastic. Moreover, they shift attention to a closer position much more frequently than to a far position when scanning an image [13]. STC-C model imitates the behavior by filtering the transition probability $\phi(\boldsymbol{x}, t)$ with a Cauchy distribution:

$$\hat{\phi}(\boldsymbol{x}, t) = \frac{D_C^2}{\|\boldsymbol{x} - \boldsymbol{x}_F(t-1)\|^2 + D_C^2}\phi(\boldsymbol{x}, t), \tag{6}$$

which serves as a new probability for the attention selection. That is, the attention of STC-C model is simply narrowed by D_C regardless of whether there is any movement in the peripheral vision. Note that the saliency for STC-C model is computed from the sharp input image.

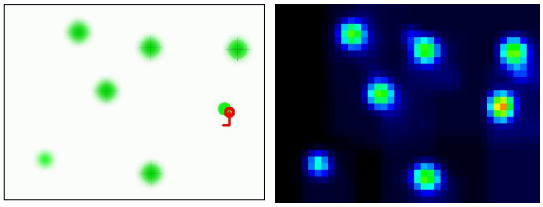
C. Results

Fig. 6 shows the result for STC-R model, whose parameters were set as $D = 40.0$, $\beta = 0.02$, and $T = 5.0$; (a) is the retinal image with the attentional point (left) and the saliency map (right) captured at 290 [frame], and (b) is the attentional transition of the model. In the scene, Target-2 moving upward was attended to by the model because of the higher saliency produced by the movement. Target-1 as well as the distractors, by contrast, had less saliency than in Fig. 5 (a) due to the blurred image. Comparing Fig. 6 (b) with Fig. 5 (b) demonstrates that STC-R model well reproduced the desired transition of attention. The model stably gazed at a moving target and continuously fixated on it while quickly responding to a new target.

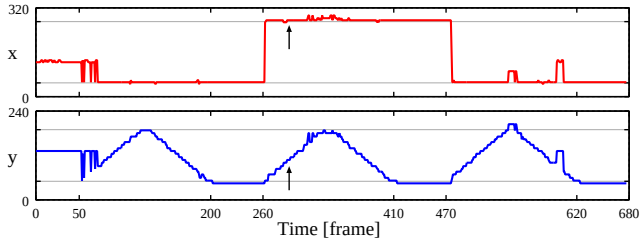
Fig. 7 shows the results for the comparative models: (a) WTA model, (b) WTA-R model, and (c) to (e) STC-C model with different D_C. The smaller D_C is, the narrower the transition probability is. $D_C = 40.0$ is comparable to D for the retinal filtering in terms of the size for the filter. The other parameters for STC-C model (i.e., β and T) were the same with those for STC-R model.

First, Fig. 7 (a) shows that WTA model could not keep fixating a target during the motion intervals. It always shifted the attention to the most salient distractor when no target was moving. In action learning, it causes a problem that the perception of demonstrated actions is often interrupted.

[1]In order to avoid the anchoring, attentional models have to inhibit the saliency for attended objects, but not for locations, for which a mechanism is still an open challenge to develop.



(a) Retinal image (left) and saliency map (right) captured at 290 [frame]. The model attends to Target-2 moving upward.



(b) Transition of attention, where arrows denote 290 [frame]

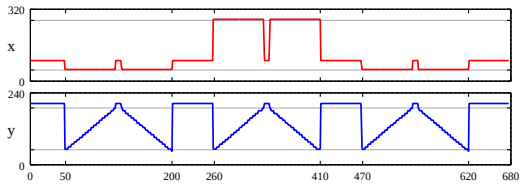
Fig. 6. Result for STC-R model with simulated scene

For example, the movement of lifting up an object and then putting it down would independently be extracted, which makes it difficult to learn the coordination between the movements. Fig. 7 (b), by contrast, shows the effect of the retinal filtering on preventing such interruptions. From 340 to 680 [frame], WTA-R model continuously gazed at the moving and the pre-moving target unlike WTA model. However, there is a drawback produced by the retinal filtering: From 140 to 320 [frame], the model could not rapidly respond to a new movement or even shift the attention back to the pre-attended movement. The attention distracted by the upper right-hand distractor was anchored there until Target-2 approached it. It shows that the attention of WTA-R model was too strongly stabilized by the retinal filtering. Note that such uncertain attentional shift can happen even with the deterministic algorithm because the movement of the attentional point (i.e., the center for the retinal filtering) causes relative change in the saliency.

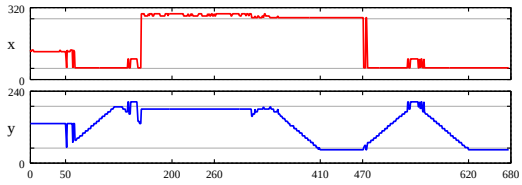
Second, the results shown in Fig. 7 (c) to (e) demonstrate less sensitivity of STC-C model with a smaller D_C and instability of it with a larger D_C. Adjusting D_C did not help the model meet the contradictory requirements of stability and sensitivity for attention. The reason is that the model does not take into account the difference in the source for the saliency (i.e., static or dynamic) but simply decreases the transition probability to distant locations. The better performance of STC-R model over STC-C model demonstrates the advantage of combining the stochastic algorithm with the retinal filtering.

V. Experiment II: Natural Scene

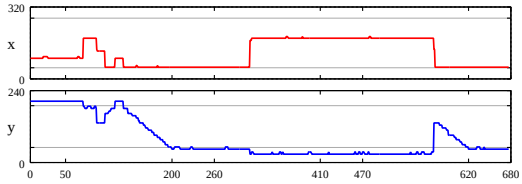
The four models (i.e., the proposed model and the three comparative ones) were applied to a pre-recorded video, in which a person is demonstrating a cup-stacking task to his infant [16]. Fig. 8 shows the results: (a) STC-R model, (b) STC-C model ($D_C = 16.0$), (c) WTA-R model, and (d) WTA



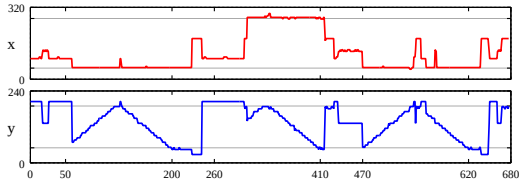
(a) WTA model



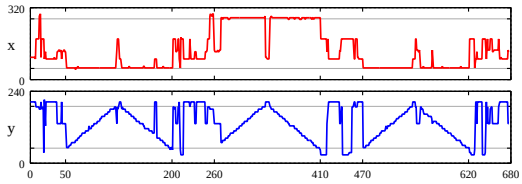
(b) WTA-R model



(c) STC-C model ($D_C = 8.0$)



(d) STC-C model ($D_C = 16.0$)



(e) STC-C model ($D_C = 40.0$)

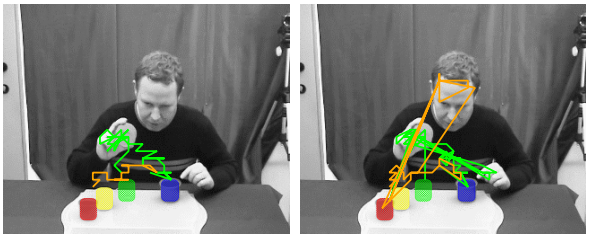
Fig. 7. Transition of attention for comparative models

model. The parameters for STC-R/C model were the same as in Experiment I. The green and the orange lines indicate the attentional transition of the models recorded when the person was moving the green cup into the blue one and when he took a pause after it, respectively.[2] He lifted up the green cup to closely present it to his infant as just shown in the picture, and then put it down into the blue one. After it, he took a short pause by putting his right hand behind the yellow cup in order to examine the attention of his infant.

First, comparing Fig. 8 (c) with (d), we can see the effect of stabilized attention by the retinal filtering. WTA-R model stably followed the movement of the person's right hand while he was handling the cup and then taking a pause. The attention was not distracted, for example, by other cups or the person's face. Similarly, comparing Fig. 8 (a) with

[2]The rough transition of attention was caused by the low resolution of the saliency map.

(a) STC-R model (b) STC-C model ($D_C = 16.0$)



(c) WTA-R model (d) WTA model

Fig. 8. Transition of attentional location in natural scene. The green and the orange lines indicate the scan paths of the models when the person was handling the green cup and then taking a pause, respectively.

(b) demonstrates the stabilization of attention by the retinal filtering. Especially when the person took a pause (the orange line), STC-R model kept attending to his right hand whereas STC-C model did not, which indicates that STC-R model can easily detect the following action. Next, the comparison of Fig. 8 (a) with (c) allows us to examine the sensitivity of attention achieved by the stochastic algorithm. The person's left hand pointing the blue cup was detected by STC-R model (and also by STC-C and WTA models) but not by WTA-R model. His pointing was indicating the goal position of the green cup and therefore is relevant to the task. STC-R model could direct attention to the pointing gesture because its movement was large enough to be extracted after the retinal filtering. These results suggest that combining the stochastic algorithm with the retinal filtering enables robots to meet both stability and sensitivity for attention required in dynamic scene analysis.

VI. Conclusion and Discussion

This paper presented an architecture extending bottom-up attention for dynamic scene analysis. When observing actions, robots have to stably extract the relevant movement by disregarding noises but still maintain sensitivity to a new movement in the surroundings. The proposed model fulfilled the contradictory requirements by retinal filtering and stochastic attention selection inspired by human vision.

It is notable that using only dynamic features for saliency computation, instead of applying retinal filtering and stochastic attention selection, cannot cope with the problem. If robots determine where to attend based only on the motion signals, they would gaze at the demonstrator's arms as often as his hands because the arms are also involved in the action. However, it is known that when watching an action, people focus more on the hand than on the arm [17]. They extract the trajectory of the hand and then reproduce it by calculating the arm posture using their motor primitives. The model proposed here imitates human attention: It extracts the movement of the demonstrator's hand by taking into account the static features as well as the dynamic. Edge features extracted from the fingers contributes to higher saliency for the hand than for the arm, which might explain why humans look at the hands. The model has also been evaluated with respect to the capability to extract key actions from task demonstrations [18]. Next steps are to apply the model to various task learning and to test it on an embodied robot, which produces dynamics by itself.

Acknowledgment

The author gratefully acknowledges the financial support from Honda Research Institute Europe.

References

[1] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[2] L. Itti, N. Dhavale, and F. Pighin, "Realistic avatar eye and head animation using a neurobiological model of visual attention," in *Proc. of the SPIE 48th Annual Intl. Symp. on Optical Science and Technology*, vol. 5200, 2003.

[3] L. Aryananda, "Attending to learn and learning to attend for a social robot," in *Proc. of the 2006 6th IEEE-RAS Intl. Conf. on Humanoid Robots*, 2006, pp. 618–623.

[4] C. Muhl and Y. Nagai, "Does disturbance discourage people from communicating with a robot?" in *Proc. of the 16th IEEE Intl. Symp. on Robot and Human Interactive Communication*, 2007, pp. 1137–1142.

[5] N. J. Butko, L. Zhang, G. W. Cottrell, and J. R. Movellan, "Visual saliency model for robot cameras," in *Proc. of the 2008 IEEE Intl. Conf. on Robotics and Automation*, 2008, pp. 2398–2403.

[6] C. C. Kemp and A. Edsinger, "What can i control?: The development of visual categories for a robot's body and the world that it influences," in *Proc. of the 5th Intl. Conf. on Development and Learning*, 2006.

[7] M. Hikita, S. Fuke, M. Ogino, T. Minato, and M. Asada, "Visual attention by saliency leads cross-modal body representation," in *Proc. of the 7th IEEE Intl. Conf. on Development and Learning*, 2008, pp. 157–162.

[8] J. Ruesch, M. Lopes, A. Bernardino, J. Hoernstein, J. Santos-Victor, and R. Pfeifer, "Multimodal saliency-based bottom-up attention: A framework for the humanoid robot icub," in *Proc. of the 2008 IEEE Intl. Conf. on Robotics and Automation*, 2008, pp. 962–967.

[9] C. Breazeal and B. Scassellati, "A context-dependent attention system for a social robot," in *Proc. of the 16th Intl. Joint Conf. on Artificial Intelligence*, 1999, pp. 1146–1153.

[10] J. Moren, A. Ude, A. Koene, and G. Cheng, "Biologically based top-down attention modulation for humanoid interactions," *Intl. Jnl. of Humanoid Robotics*, vol. 5, no. 1, pp. 3–24, 2008.

[11] F. Shic and B. Scassellati, "A behavioral analysis of computational models of visual attention," *Intl. Jnl. of Computer Vision*, vol. 73, no. 2, pp. 159–177, 2007.

[12] S. Coren, C. Porac, and L. M. Ward, *Sensation and perception*. Orlando: Academic Press, 1984.

[13] D. Brockmann and T. Geisel, "The ecology of gaze shifts," *Neurocomputing*, vol. 32-33, pp. 643–650, 2000.

[14] G. Boccignone and M. Ferraro, "Modelling gaze shift as a constrained random walk," *Physica A*, vol. 331, no. 1, pp. 207–218, 2004.

[15] H. Martinez, M. Lungarella, and R. Pfeifer, "Stochastic extension of the attention-selection system for the icub," University of Zurich, Tech. Rep., 2008.

[16] Y. Nagai and K. J. Rohlfing, "Computational analysis of motionese toward scaffolding robot action learning," *IEEE Trans. on Autonomous Mental Development*, vol. 1, no. 1, pp. 44–54, 2009.

[17] M. J. Mataric and M. Pomplun, "Fixation behavior in observation and imitation of human movement," *Cognitive Brain Research*, vol. 7, no. 2, pp. 191–202, 1998.

[18] Y. Nagai, "From bottom-up visual attention to robot action learning," in *Proc. of the 8th IEEE Intl. Conf. on Development and Learning*, 2009.