From Bottom-Up Visual Attention to Robot Action Learning

Yukie Nagai

***Abstract*—This research addresses the challenge of developing an action learning model employing bottom-up visual attention. Although bottom-up attention enables robots to autonomously explore the environment, learn to recognize objects, and interact with humans, the instability of their attention as well as the poor quality of the information detected at the attentional location has hindered the robots from processing dynamic movements. In order to learn actions, robots have to stably attend to the relevant movement by ignoring noises while maintaining sensitivity to a new important movement. To meet these contradictory requirements, I introduce mechanisms for retinal filtering and stochastic attention selection inspired by human vision. The former reduces the complexity of the peripheral vision and thus enables robots to focus more on the currently-attended location. The latter allows robots to flexibly shift their attention to a new prominent location, which must be relevant to the demonstrated action. The signals detected at the attentional location are then enriched based on the spatial and temporal continuity so that robots can learn to recognize objects, movements, and their associations. Experimental results show that the proposed system can extract key actions from human action demonstrations.**

***Index Terms*—learning from demonstration, bottom-up visual attention, what to imitate, key actions**

I. Introduction

LEARNING from demonstration and learning by imitation are widely accepted methodologies for robot action learning [1]. They investigate how robots can detect relevant information from action demonstrations and how they translate the information into their motor commands so as to reproduce the actions (see [2], [3] for a deeper discussion). The former issue is states as "what to imitate" whereas the latter is "how to imitate." This research focuses on "what to imitate," which mainly concerns robots' attention to extract task relevant features from visual input [4]. Although several approaches have been proposed to address the issue, they either predefined the candidate variables for robots to imitate (e.g., [5]–[9]) or employed top-down architectures to modulate robots' attention (e.g., [10], [11]).

Bottom-up visual attention (e.g., [12], [13]), on the other hand, has been increasingly exploited in the field of developmental robotics. It enables robots to autonomously explore the environment [14], [15], detect and interact with humans [16]–[18], and learn to recognize objects and their own bodies in their visual field [19], [20]. It has also contributed to uncovering the development of human attention [21]–[23]. Bottom-up attention is supposed to play an important role in infant vision. These great successes indicate the potential of bottom-up attention for coping with "what to imitate"; however, there has been no attempt to address the issue.

One reason is the instability of robots' attention. Bottom-up attention can be easily distracted by a noise or an irrelevant feature [17], which is often observed in dynamic scenes. Moreover, the dynamic properties of the relevant movement, which are important for learning actions, increase the difficulty considerably. In action learning, robots have to stably attend to the relevant movement by ignoring noises while maintaining the sensitivity to a new important movement. Compared to object learning, a big challenge for robots to learn actions is to satisfy these contradictory requirements.

Another difficulty in using bottom-up attention is the poor quality of the information detected at the attentional locations. For example, attentional points do not always correspond to a certain object: only a part of an object and sometimes even parts of several objects are included. Robots, therefore, have to enrich the information detected at the locations so as to learn to recognize objects involved in the demonstrated task. In addition, the transition of attentional points is not smooth or continuous. The attention suddenly jumps from one object to another, and often moves within an object. Thus, robots cannot simply trace the trajectory of their attention, but instead have to generatively learn to produce smooth movement.

In order to address the challenges, I introduce three key ideas: retinal filtering, stochastic attention selection, and continuity detection, inspired by biological and developmental evidence. Retinal filtering generates an image as in human vision: The acuity of an image is high in the fovea (i.e., the center of the image) whereas it gets drastically low in the peripheral vision. This mechanism reduces the complexity of the visual input, especially in the periphery, and thus enables robots to enhance their focus on the currently-attended location. A stochastic process for the attention selection then allows robots to flexibly shift their attention to an even less prominent location in the periphery. Such a salient location after the filtering must be relevant to the demonstrated task. Together with the retinal filtering, the stochastic attention fulfills the contradictory requirements, i.e., stability and sensitivity. Continuity detection subsequently enables robots to enrich the information detected at the attentional point. It leads to examining the relevance of the information to the task and associating it with respect to time and space. Primitive features like color and motion are used to calculate the continuity.

The following sections describe the proposed architecture as well as biological and developmental evidence: Section II gives an overview of the proposed system employing bottom-

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

978-1-4244-4118-1/09/$25.00

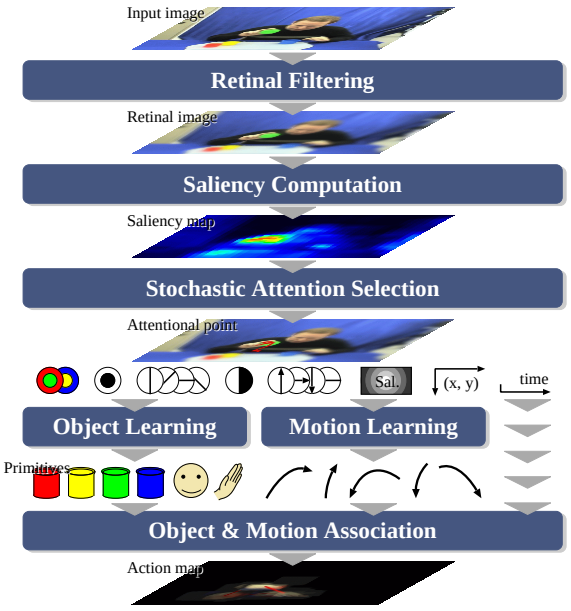


Fig. 1. A system architecture for action learning employing bottom-up visual attention

up visual attention. An attentional mechanism integrated with retinal filtering and stochastic attention selection is first explained in Section III. Then, a mechanism for learning objects and movements based on continuity is described in Section IV. Section V presents experimental results, where the system was applied to action learning from human demonstrations. How the system extracts key actions from task demonstrations is explained there. Finally, Section VI concludes the paper with discussion and future issues.

II. System Architecture for Bottom-Up Action Learning

Fig. 1 shows the system architecture for robot action learning employing bottom-up visual attention. The three key ideas inspired by human vision (i.e., retinal filtering, stochastic attention selection, and continuity detection) are applied in order for robots to enhance and enrich their attention.

The system consists of six modules: The first three are the retinal filtering, the saliency computation, and the stochastic attention selection, which are responsible for selecting the attentional location in robots' vision. They receive an image capturing a person demonstrating a task and determine where robots should attend. Here an architecture based on visual saliency is adopted as the basis for bottom-up attention. The following three modules are the object learning, the motion learning, and their association. They extract both objects and movements from the attentional location and then associate them by examining their temporal and spatial continuity. The system finally builds an action map, where key points in the demonstrated task are represented. For example, in a cup-stacking task, which was used in the present experiment, the action to grasp a cup, to lift it, and then to put it down into another cup can be extracted as key actions for accomplishing the task. It is suggested that the detection of such key actions

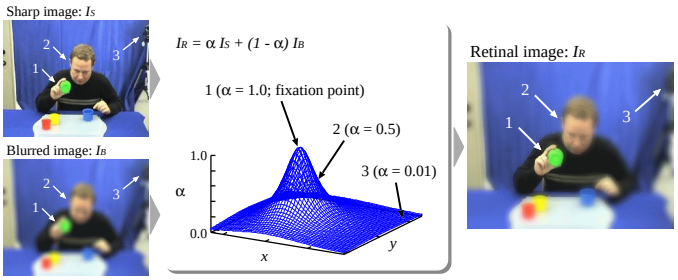


Fig. 2. Retinal filtering for enhancing focus of robots' attention on fovea

allows robots to learn to imitate the task not only at the motion level but also at the goal level. The following sections explain the six modules.

III. Attention Selection Based on Saliency

A. Retinal Filtering

In order to reduce the complexity of visual input and hence to stabilize bottom-up attention, a mechanism for retinal filtering inspired by human vision is introduced. Human vision has different acuity depending on the retinal area: in the fovea, humans see 100 % acuity of an image whereas it rapidly drops to 10 % or less in the peripheral vision [24]. This mechanism causes the loss of information in the peripheral area but allows humans to focus more on the signals perceived in the fovea.

The proposed system imitates this mechanism by filtering the input image (see Fig. 2). The retinal image $\boldsymbol{I}_R$ is created by combining two different sharpness of the input image: a sharp one $\boldsymbol{I}_S$, which is directly captured from a robot's camera, and a blurred one $\boldsymbol{I}_B$, which is generated by smoothing $\boldsymbol{I}_S$ with a Gaussian filter. Let $\boldsymbol{x}_F(t-1) = (x_F,\ y_F)$ be the fixation point in the image at time $t-1$. The image value $I_R(\boldsymbol{x}, t)$ at the location $\boldsymbol{x}$ is calculated by summing $I_S(\boldsymbol{x}, t)$ and $I_B(\boldsymbol{x}, t)$ using weights with respect to the distance from $\boldsymbol{x}_F(t-1)$:

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

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

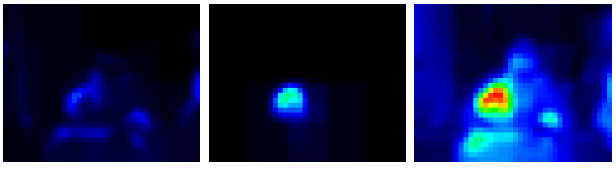
The weight $\alpha(\boldsymbol{x}, t)$ is a function of Cauchy distribution whose center is $\boldsymbol{x}_F(t-1)$, amplitude 1.0, and diameter D.

The resulting image is shown in the 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 demonstrator'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 $\boldsymbol{I}_B$.

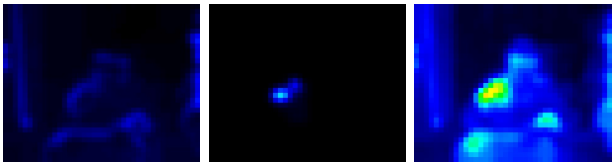
B. Saliency Computation

The effect of the retinal filtering can be observed in visual saliency. The system next computes the saliency for $\boldsymbol{I}_R$ using the model proposed by Itti et al. [12], [13]. This section gives only a brief description of the model. Refer to [12], [13] for a more detailed explanation.

The saliency is calculated as the difference between the focused pixels and the surroundings. Since robots are supposed



(a) Saliency maps calculated from *retinal* image



(b) Saliency maps calculated from *sharp* image

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

not to be able to distinguish demonstrators or objects from the background, all image pixels are processed equally. The system proposed here utilizes five features for the calculation of saliency: color, intensity, orientation, flicker, and motion. The first three are responsible for static features whereas the last two are for dynamic.

Fig. 3 shows the saliency maps derived from the retinal image (a) and from the sharp input image (b). The three maps correspond to the orientation feature (left), the motion (center), and the sum of the five features (right). The comparison of Fig. 3 (a) with (b) demonstrates that the retinal filtering can enhance the saliency in the fovea (i.e., the demonstrator's right hand with the green cup) while suppressing it in the periphery, which facilitates the stabilization of robots' attention.

C. Stochastic Attention Selection

The system then selects an image location to attend to based on the saliency. In order to maintain the sensitivity to a new prominent target, a stochastic algorithm for attention selection [25], [26] is adopted. Human attention is known to be stochastic rather than deterministic [27]. Even when humans scan the same picture, the paths of their attention are different between trials. It indicates that humans are sensitive to new signals and flexibly change their attention, which enables them to efficiently explore the environment.

First, the module for attention selection calculates the transition probability $\phi(\boldsymbol{x}, t)$ for all image locations, which defines the probability for robots to shift their 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 a positive value. That is, the probability becomes high if the saliency for $\boldsymbol{x}$ is higher than that for $\boldsymbol{x}_F(t-1)$. The module 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

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

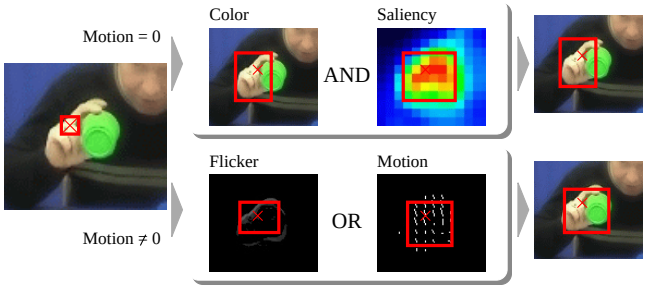


Fig. 4. Selective object extraction by spatial continuity. The cross and box indicate the attentional location and extracted object area, respectively.

Otherwise $\boldsymbol{x}_{F'}$ is accepted only at the 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. This mechanism enables robots to shift their attention to a more salient location but also to a less salient one with lower probability. The system repeats this process until a newly selected location satisfies the above condition.

IV. Object and Motion Learning Based on Continuity

The system next learns objects and motion based on the temporal and spatial continuity. Continuity is a basic physical principle, which infants can also detect in terms of color, size, and motion [28]. The mechanism for detecting continuity enables robots not only to enrich the information detected at the attentional location but also to examine its relevance to the demonstrated task.

A. Object Learning

The system extracts objects from the visual input based on the spatial continuity. Since the location selected by the above attentional mechanism is a small image area (8×8 pixels) and does not correspond to a certain object (e.g., it can contain only a part of an object or sometimes parts of several objects), it has to be expanded and enriched in terms of space.

Fig. 4 illustrates the mechanism. The system selectively uses the features to examine the spatial continuity. When the attentional location is static (the upper in Fig. 4), the system refers to the color and the saliency. If the neighbor regions have the same properties of these features, the object area denoted by a box is expanded to the regions. The reason for using the saliency in addition to the color is to eliminate irrelevant information. Because the attentional point is often on the contour of an object (e.g., the upper edge of the demonstrator's hand), the color continuity might include irrelevant features as well as the relevant (e.g., the demonstrator's body behind his hand). In such a case, the saliency helps robots discriminate the relevant information (i.e., the hand) from irrelevant (i.e., the body) because irrelevant locations likely have much lower saliency. When the attentional location is dynamic (the lower in Fig. 4), the continuity in terms of the flicker and the motion information are examined instead. If the neighbor regions have

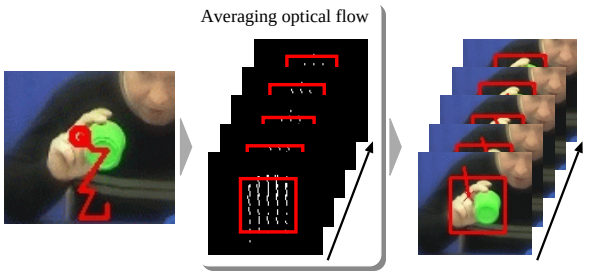


Fig. 5. Generative motion segmentation by spatial continuity. The optical flows detected in the object area are averaged to create a motion vector.

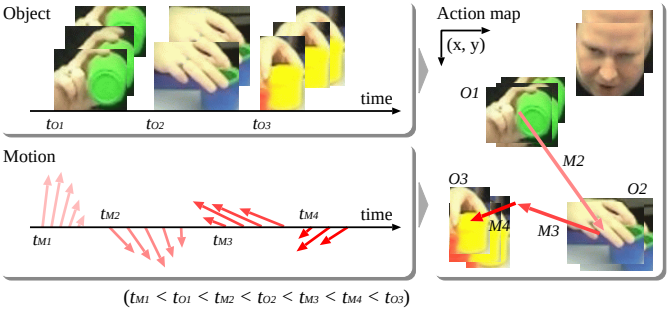


Fig. 6. Object and motion association by temporal and spatial continuity. The object chunks and the motion chunks are associated if they are relevant to each other.

the same amount of flicker or the same motion direction, the object area is expanded to the regions. This mechanism enables robots to extract an object together with the demonstrator's hand manipulating the object, which is important for robots to recognize the goal-orientedness of the task.

The extracted objects are then organized by the temporal continuity (see the upper left in Fig. 6). As long as the object images maintain the same color histogram, they are categorized into the same object chunk. For example, if the demonstrator's hand with the green cup is continuously detected, it creates an object chunk, whose histogram consists of the skin and the green color. When the demonstrator releases the green cup into the blue one and then starts handling the yellow cup, two new chunks are created respectively.

B. Motion Learning

The system also learns the motion concerning the object. Here a difficulty is that the trajectory of the attentional point is not smooth or continuous because the same object might not always be detected. Therefore, robots cannot simply trace the path of their attention but instead have to generatively learn smooth and meaningful motion segments.

Fig. 5 and the lower left part in Fig. 6 illustrate the mechanism for motion learning. The system first examines the spatial continuity to extract a motion vector. The optical flows detected from the object area are averaged in space. The vector is then organized by the temporal continuity as the object images. As long as the motion maintains the same direction, the vectors are categorized into the same chunk and cumulated into a meaningful motion segment. For example, the movement of lifting a cup (i.e., upward vectors), putting it down (i.e., downward vectors), and moving to the next cup (i.e., leftward vectors) can form each motion segment.

C. Object and Motion Association

The system finally associates the object and the motion in order to build an action map. Note that there is still *no* guarantee that the extracted information is relevant to the task. For example, the demonstrator's face might be extracted as an object chunk when he is talking to a robot. Such a social signal, on the one hand, assists robots in finding the action segment [29] but, on the other hand, has to be ignored when robots reproduce the task.

The key idea is the continuity across the modalities. Associating the object chunks with the motion chunks in terms of time and space enables robots to find the task relevance of the information. Fig. 6 illustrates the process, where the three object chunks and four motion chunks are already extracted. Let t_{Oi} and t_{Mj} be the time when the i-th object chunk and the j-th motion chunk are created, respectively. The displacement $\boldsymbol{X}(i)$ between the two consecutive object chunks is calculated by

$$\boldsymbol{X}(i) = \boldsymbol{x}_F(t_{Oi}) - \boldsymbol{x}_F(t_{Oi-1}). \tag{6}$$

The corresponding cumulative motion $\boldsymbol{M}(i)$ is defined as

$$\boldsymbol{M}(i) = \sum_{t_{Oi-1} < t_{Mj'} \leq t_{Oi}} \boldsymbol{m}(j'), \tag{7}$$

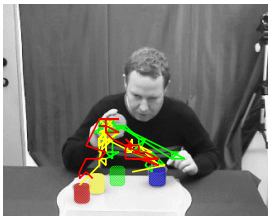
where $\boldsymbol{m}(j)$ is the cumulative vector for the j-th motion chunk. The association between the object chunks and the motion chunks is established if

$$\arccos \frac{\boldsymbol{X}(i) \cdot \boldsymbol{M}(i)}{\|\boldsymbol{X}(i)\| \, \|\boldsymbol{M}(i)\|} < \theta, \tag{8}$$

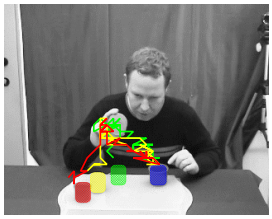
where θ is a threshold. That is, if the displacement of the object chunk is the same as the direction of the cumulative movement, the two object chunks are associated by the motion vector. In the example shown in Fig. 6, the green-cup chunk is connected with the blue-cup by the second motion chunk (i.e., the motion of putting down). The blue-cup is then connected with the yellow-cup by the third and fourth motion chunks (i.e., the motion of moving to the next cup). These associations finally build an action map, where only the relevant information is connected to each other. Note that even if the demonstrator's face is detected as an object chunk during this period, it can be ignored as irrelevant. The face chunk would not be connected with any other object chunks because the displacement between the demonstrator's face and the objects does not correspond to the cumulative movements.

V. Experiment

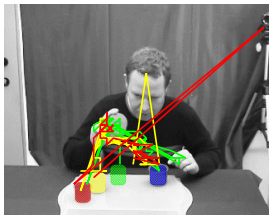
The proposed system was evaluated using pre-recorded videos where a father presented a cup-stacking task to his infant. The author has investigated parental scaffolding for infants' action learning as well as for robots' [18], [23]. The



(a) Stochastic algorithm *with* retinal filter

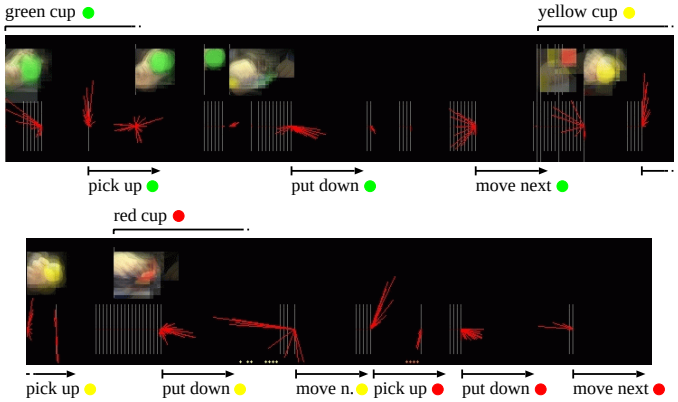


(b) Winner-take-all algorithm *with* retinal filter

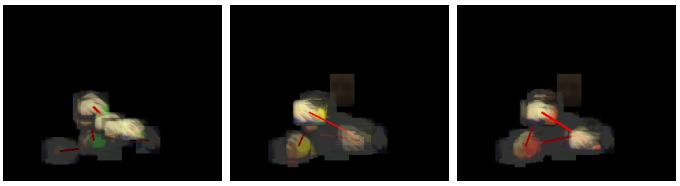


(c) Winner-take-all algorithm *without* retinal filter

Fig. 7. Transition of attention of proposed model (a) and two comparative models, (b) and (c). The line color corresponds to the cup color.



(a) Object and motion chunks created through task demonstration



(b) Action map for moving green, yellow, and red cups

Fig. 8. Extracted key actions and their association represented in time and space

focus of this experiment, however, is on the evaluation of the proposed system rather than on the analysis of parental teaching.

A. *Stability and Sensitivity of Bottom-Up Attention*

The first experiment focused on the stability and the sensitivity of the bottom-up attention. The effects of the retinal filtering and the stochastic attention selection were evaluated qualitatively.

Fig. 7 shows the result for the proposed model (a) as well as for two comparative models: a winner-take-all[1] model with the retinal filtering (b) and a model without the filtering (c). The three colored lines (red, yellow, and green) indicate the transition of the models' attention when the corresponding colored cups were moved into the blue one. First, comparing the result shown in Fig. 7 (b) with (c), we can see the stabilized attention achieved by the retinal filtering. Fig. 7 (b) demonstrates that the model's attention stably followed the cup-handling movement and never focused on irrelevant objects like a camera at the upper-right corner of the image. Note also that in (b) each colored trajectory reached only the corresponding cup, whereas this was not the case in (c) (i.e., the red cup attracted the attention even while the other cups were being manipulated). This was due to the high color saliency for the red cup although the cup was irrelevant at that moment. The retinal filtering could prevent such an undesired situation by enhancing the saliency in the foveal vision.

Next, comparing the result shown in Fig. 7 (a) with (b), we can see the sensitivity achieved by the stochastic attention. In this scene, the demonstrator was pointing to the blue cup with his left hand in order to indicate the goal position for the holding cup. The proposed model (a) as well as the comparative one without the retinal filtering (c) could attend to the pointing hand, whereas (b) could not. In action learning, robots should be able to rapidly respond to such a new movement because it is likely related to the current action. The experimental result showed that the stochastic attention selection coupled with the retinal filtering could meet the contradictory requirements of stability and sensitivity.

[1]Winner-take-all is an algorithm to select the most salient location as an attentional location in the image.

B. *Extracting Key Actions Based on Continuity*

The second experiment evaluated the system's ability to extract and associate key points in the demonstrated action.

Fig. 8 shows the result: (a) the object and the motion chunks created through the demonstration, and (b) their spatial association in the action map. First, we can see from Fig. 8 (a) that the proposed system adequately created the chunks. The green, the yellow, and the red cups with the demonstrator's hand were sequentially extracted as the object chunks. The motion chunks corresponded to the objects; for each object, the picking-up (i.e., upward vectors), the putting-down (i.e., downward vectors from left to right), and the moving-next chunks (i.e., leftward vectors) were sequentially extracted. Note that the creation for the motion chunks did not always coincide with for the objects. For example, an image of the red cup can be seen on the left side of the motion chunk for putting-down the yellow cup. This indicates that the system continuously traced the movement of the demonstrator's hand, and thus the object images changed seamlessly enough to be recognized as the same object.

Fig. 8 (b) shows the corresponding action maps, where the object chunks were associated by the motion chunks. The map for each cup is separately represented for clarity. We can see from the result that a triangular association, whose corners were the key points in the action, was generated for each cup.

The bottom left corner of the triangle corresponded to the action of grasping a cup, the top to lifting it, and the bottom right to releasing it into the blue one. These three were the important actions to achieve the cup-stacking task and were strongly related to the goal (or the sub-goals) of the task. The result indicated that the system was able to learn the task not only at the motion level but also at the goal level.

VI. Conclusion

Investigating potentials of bottom-up faculties is a common subject in developmental robotics. It would lead to a better design of cognitive models as well as to a deeper understanding of human development [30]. This research has addressed the challenge of developing an action learning model employing bottom-up visual attention. The proposed system overcame the difficulties by introducing biologically- and developmentally-inspired mechanisms. The retinal filtering combined with the stochastic algorithm achieved the stability and the sensitivity of bottom-up attention. They enhanced the system's attention to the fovea while maintaining its sensitivity to the peripheral vision. Examining continuity enabled the system to enrich the information detected at the attentional location. The system could extract meaningful segments of objects and movements, and associate them to build a map representing key actions.

Integrating a mechanism for learning primitives would improve the performance. Object and motion features can be clustered to form primitives. The primitives could then be used to learn their association not only at the temporal or spatial level but also at the conceptual level, which facilitates the recognition and the imitation of actions. Another interesting issue from a developmental perspective is to analyze different tasks as well as different teaching-learning scenarios. My hypothesis is that key points in a demonstrated action will be extracted differently depending on the goal-orientedness of the task and on the ability of learners [23]. Such analytical studies employing bottom-up architectures would contribute to uncovering human development.

Acknowledgment

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

References

[1] A. Billard and R. S. Eds., "Special issue: Robot learning from demonstration," *Robotics and Autonomous Systems*, vol. 47, no. 2–3, 2004.

[2] C. L. Nehaniv and K. Dautenhahn, "Like me? measures of correspondence and imitation," *Cybernetics and Systems: An International Journal*, vol. 32, pp. 11–51, 2001.

[3] C. Breazeal and B. Scassellati, "Challenges in building robots that imitate people," in *Imitation in Animals and Artifacts*, K. Dautenhahn and C. L. Nehaniv, Eds. MIT Press, 2002, pp. 363–389.

[4] B. Scassellati, "Knowing what to imitate and knowing when you succeed," in *Proceedings of the AISB'99 Symposium on Imitation in Animals and Artifacts*, 1999, pp. 105–113.

[5] A. Alissandrakis, C. L. Nehaniv, and K. Dautenhahn, "Imitation with alice: Learning to imitate corresponding actions across dissimilar embodiments," *IEEE Transactions on Systems, Man, and Cybernetics–Part A: Systems and Humans*, vol. 32, no. 4, pp. 482–496, 2002.

[6] A. Billard, Y. Epars, S. Calinon, S. Schaal, and G. Cheng, "Discovering optimal imitation strategies," *Robotics and Autonomous Systems*, vol. 47, pp. 69–77, 2004.

[7] A. G. Billard, S. Calinon, and F. Guenter, "Discriminative and adaptive imitation in uni-manual and bi-manual tasks," *Robotics and Autonomous Systems*, vol. 54, no. 5, pp. 370–384, 2006.

[8] S. Calinon, F. Guenter, and A. Billard, "Goal-directed imitation in a humanoid robot," in *Proceedings of the International Conference on Robotics and Automation*, 2005.

[9] S. Calinon and A. Billard, "A framework integrating statistical and social cues to teach a humanoid robot new skills," in *Proceedings of the ICRA Workshop on Social Interaction with Intelligent Indoor Robots*, 2008.

[10] Y. Demiris and B. Khadhouri, "Hierarchical attentive multiple models for execution and recognition of actions," *Robotics and Autonomous Systems*, vol. 54, pp. 361–369, 2006.

[11] I. D. Goga and A. Billard, "Attention mechanisms for the imitation of goal-directed action in developmental robots," in *Proceedings of the 6th International Conference in Epigenetic Robotics*, 2006.

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

[13] L. Itti, N. Dhavale, and F. Pighin, "Realistic avatar eye and head animation using a neurobiological model of visual attention," in *Proceedings of the SPIE 48th Annual International Symposium on Optical Science and Technology*, 2003, pp. 64–78.

[14] J. Ruesch, M. Lopes, A. Bernardino, J. Hornstein, J. Santos-Victor, and R. Pfeifer, "Multimodal saliency-based bottom-up attention: A framework for the humanoid robot iCub," in *Proceedings of the IEEE International Conference on Robotics and Automation*, 2008, pp. 962–967.

[15] N. J. Butko and J. R. Movellan, "I-pomdp: An infomax model of eye movement," in *Proceedings of the IEEE International Conference on Development and Learning*, 2008.

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

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

[18] Y. Nagai, C. Muhl, and K. J. Rohlfing, "Toward designing a robot that learns actions from parental demonstrations," in *Proceedings of the 2008 IEEE International Conference on Robotics and Automation*, 2008, pp. 3545–3550.

[19] 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 *Proceedings of the 5th IEEE International Conference on Development and Learning*, 2006.

[20] M. Hikita, S. Fuke, M. Ogino, T. Minato, and M. Asada, "Visual attention by saliency leads cross-modal body representation," in *Proceedings of the 7th IEEE International Conference on Development and Learning*, 2008.

[21] M. Schlesinger, D. Amso, and S. P. Johnson, "The neural basis for visual selective attention in young infants: A computational account," *Adaptive Behavior*, vol. 15, no. 2, pp. 135–148, 2007.

[22] F. Shic, B. Scassellati, D. Lin, and K. Chawarska, "Measuring context: The gaze patterns of children with autism evaluated from the bottom-up," in *Proceedings of the 6th IEEE International Conference on Development and Learning*, 2007.

[23] Y. Nagai and K. J. Rohlfing, "Parental action modification highlighting the goal versus the means," in *Proceedings of the IEEE 7th International Conference on Development and Learning*, 2008.

[24] S. Core, C. Porac, and L. M. Ward, *Sensation and perception*. Academic Press: New York, 1978.

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

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

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

[28] F. Vital-Durand, J. Atkinson, and O. J. Braddick, Eds., *Infant Vision*. Oxford University Press, 1996.

[29] Y. Nagai and K. J. Rohlfing, "Parental signal indicating significant state change in action demonstration," in *Proceedings of the 7th International Conference on Epigenetic Robotics*, 2007, pp. 205–206.

[30] M. Asada, K. F. MacDorman, H. Ishiguro, and Y. Kuniyoshi, "Cognitive developmental robotics as a new paradigm for the design of humanoid robots," *Robotics and Autonomous Systems*, vol. 37, pp. 185–193, 2001.