Estimation of User's Attention and Awareness in Occlusion-Rich Environments Using RGB-D Cameras

Jun-Ichi Imai, Masanori Nemoto

Estimation of User's Attention and Awareness in Occlusion-rich Environments Using RGB-D Cameras

Jun-ichi Imai† and Masanori Nemoto

Chiba Institute of Technology
2–17–1 Tsudanuma, Narashino-shi, Chiba 275–0016, Japan
† imai@cs.it-chiba.ac.jp
http://www.imai.cs.it-chiba.ac.jp/

Abstract. Objective recognition by systems often does not agree with subjective recognition by users. Therefore, it is an important to estimate users' subjective states appropriately. Especially, in occlusion-rich environments, information on what a user can/cannot see, what he/she pays attention to, and what he/she is aware of or not in the environments is one of important clues to estimate his/her subjective states and predict next actions. In this paper, we propose a system for estimating maps of a user's attention and awareness in such environments based on the view estimation system using RGB-D cameras. The proposed system can estimate what the user sees, what he/she pays attention to, and what he/she is aware of in environments in pixels of captured images. Experimental results in a real environments show effectiveness of the proposed system. Furthermore, we discuss an extension of the proposed system to estimation for multiple users.

Keywords: Attention, Awareness, Occlusion-rich Environment, RGB-D Camera, View Estimation

1 Introduction

Human-symbiotic systems, which share humans' living spaces and assist with their activities by various means, have been increasingly studied in recent years (e.g. [1]). In order to realize such systems, a lot of techniques for recognizing states of users and their circumstances have been proposed. However, such *objective* recognition by the systems often does not agree with *subjective* recognition by users. Such a perception gap will cause disagreement between assistance provided by the systems and one which users really need. Therefore, it is an important problem to estimate users' subjective states appropriately.

Generally, in our living spaces, there are a lot of objects which can cause visual occlusion. Since a system and a user will observe the environment from different positions, it often occurs that one cannot see an object by occlusion while the other can. This is a typical example of the perception gap between the systems and users [2]. In such environments, information on what a user

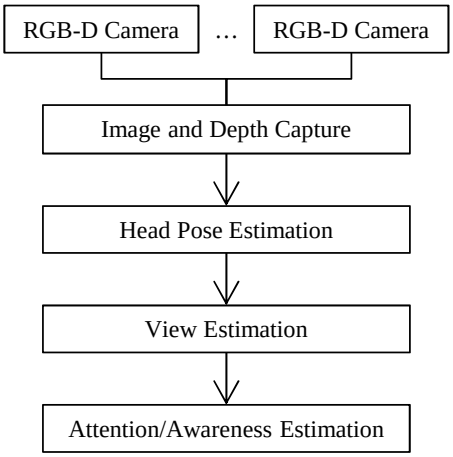


Fig. 1. Processing Flow of Proposed System

can/cannot see, what he/she pays attention to, and what he/she is aware of in the environments will be one of important clues to estimate his/her subjective states and predict the next action.

In this paper, based on the view estimation using multiple RGB-D cameras [3], we propose a system for estimating maps of a user's attention and awareness in occlusion-rich environments. We simply define a user's *attention* as watching something continuously during a short period, and *awareness* as having seen something before and knowing that it is there. The proposed system can estimate what the user sees, what he/she pays attention to, and what he/she is aware of in environments in pixels of captured images. The system can estimate them without his/her wearing cameras or other equipments. This point is the advantage of our proposed method.

Fig. 1 shows the processing flow of the proposed system. The flow can be divided into four parts; the image and depth capture module, the head pose estimation module, the view estimation module, and the attention/awareness estimation module. After capturing images and depth information from the cameras, the system estimates the user's head pose horizontally and vertically. Then, the user's field of view is estimated using the specified head pose. And finally, based on the specified field of view, maps of the user's attention and awareness in environments are estimated.

2 System Architecture

2.1 Hardware

The proposed system consists of multiple RGB-D cameras put in the environment and a PC. As RGB-D cameras, we adopt *Kinects*, manufactured by Microsoft Corporation. We assume that the system knows the number of cameras

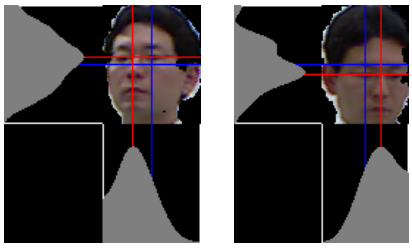


Fig. 2. Examples of Head Pose Estimation

N, their positions and poses described in the world coordinate system. The cameras are put so that their optical axes will be parallel to the x-z plane.

2.2 Image and Depth Capture

In the first module, color images and corresponding depth information are captured from synchronized multiple RGB-D cameras. After the capture, three-dimensional Cartesian camera coordinates, which origin is placed in the position of the corresponding camera, are calculated for each pixel in each captured image using the measured depth z.

2.3 Head Pose Estimation

In the second module, the user's head pose is estimated from a set of images and depth information captured in the first module.

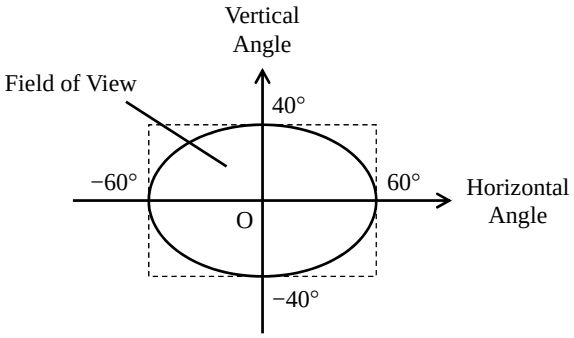
First, the position of the user's head is detected. We adopt the particle filter [4] for tracking the user's head. Then, based on the specified head position, the user's head pose $(\theta_{\text{pitch}}, \theta_{\text{yaw}})$ is estimated. In this paper, we assume that the roll angle $\theta_{\text{roll}} = 0$. They are estimated by using vertical integral projections v_x and v_y of horizontal edge components in the extracted head image are calculated as follows [5]:

$$v_x(x) = \sum_y |e_h(x, y)|, \quad v_y(y) = \sum_x |e_h(x, y)|, \tag{1}$$

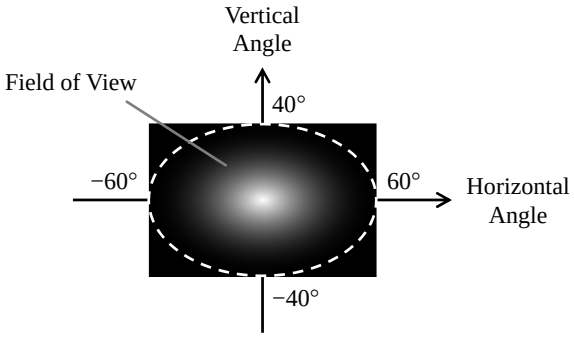
where $e_h(x, y)$ denotes the horizontal edge component at the pixel (x, y). Then, histograms of v_x and v_y are smoothed ten times to extract rough features of the face. The positions of the maximum values of v_x, v_y are calculated as

$$x_{\max} = \arg\max_x v_x(x), \quad y_{\max} = \arg\max_y v_y(y). \tag{2}$$

Fig. 2 shows examples of the histograms of v_x and v_y. Blue lines denote the center of the head image and red lines denote positions of $x_{\max}$ and $y_{\max}$. We can see from these examples that the point $(x_{\max}, y_{\max})$ corresponds to the center of the user's face.



(a) Horizontal and Vertical Angles of View



(b) Distribution of Relative Acuity of Visual Perception

Fig. 3. Model of Field of View

Although a lot of methods have been proposed for the head pose estimation [6], we give priority to achievement of estimation at high frame rate because this advantage enables the system to recognize even the user's glancing. Therefore, we adopt a simple method for the head pose estimation.

2.4 View Estimation

Next, based on the specified head pose, the user's field of view in pixels from three-dimensional point cloud data [3].

In this paper, we simply assume that the horizontal and vertical angles of view of human's eye are 120° and 80° respectively, as shown in Fig. 3 (a). The origin of the coordinates in Fig. 3 (a) denotes the user's head orientation.

Humans can see objects clearly near the center of view, but blurredly near the border. So we also assume that relative acuity of visual perception in the field of view is approximated as shown in Fig. 3 (b). The brighter color in Fig. 3 (b) denotes the higher relative acuity and the origin of the coordinates has the highest acuity 1.

The user's field of view in an environment is estimated according to the following procedure. This procedure is performed for each captured image.

1. The camera coordinate system which origin is placed in the position of the camera is transformed into the viewing coordinate system such that its origin is placed in the center of gravity of the user's head and its z-axis corresponds to the user's head orientation, using the specified head pose parameters θ_{pitch} and θ_{yaw}.
2. The Cartesian coordinate system (x, y, z) is transformed into the polar coordinate system (ρ, θ, ϕ). In this coordinate system, the field of view shown in Fig. 3 corresponds to $0^\circ \leq \theta \leq T_\theta(\phi)$ and $0^\circ \leq \phi \leq 360^\circ$, where

$$T_\theta(\phi) = (2 + |\cos\phi|) \times 20^\circ. \tag{3}$$

   Furthermore, the relative acuity a shown in Fig. 3 (b) is formulated as

$$a(\theta, \phi) = \left(\frac{\theta - T_\theta(\phi)}{T_\theta(\phi)} \right)^2. \tag{4}$$

3. The θ-ϕ plane ($0^\circ \leq \theta \leq T_\theta(\phi)$, $0^\circ \leq \phi \leq 360^\circ$) is divided into bins at 0.5° intervals. For each bin, the nearest pixel to the user's eye, which has the smallest value of ρ, is decided. The set of these nearest pixels is defined as the user's *field of view*, that is, the regions that the user can observe from his/her position. Conversely, pixels out of the field of view are defined as the user's *blind regions*.

Pixels in the field of view are classified into the following three categories according to the positional relationship between the user's line of sight and the environment [3].

(A) Pixels that the corresponding camera and the user observe the same side.
(B) Pixels that the corresponding camera and the user observe the opposite side to each other.
(C) Pixels that the corresponding camera and the user *will* observe the same side, but it is also possible that the user cannot observe it by occlusion due to an obstacle in the camera's blind region.

2.5 Attention/Awareness Estimation

Finally, based on the estimated field of view, the maps of the user's attention and awareness in the environment on the captured images are estimated as follows:

1. The visual acuity map M_{acuity} at time step t is defined as

$$M_{\text{acuity}}(i, j, t) = a(\theta(i, j), \phi(i, j), t), \tag{5}$$

   where (i, j) denotes a pixel in the map, $\theta(i, j)$ and $\phi(i, j)$ denote the parameters in the polar coordinate system for the point which corresponds to (i, j). M_{acuity} denotes a distribution of the user's visual acuity among his/her field of view. That is, it represents how well the user sees each pixel at the moment.

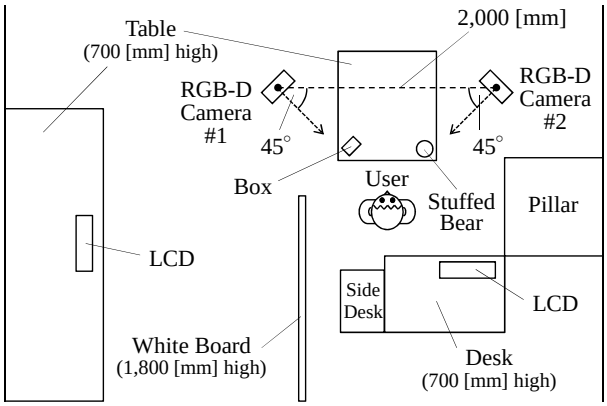


Fig. 4. Experimental Environment

2. The user's attention map M_{att} at time step t is defined as

$$M_{\text{att}}(i,j,t) = \beta \cdot M_{\text{acuity}}(i,j,t) + (1-\beta) \cdot M_{\text{att}}(i,j,t-1), \tag{6}$$

where β denotes a mixture rate. In this paper, we set $\beta = 0.1$. M_{att} is a temporal accumulation of M_{acuity}, and it represents how well and how continuously the user sees each pixel for a period. So we define this map as the user's *attention*.

3. The user's awareness map M_{awr} at time step t is defined as

$$M_{\text{awr}}(i,j,t) = \max\left\{M_{\text{acuity}}(i,j,t), M_{\text{awr}}(i,j,t-1)\right\} \tag{7}$$

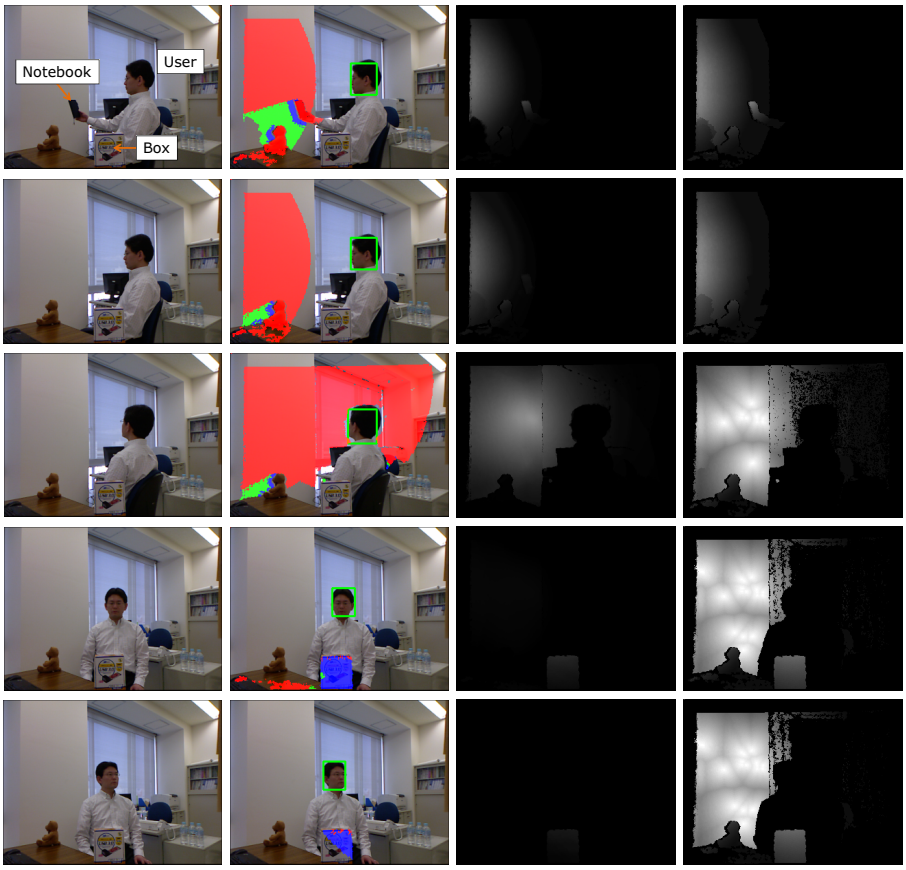
If depth z for the pixel (i,j) makes some change (we set a threshold at 50 [mm] in this paper) from the previous time step, then $M_{\text{awr}}(i,j,t)$ is reset at 0. M_{awr} denotes a map of maximum values of visual acuity until then. It represents pixels which has been included in the user's view. Therefore, it is probable that the user is aware of the objects and information corresponding to those pixels. So we define this map as the user's *awareness*.

3 Experiment

We carry out an experiment to confirm effectiveness of the proposed system.

3.1 Experimental settings

Fig. 4 shows the experimental environment. The two RGB-D cameras are set in front of a user, who sits on a swivel chair, at interval of 2,000 [mm]. These cameras are set at a height of 1,050 [mm]. There are a white board and a pillar on the left and right side of the user respectively. There are also a box and a



(a) Original Image (b) Field of View (c) Attention Map (d) Awareness Map

Fig. 5. Experimental Results (Camera #1)

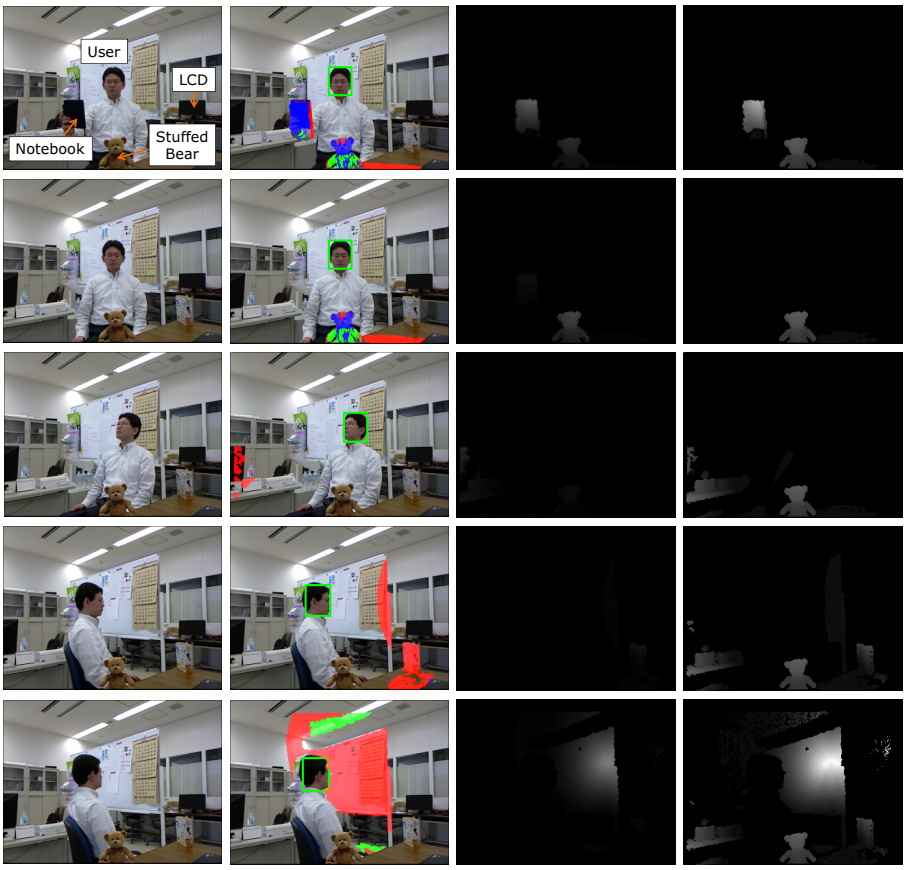
stuffed bear on the table in front of the user, an LCD on the desk behind him, and another LCD on the left table.

The size of both of image and depth information captured from the RGB-D camera is 640×480 pixels. The estimation of the maps of attention and awareness is performed to the resized image of 320×240 pixels to reduce computational time. The frame rate of the system is about 5–7 [fps] without any special optimization on a normal notebook PC (Intel Core i7, 2.60 [GHz]). The latency time from data capture to output of the estimated field of view is about 200 [msec].

3.2 Results

Fig. 5 and 6 show examples of the experimental results.

Fig. 5 (a) and 6 (a) show the original images of the same scene captured from the camera #1 and #2 respectively. These figures are arranged in time order.



(a) Original Image (b) Field of View (c) Attention Map (d) Awareness Map

Fig. 6. Experimental Results (Camera #2)

In the top figures, the user held his notebook and looked to it. Then he lowered the notebook and looked down to the stuffed bear (in the second rows). Next, in the third rows, he turned to his right and saw the pillar. Then he turned to his left and looks to the box (in the fourth rows). And finally, he further turned to his left and looked to the right side of the white board (in the bottom).

Fig. 5 (b) and 6 (b) show the estimated fields of view in the images captured from the camera #1 and #2 respectively. Red, blue and green regions in these figures correspond to Category (A), (B) and (C) respectively. Fig. 5 (c) and 6 (c) show the estimated attention maps, 5 (d) and 6 (d) show the estimated awareness maps on the images captured from the camera #1 and #2 respectively. (In Fig. 5 and 6, the field of view and the maps of attention and awareness are not estimated in the upper and left regions because the RGB-D cameras cannot capture the corresponding depth data.)

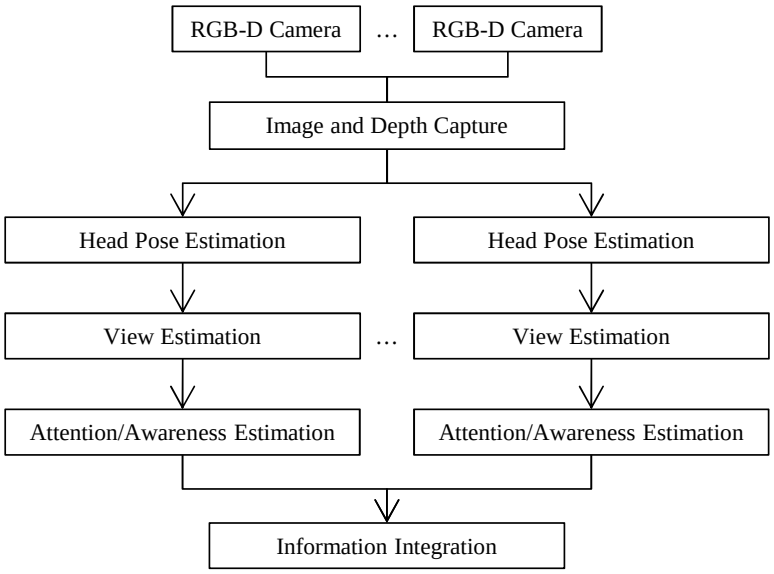


Fig. 7. Flow of Estimation for Multiple Users

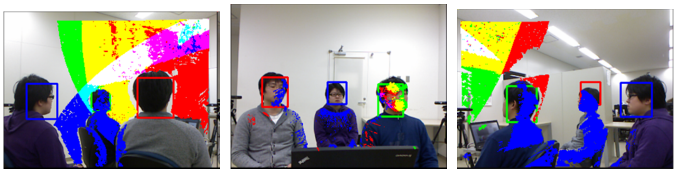
We can see from Fig. 5 and 6 that the proposed system could estimate the maps of the user's attention and awareness well. For example, in the top figures in Fig. 6 (c) and (d), the pixels which corresponds to the notebook and the stuffed bear near it are colored brighter gray because they must be seen clearly by the user. In the second rows, the notebook are removed from both maps because it moved and went out of his view. Furthermore, in the fourth and bottom figures, the LCD put on the table is not colored even if the user tuned to his left because it was occluded by the white board and out of his view. Therefore, the system could recognize that the user was not aware of the LCD. We can see from the attention maps (Fig. 5 (c) and 6 (c)) that the system can recognize the regions to where the user looks, that is, pays attention at the moment well. We can also see from the awareness maps (Fig. 5 (d) and 6 (d)) that the system can store pixels which the user has seen in the environment well. These stored pixels are expected to correspond to the objects of which the user will be aware.

These experimental results agree well with the user's subjective evaluation.

4 Extension to Estimation for Multiple Users

The proposed method can be extended to the estimation of joint attention among multiple users. Fig. 7 shows the extended processing flow. The processes of the head pose estimation and the view/attention/awareness estimation are parallelized into threads for multiple users. The results of estimation are finally integrated.

Fig. 8 shows an example of the view estimation for multiple users. In this



(a) Camera #1 (b) Camera #2 (c) Camera #3

Fig. 8. Example of View Estimation for Multiple Users

example, the system has three cameras and estimate three users' field of view simultaneously. In Fig. 8, the colors of red, green and blue are assigned to fields of view for each user. Furthermore, the overlapped regions in their view are painted the mixed color of their assigned colors. The system can obtain the regions where all users can see (i.e. their *joint attention*), ones where the only one user can see while the others cannot, and so on.

5 Conclusions

In this paper, we propose a system for estimating maps of a user's attention and awareness in occlusion-rich environments based on the view estimation system using multiple RGB-D cameras. The proposed system can estimate what the user sees, what he/she pays attention to, and what he/she is aware of in environments in pixels of captured images. Experimental results in a real environments show effectiveness of the proposed system. Furthermore, we discuss an extension of the proposed system to estimation for multiple users.

The maps of attention and awareness estimated by the proposed system are still only rough estimation. As a future task, we plan to investigate the accuracy of the proposed system through the subjective evaluation. It is expected that we can obtain the user's attention and awareness by subjective questionnaire. Furthermore, we also plan to introduce the saliency map to the system in order to estimate more detailed information on users' attention and awareness in environments.

References

1. Maeda, Y., Katagami, D., eds.: Special Issue on Human Symbiotic System. Journal of Advanced Computational Intelligence and Intelligent Informatics, vol. 14, no. 7 (2010)
2. Imai, J., Kaneko, M.: Human Interactions with a Robot that Recognizes Differences between Fields of View. Kansei Engineering International Journal, vol. 10, no. 1, pp. 59–68 (2010)
3. Imai, J., Tamegai, M.: Three-dimensional Estimation of User's Field of View Using Multiple RGB-D Cameras. In: Proc. SICE Annual Conference 2013, pp. 2347-2352 (2013)

4. Doucet, A., de Freitas, N., Gordon, N., eds.: Sequential Monte Carlo Methods in Practice, Springer, New York (2001)
5. Brunelli, R., Poggio, T.: Face recognition: Features versus templates, IEEE Trans. Pattern Anal. Mach. Intell., vol. 15, no. 10, pp. 1042–1052 (1993)
6. Murphy-Chutorian, E., Trivedi, M.M.: Robust Head Pose Estimation in Computer Vision: A Survey, IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 4, pp. 607–626 (2009)