2014 IEEE International Conference on Robotics & Automation (ICRA)
Hong Kong Convention and Exhibition Center
May 31 - June 7, 2014. Hong Kong, China

Visual Attention Model for Manipulating Human Attention by a Robot*

Yusuke Tamura[1], Shiro Yano[2] and Hisashi Osumi[1]

***Abstract*— For smooth interaction between human and robot, the robot should have an ability to manipulate human attention and behaviors. In this study, we developed a visual attention model for manipulating human attention by a robot. The model consists of two modules, such as the saliency map generation module and manipulation map generation module. The saliency map describes the bottom-up effect of visual stimuli on human attention and the manipulation map describes the top-down effect of face, hands and gaze. In order to evaluate the proposed attention model, we measured human gaze points during watching a magic video, and applied the attention model to the video. Based on the result of this experiment, the proposed attention model can better explain human visual attention than the original saliency map.**

I. Introduction

In human-human interaction, we usually estimate other's intention and predict his/her behaviors according to the observation. Based on the estimation and prediction, we decide our own behaviors.

In the research field of human-robot interaction, estimation of human intention has been actively studied [1]–[5].

However, such "estimation-first" approaches are not enough for smooth interaction between human and robot. Humans sometimes actively act before observation to manipulate other's behaviors. In order for robots to coexist and smoothly interact with humans, the ability to manipulate human's behaviors is required. Overview of the smooth interaction process between robot and human is shown in Fig. 1.

Furthermore, in order to control human's behaviors, it is necessary for the robots to manipulate human's attention. To manipulate human's attention, the robots must know how their behaviors and surrounding environments affect human's attention. In this study, therefore, we aim for developing a model of human visual attention.

One of the most popular studies on a computational model of human visual attention is the saliency map model by Itti *et al* [6]. The model calculates saliency in a bottom-up manner based on colors, intensity and orientations of an input image. The model could reasonably simulate human's visual attention to some extent. In real situation, however, human attention can be affected not only by bottom-up factors but also by top-down factors, such as meaning of surrounding objects, other person's existence and behaviors, and context. Cerf *et al.* proposed a model integrating saliency and face detection [7]. Ozeki *et al.* proposed a spotlight-type attention model based on a particle filter [8]. The model considered visual saliency of the input image and position and direction of human face.

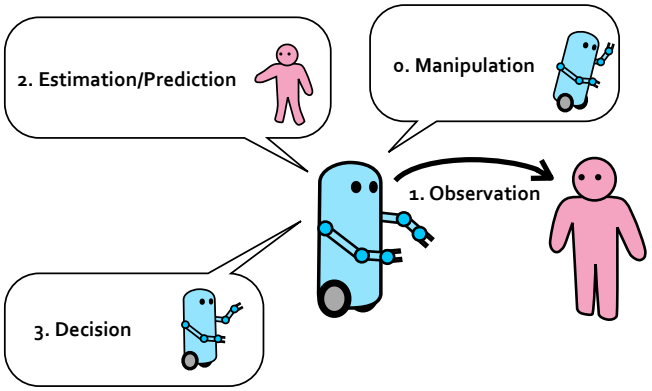

Fig. 1. Smooth interaction process between robot and human

From a perspective of manipulating other's attention, magicians are experts [9], [10]. The magic performance requires a *method* (how the trick works) to achieve an *effect* (what the spectator perceives) [11]. Magicians manipulate spectators' attention in order to prevent them from uncovering the methods by various skills, such as misdirection [12], [13]. Spectators' attention is effectively controlled by skillful magicians' eye gaze [13]. Moreover, even in the same magic trick, curved hand motions can attract more human attention than straight motions [14]. We have analyzed human eye gaze during watching a video of magic trick, and the result suggested that the spectators' attention is affected by the relationship between magician's gaze and hand [15]. In other words, human hand and gaze do not work independently for manipulating other's attention; their relationship is important.

Based on these facts, the objective of this study is to develop a model of human visual attention, which considers the saliency as a bottom-up factor and the relationship between human face, hands and gaze as a top-down factor.

In section II, we present our proposed model of visual attention. Section III explains the measurement of human gaze during watching a magic trick. In section IV, the evaluation of the proposed model is conducted based on the measurement of human gaze. Section V concludes the paper and addresses our future plans.

*This work was supported by the Japan Society for the Promotion of Science, Grant-in-Aid for Young Scientists (B), 24700190.

[1]Y. Tamura and H. Osumi are with the Department of Precision Mechanics, Faculty of Science and Engineering, Chuo University, Japan (e-mail: tamura@mech.chuo-u.ac.jp)

[2]S. Yano is with Research Organization of Science and Technology, Ritsumeikan University, Japan.

II. Model of visual attention

The proposed visual attention model consists of the following two modules:

- Saliency map generation module
- Manipulation map generation module

The former module calculates visual saliency of input images in a bottom-up manner, and the latter module calculates a manipulation map in a top-down manner, which describes the effect of the relationship between human face, hand and gaze. Figure 2 shows an overview of the proposed visual attention model.

A. Saliency map generation module

The saliency map generation module considers motion in addition to the original saliency map [6]. In other words, the module calculates the following four types of feature maps:

- intensity
- colors ($\times 2$)
- orientations ($\times 4$)
- motion

The feature maps of intensity, colors and orientations are generated in the same manner as the original saliency map. The feature of motion is calculated as magnitudes of optical flow [16]. The saliency map is obtained as a linear sum of normalized these features.

$S_s(\boldsymbol{x})$, the saliency value of position $\boldsymbol{x}$ at time s, is calculated as follows:

$$S_s(\boldsymbol{x}) = \frac{1}{4}\left(\bar{I}_s(\boldsymbol{x}) + \bar{C}_s(\boldsymbol{x}) + \bar{O}_s(\boldsymbol{x}) + \bar{M}_s(\boldsymbol{x})\right) \quad (1)$$

where $\bar{I}_s(\boldsymbol{x})$, $\bar{C}_s(\boldsymbol{x})$, $\bar{O}_s(\boldsymbol{x})$ and $\bar{M}_s(\boldsymbol{x})$ are normalized feature value of intensity, colors, orientations and motion of position $\boldsymbol{x}$ at time s, respectively.

B. Manipulation map generation module

The manipulation map is to describe an effect of the relationship between face, hands and gaze on other's attention. The map is obtained as a linear sum of face-hand map and gaze map. The face-hand map describes the positions of human face and hands in the input image. The gaze map describes the human gaze direction.

Figure 3 shows the parameter definition for generating manipulation map.

Here, $\boldsymbol{x}_{\mathrm{F}}$ is a position of the face, and $\boldsymbol{x}_{\mathrm{R}}$ and $\boldsymbol{x}_{\mathrm{L}}$ are positions of the right and left hand, respectively.

The face-hand map is generated through the following calculation at each pixel position $\boldsymbol{x}$.

$$F(\boldsymbol{x}) = \sum_i \frac{1}{2\pi\sqrt{|\boldsymbol{S}_i|}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_i)^T \boldsymbol{S}_i^{-1}(\boldsymbol{x}-\boldsymbol{x}_i)\right\}, \quad (2)$$

where $i = \mathrm{F}, \mathrm{R}, \mathrm{L}$ and $\boldsymbol{S}_i$ is a covariance matrix of $\boldsymbol{x}_i$. In other words, the face-hand map is calculated as the superposition of the two-dimensional normal distribution for face and both hands.

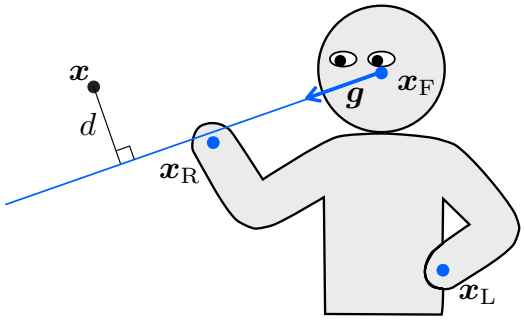

Fig. 3. Parameter definition for generation of manipulation map

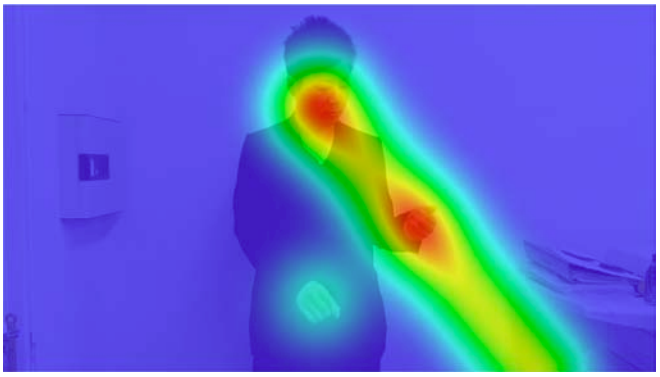

Fig. 4. An example of the manipulation map

The initial point of the gaze vector $\boldsymbol{g}$ is the face position $\boldsymbol{x}_{\mathrm{F}}$. Defining $d(\boldsymbol{x})$ as the distance between $\boldsymbol{x}$ and the gaze half line from $\boldsymbol{x}_{\mathrm{F}}$, the gaze map is generated through the following calculation at each pixel position $\boldsymbol{x}$.

$$G(\boldsymbol{x}) = \frac{1}{\sqrt{2\pi}\sigma_d} \exp\left(-\frac{d(\boldsymbol{x})^2}{2\sigma_d^2}\right) \quad (3)$$

$H_s(\boldsymbol{x})$, the manipulation value of position $\boldsymbol{x}$ at time s, is calculated as follows:

$$H_s(\boldsymbol{x}) = k_1 F_s(\boldsymbol{x}) + (1-k_1) G_s(\boldsymbol{x}), \quad (4)$$

where $F_s(\boldsymbol{x})$ and $G_s(\boldsymbol{x})$ are the values of face-hand map and gaze map of $\boldsymbol{x}$ at time s, respectively. k_1 is a weight value.

Figure 4 shows an example of the manipulation map overlaid to the input image.

In the figure, because the person looks at his left hand, the calculated manipulation value is high around his face and left hand.

C. Integration of saliency map and manipulation map

The proposed attention map is obtained as a linear sum of the saliency map and the manipulation map. That is, $V_s(\boldsymbol{x})$, the attention value of position $\boldsymbol{x}$ at time s, is calculated as follows:

$$V_s(\boldsymbol{x}) = k_2 S_s(\boldsymbol{x}) + (1-k_2) H_s(\boldsymbol{x}), \quad (5)$$

where k_2 is a weight value.

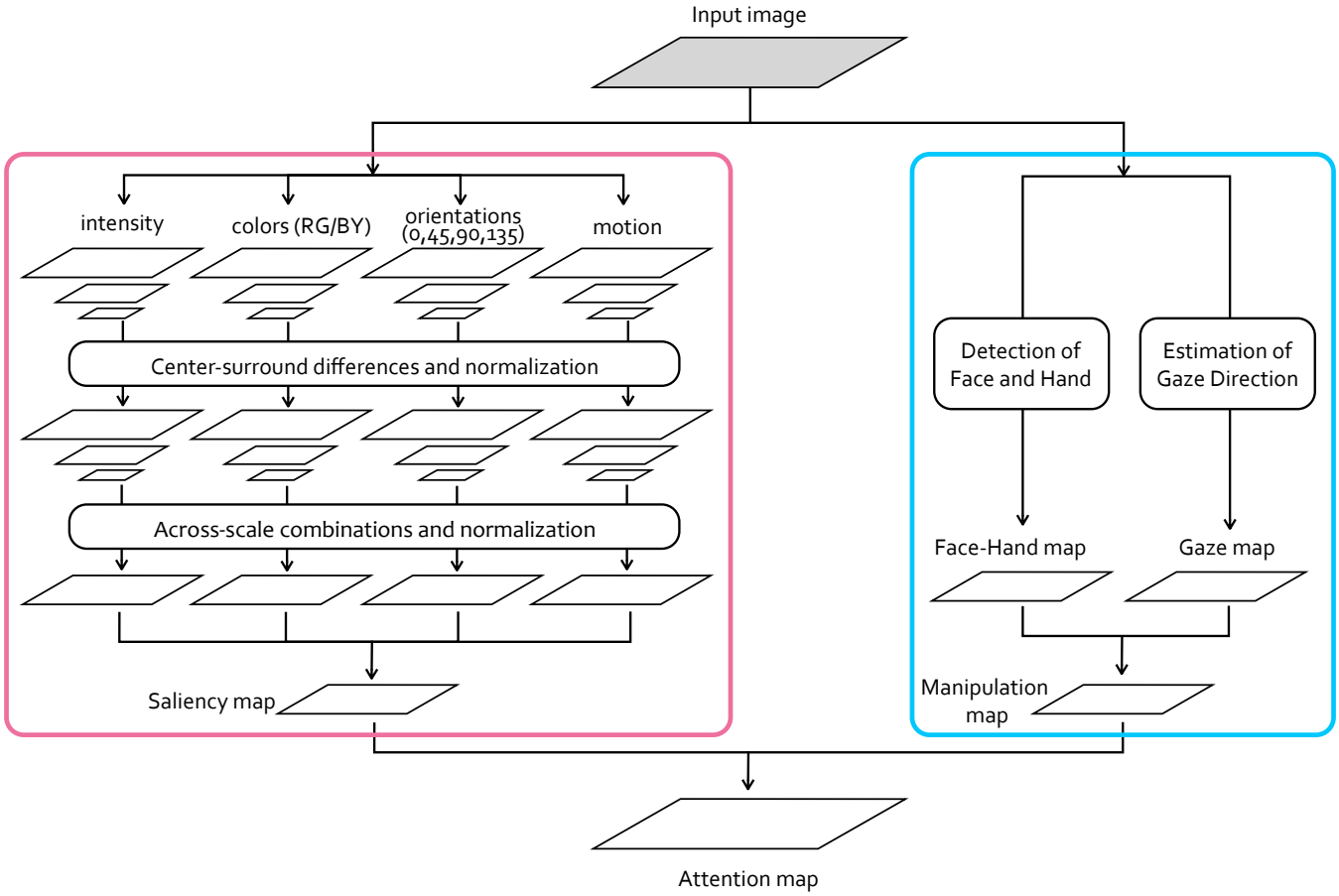

Fig. 2. Proposed visual attention model considering saliency and manipulation of attention

III. Measurement of Eye Gaze during Watching a Magic Video

Generally, attention is classified into two types, such as covert attention and overt attention [17]. Covert attention independently moves with eye movements and overt attention follows the gaze points. Because direct measurement of covert attention is impossible, we dealt with human overt attention in this paper.

To investigate how human attention was manipulated, therefore, we carried out an experiment of measuring human gaze points during watching a magic video.

A. Magic video

In the video, a magician performs "vanishing a ball" trick (Fig. 5). The magician shows a pink-colored sponge ball (43 mm in a diameter) in his right hand, then apparently passes it from his right hand to the left hand, but secretly retains it in the right hand. After that, he opens his left hand and the ball apparently vanished.

As mentioned in section I, magic tricks consist of *methods* and *effects*. These factors of the experimental magic trick are as follows:

- **method:** a false transfer of the ball
- **effect:** vanish of the ball

Magicians usually manipulate spectators' attention not to be detected the method by using *misdirection*. In this video, however, the magician dare to *direct* spectators' attention to his hands where the method is performed.

The time length of the video clip is about 10 seconds, and the resolution is 1920 $\times$ 1080 pixels.

B. Measurement of eye gaze

In the experiment, five healthy male volunteers (aged 21–24 years) participated.

The gaze points of the participants were recorded while they watched the magic video. As shown in Fig. 6, each participant was required to seat in front of a 23-inch computer screen and watch the video. The distance between the screen and participant was about 0.86 m.

The participants' head and eye movements were measured by a contactless eye tracking system (faceLAB 4.2, Seeing Machines) with two CCD cameras (FCB-EX480B, SONY). The system measured the head and eye movements at 60 Hz, and calculated the gaze points based on the measurement.

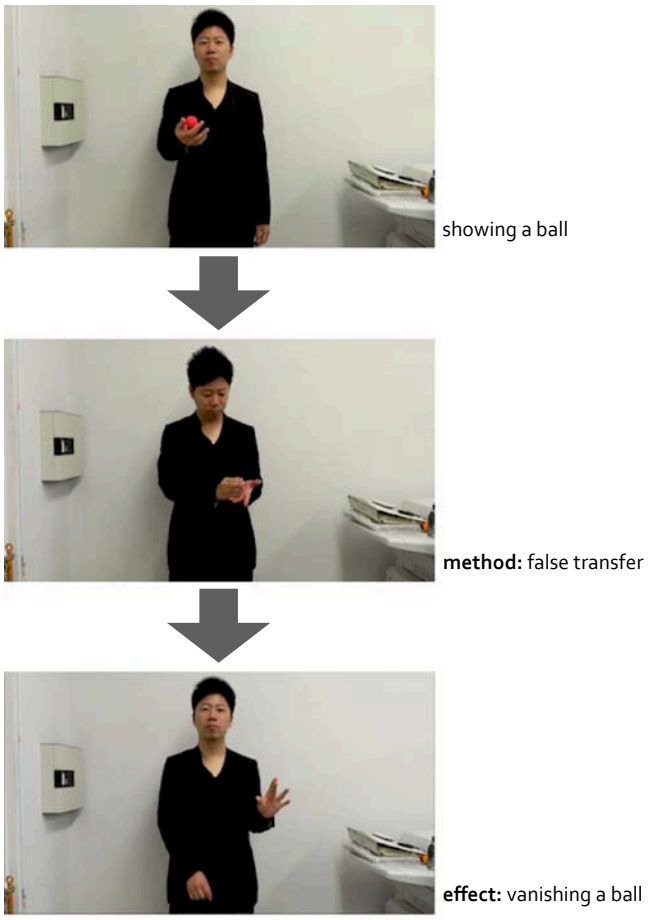

Fig. 5. "Vanishing a ball" trick

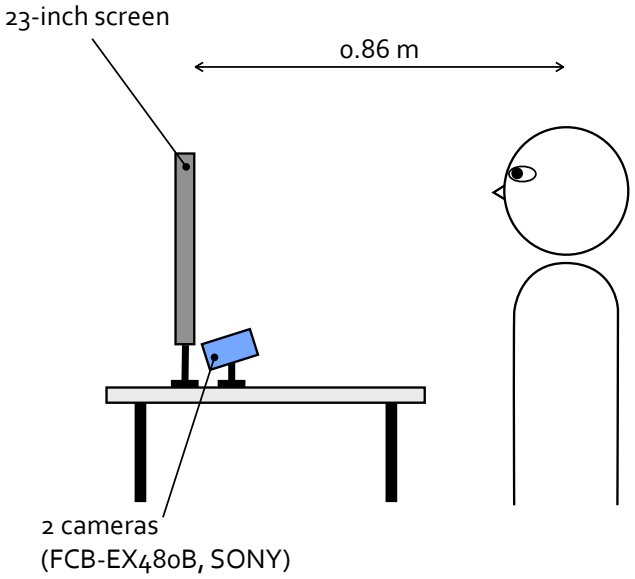

Fig. 6. Measurement setup

IV. Evaluation of the Proposed Attention Model

A. Evaluation method

We compared the proposed attention model to the saliency map. The evaluation index was the ratio of the attention value at the gaze point divided by the overall average at the same timing. In other words, the evaluation index of the proposed attention map at time s is

$$\frac{V_s(\boldsymbol{x}_g)}{\bar{V}_s}, \tag{6}$$

and that of the saliency map is

$$\frac{S_s(\boldsymbol{x}_g)}{\bar{S}_s}, \tag{7}$$

where $\boldsymbol{x}_g$ is the gaze point, and $\bar{V}_s$ and $\bar{S}_s$ are the average attention value and saliency value at time s, respectively. In this paper, the both weight values k_1 and k_2 are 0.5. In other words, the face-hand map and gaze map carry the same weight, and the manipulation map and saliency map also carry the same weight.

In this evaluation, the positions of magician's face and hands were calculated through image processing with OpenCV 2.4. Concretely, the input image was converted from RGB to HSV, and thresholded for skin color extraction. After that, the skin color areas were corresponded to the face and hands according to the positions of the areas.

For simplicity, the magician's gaze direction was determined based on the interview with the magician. Each frame of the magic video was classified into two scenes, such as "facing front" and "looking at the left hand." In the "facing front" scene, the gaze vector $\boldsymbol{g} = \boldsymbol{0}$. In the "looking at the left hand" scene, the gaze vector was calculated as follows:

$$\boldsymbol{g} = \frac{\boldsymbol{x}_{\mathrm{L}} - \boldsymbol{x}_{\mathrm{F}}}{\|\boldsymbol{x}_{\mathrm{L}} - \boldsymbol{x}_{\mathrm{F}}\|} \tag{8}$$

Here, the covariance matrix in eq. (2) was defined as follows:

$$\boldsymbol{S}_i = \mathrm{diag}[\sigma_{x_i}^2, \sigma_{y_i}^2], \tag{9}$$

where σ_{x_i}, σ_{y_i} and σ_d were empirically set to 96.

B. Result

The sample frames of the original video and the three maps are shown in Fig. 7.

Generally, the saliency values of the areas with high-contrast were higher than other areas. On the other hand, the magician's face and left hand scored high manipulation value as we designed. Integrating the saliency map and manipulation map, the proposed attention map seems to be reasonable.

In the scene 1, the magician raised his right hand with a ball in the hand and looks straight ahead. The highest manipulation value and attention value were around his face. In the scene 2, the magician showed the ball to spectators. The saliency around his right hand and the ball scored high value. In the scene 3, he raised his left hand and turned his eyes to the left hand. Therefore, the manipulation value around his face and left hand scored highest. In the scene 4, he "passed" the ball from his right hand to the left hand.

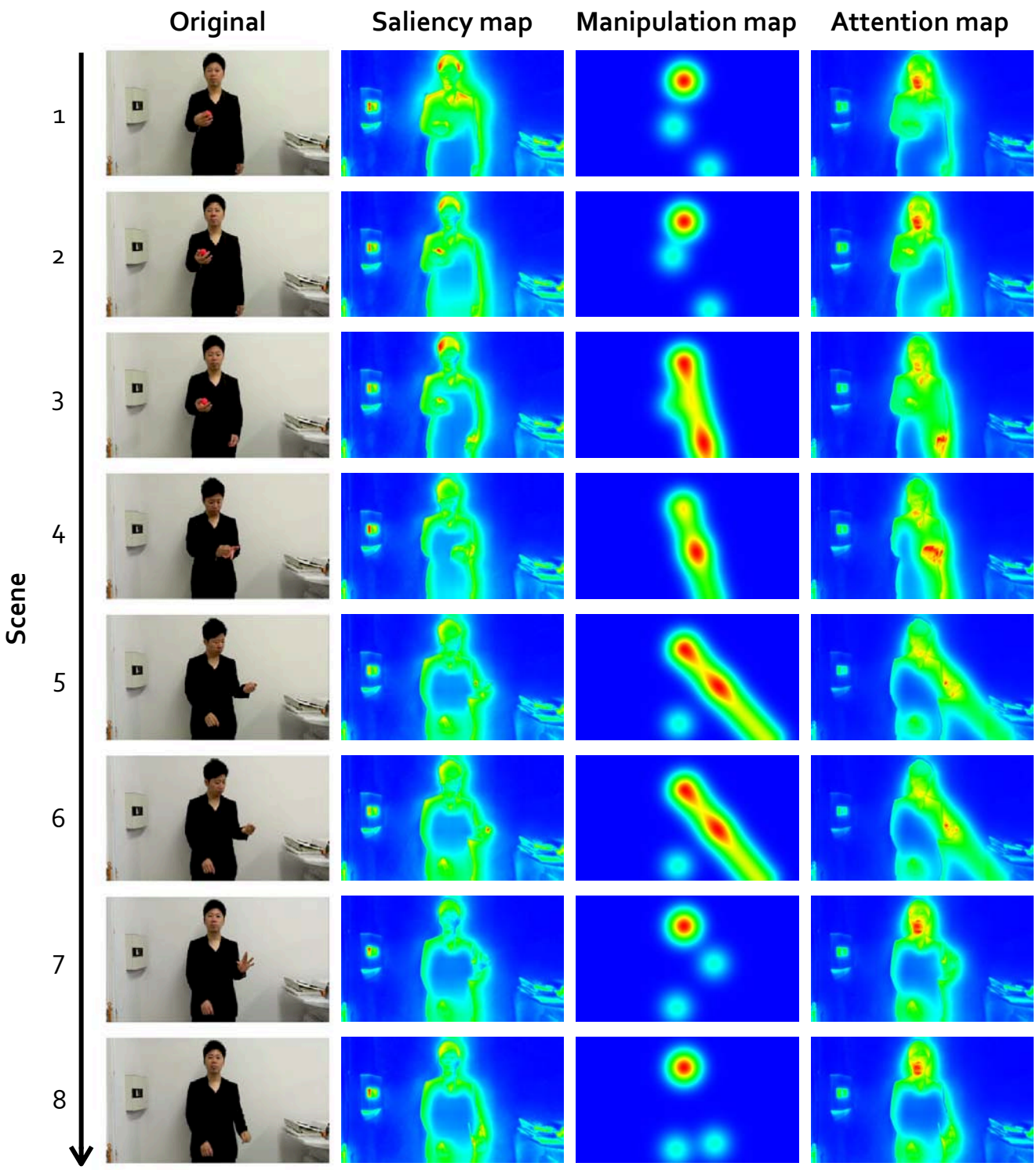

Fig. 7. Sample frames of the original video and the saliency, manipulation and attention maps

In reality, the ball remained in his right hand. In this scene, the highest manipulation and attention values were around his hands. In the scene 5, he lowers his right hand and simultaneously raises the left hand. After that, he rubs the fingers of his left hand to "vanish" the ball in the scene 6. In these scenes, the attention value around his right hand scored higher than the other area. In the scene 7, he turned his eyes from his left hand to the front and opens his left hand to show the effect. Finally he lowered his left hand and the video ends after the scene 8. In these scenes, the highest

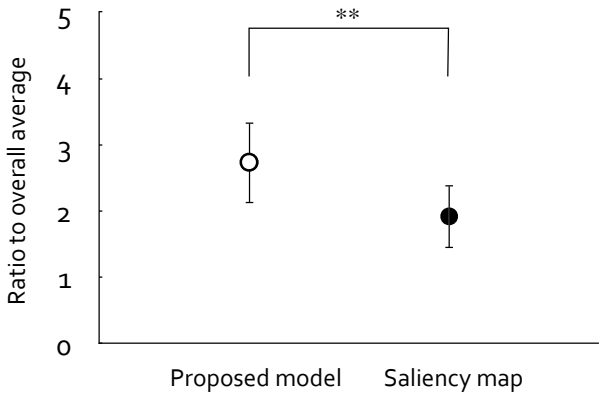

Fig. 8. Comparison of the evaluation indexes at participants' gaze point between proposed method and saliency map

manipulation and attention values were around his face like in the scene 1.

Figure 8 shows the evaluation indexes at participants' gaze point of the proposed method and saliency map described in eqs. (6) and (7).

As the figure shows, the evaluation index of the proposed model was higher than that of the saliency map. According to the paired t-test, there was a significant difference between them ($p < 0.01$). Based on the fact, the proposed visual attention model can better explain human attention than the saliency map.

V. Conclusion

In this study, we aim for the development of robot that can manipulate human attention. In this paper, we proposed and developed the visual attention model integrated the saliency map with the manipulation map, which described the effect of relationship between face, hands and gaze.

Through the measurement of gaze points of spectators who were watching a magic video, the proposed model can better describe human attention than the original saliency map.

For future works, we plan to implement the proposed model into a real robot, and develop a method to estimate how the robot's behavior affects surrounding human attention.

References

[1] T. Sato, Y. Nishida, J. Ichikawa, Y. Hatamura, and H. Mizoguchi, "Active Understanding of Human Intention by a Robot Through Monitoring of Human Behavior," *Proceedings of the IEEE/RSJ/GI International Conference on Intelligent Robots and Systems '94*, pp.405–414, 1994.

[2] K. Sakita, K. Ogawara, S. Murakami, K. Kawamura, and K. Ikeuchi, "Flexible Cooperation between Human and Robot by Interpreting Human Intention from Gaze Information," *Proceedings of the 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp.846–851, 2004.

[3] Y. Tamura, M. Sugi, J. Ota, and T. Arai, "Deskwork Support System Based on the Estimation of Human Intentions," *Proceedings of the 13th IEEE International Workshop on Robot and Human Interactive Communication*, pp.413–418, 2004.

[4] K. A. Tahboub, "Intelligent Human-Machine Interaction Based on Dynamic Bayesian Networks Probabilistic Intention Recognition," *Journal of Intelligent and Robotic Systems*, vol.45, no.1, pp.31–52, 2006.

[5] A. J. Schmid, O. Weede, and H. Worn, "Proactive Robot Task Selection Given an Human Intention Estimate," *Proceedings of the 16th IEEE International Symposium on Robot and Human Interactive Communication*, pp.726–731, 2007.

[6] 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.1254–1258, 1998.

[7] M. Cerf, J. Harel, W. Einhäuser, and C. Koch, "Predicting Human Gaze Using Low-Level Saliency Combined with Face Detection," *Advances in Neural Information Processing Systems*, vol.20, pp.241–248, 2008.

[8] M. Ozeki, Y. Kashiwagi, M. Inoue, and N. Oka, "Top-Down Visual Attention Control Based on a Particle Filter for Human-Interactive Robots," *Proceedings of the 4th International Conference on Human System Interaction*, pp.188–194, 2011.

[9] G. Kuhn, A.A. Amlani, and R.A. Rensink, "Towards a science of magic," *Trends in Cognitive Sciences*, Vol.12, No.9, pp.349–354, 2008.

[10] S.L. Macknik, M. King, J. Randi, A. Robbins, Teller, J. Thompson, and S. Martinez-Conde, "Attention and Awareness in Stage Magic: Turning Tricks into Research," *Nature Reviews Neuroscience*, Vol.9, No.11, pp.871–879, 2008.

[11] P. Lamont and R. Wiseman, *Magic in Theory — An Introduction to the Theoretical and Psychological Elements of Conjuring*, University of Hertfordshire Press, p.29, 1999.

[12] G. Kuhn, B.W. Tatler, J. M. Findlay, and G.G. Cole, "Misdirection in Magic: Implications for the Relationship between Eye Gaze and Attention," *Visual Cognition*, Vol.16, pp.391–405, 2008.

[13] G. Kuhn, B. W. Tatler, and G. G. Cole, "You Look Where I Look! Effect of Gaze Cues on Overt and Covert Attention in Misdirection," *Visual Cognition*, vol.17, pp.925–944, 2009.

[14] J. Otero-Millan, S. L. Macknik, A. Robbins, and S. Martinez-Conde, "Stronger Misdirection in Curved Than in Straight Motion," *Frontiers in Human Neuroscience*, vol.5, 133, pp.1–4, 2011.

[15] T. Akashi, Y. Tamura, S. Yano, and H. Osumi, "Analysis of Manipulating Other's Attention for Smooth Interaction between Human and Robot," *2013 IEEE/SICE International Symposium on System Integration*, submitted.

[16] G. Farnebäck, "Two-Frame Motion Estimation Based on Polynomial Expansion," *Proceedings of the 13th Scandinavian Conference on Image Analysis*, pp.363–370, 2003.

[17] A.R. Hunt and A. Kingstone, "Covert and Overt Voluntary Attention: Linked or Independent?," *Cognitive Brain Research*, Vol.18, pp.102–105, 2003.