Gaze Estimation Driven Solution for Interacting Children with ASD

Haibin Cai[1], *Xiaolong Zhou*[1,3], *Hui Yu*[1], *Honghai Liu*[1,2]
[1]School of Computing, University of Portsmouth, UK
[2]State Key Laboratory of Mechanical System and Vibration, Shanghai Jiao Tong University, China.
[3]College of Computer Science, Zhejiang University of Technology, China

Abstract—**This paper investigates gaze estimation solutions for interacting children with Autism Spectrum Disorders (ASD). Previous research shows that satisfactory accuracy of gaze estimation can be achieved in constrained settings. However, most of the existing methods can not deal with large head movement (LHM) that frequently happens when interacting with children with ASD scenarios. We propose a gaze estimation method aiming at dealing with large head movement and achieving real time performance. An intervention table equipped with multiple sensors is designed to capture images with LHM. Firstly, reliable facial features and head poses are tracked using supervised decent method. Secondly, a convolution based integer-differential eye localization approach is used to locate the eye center efficiently and accurately. Thirdly, a rotation invariant gaze estimation model is built based on the located facial features, eye center, head pose and the depth data captured from Kinect. Finally, a multi-sensor fusion strategy is proposed to adaptively select the optimal camera to estimate the gaze as well as to fuse the depth information of the Kinect with the web camera. Experimental results showed that the gaze estimation method can achieve acceptable accuracy even in LHM situation and could potentially be applied in therapy for children with ASD.**

Index Terms—**Gaze estimation, eye tracking, eye location, child, ASD.**

I. Introduction

As one of the most salient features of a child's face, eyes and gaze play an important role in expressing a human focus of attentions, cognitive processes, emotional states and memory. Knowing the gaze of the child and the corresponding visual focus of attention can provide useful information to the robot when interacting with the child. Thus it could potentially be applied in therapy for children with Autism Spectrum Disorders (ASD). In this paper, we explore gaze estimation solutions for interacting children with ASD.

Many approaches have been proposed to estimate gaze direction or point of regard. Video-based gaze estimation can be mainly categorized into two categories [1], namely appearance based methods and feature based methods. Appearance based methods regard gaze estimation problem as directly mapping the eye image contents to the point of regard. Sugano et al. [2] propose a method to map the eye images to the gaze points of a person watching a video clip by using Gaussian process regression and saliency of the video frames. Williams et al. [3] build a sparse and semi-supervised Gaussian process regression model which uses a mixture of different image features to improve accuracy and consistency of gaze estimation. Lu et al. [4] propose an adaptive linear regression method to select an optimal set of sparest training samples for gaze estimation, thus less training samples are required. Sugano et al. [5] propose a multi-camera system to reconstruct the 3D shape of eye regions and then use a random regression forest for person and head pose independent gaze estimation. Funese et al. [6] propose to create a face mesh by fitting a 3D Morphable Model to Kinect depth data and then crop eye images to frontal looking to obtain pose rectified eye images.

On the other hand, feature based methods rely on the local face features such as eye contours, eye corners, pupil center and eye glints. Zhu et al. [7] improve the classical pupil center corneal reflection technique by using a head mapping function to compensate head movement. Lu et al. [8] use a combination of improved pixel-pattern-based texture feature and local binary patter texture feature and support vector regressor to track the gaze direction under allowable head movement. Valenti et al. [9] propose a hybrid scheme to combine the head pose and eye location information for gaze estimation. Xiong et al. [10] propose a supervised decent method to track the facial landmarks whose 3D coordinates are provided via RGBD camera, then an eye gaze model is used for gaze estimation.

However, the majority of the existing methods can not handle large head movement (LHM) of the children with ASD. For example, in the therapy scenarios, the child is asked to looking at different objects on the table or looking at the head of the robot. The angle of his head pose is sometimes more than 90 ° and his face is sometimes not present in the camera. In this paper, we design an intervention table equipped with multiple sensors to deal with deal with LHM while achieve real time performance. The child's facial features and head poses are tracked frame to frame. The eye center is localized by using the convolution based integer-differential eye localization method based on our former work [11]. A rotation invariant gaze estimation method is proposed by combining the located facial features, eye center, head pose and the depth data captured from Kinect. A multi-sensor fusion strategy is further developed to adaptively select the optimal camera to estimate the gaze as well as to fuse the depth information of the Kinect with the web camera. By using multiple cameras, the proposed system is able to deal with LHM even when the child's face is out of the field of view of one of the cameras.

The rest of this paper is organized as follows. A brief introduction of the designed intervention table is presented in Section II. Section III describes the detail of the facial feature extraction, head pose tracking, eye center localization method, gaze estimation and the multi-sensor fusion strategy. Section IV presents the experimental results of eye center localization and gaze estimation. Finally the paper is concluded with discussions in Section V.

II. Proposed System

We design an interaction table with multi-sensor setup to deal with the LHM. The boosted cascade face detector is employed with default parameters in order to obtain the approximate location of the face [12]. Facial features and heap pose estimation have been obtained through applying a supervised decent method (SDM) [13]. The eye location method is based on our former convolution based integer-differential eye localization approach [11]. We proposed a new rotation invariant gaze estimation model by combining the located facial features, eye center, head pose and the depth data captured from Kinect. Finally we present a multi-sensor fusion strategy which can adaptively selects the optimal camera to estimate the gaze as well as to fuse the depth information of the Kinect with a web camera.

A. Hardware design

The system is designed for the therapy of children with ASD. In order to deal with LHM, a multi-camera solution is necessary and the cameras must be placed in a large distance with big difference of view angels to capture the face. Thus in some image data, the eye image only be captured by one camera, this makes it hard to reconstruct the 3D shape of eye regions by simply using the multi-view data. To solve this problem, the Kinect is used to capture the 3D positions of objects and both heads of child and robot. The cameras are used to determine the direction of the gaze. As can be seen from Fig.1, the system consists of an intervention table, two Kinects and three cameras. The first Kinect installed on the middle of the bar is used to capture the child's head position. The second Kinect installed at the middle of the top bar is used to capture the robot's head and objects. The cameras have a resolution of 1028 × 960 pixels and capturing speed of 25 fps, which are installed as shown in Fig.1. The visual focus of attention of the child is determined by combining the objects position and gaze of the child. Thus the whole system can provide useful information to the robot to interact with the child.

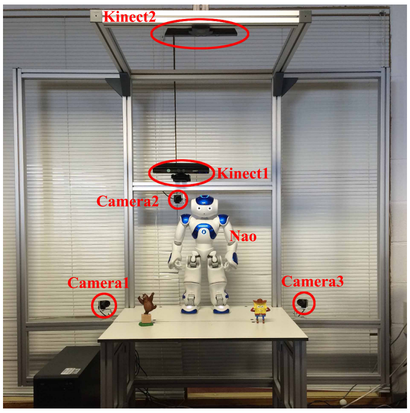
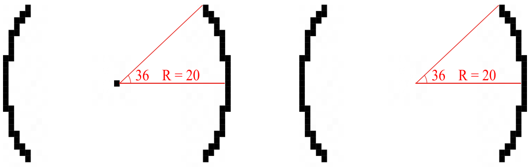


Fig. 1. Hardware setup of our system.

B. Eye center localization

The convolution based integer-differential eye center localization method [11] is used for eye center localization. The eye center is located by using the drastic intensity changes at the boundary of iris and sclera and the gray scale of eye center as well. The eye center localization method utilizes all the pixels along the circular by convoluting different sizes of circle masks with the eye region image. The gray scale of the eye center is considered by designing the convolution masks with a weight in the center point. The size of the mask is $2r + 1$ where r stands for the radius of circle. The pixels along the circular are assigned with a normalized value. Fig.2 shows the reverse value of the kernel, in which the black squares represent pixels with a weight and the white part represent pixels with a value of zero. In order to cope with the occlusion of eye lids, the upper and lower part of the masks are not assigned and the angle of the arc is set to 36 degree.



(a) (b)

Fig. 2. Two kinds of designed masks (a) The mask with center weight at a radius size of 20 pixels. (b) The mask without center weight at a radius size of 20 pixels.

The convolution based integer-differential eye localization method calculates a ratio derivative between neighbor curve magnitudes. The ratio derivative is formulated as follows.

$$\begin{cases} I_r = K_r * I \\ I'_{r+1} = K'_{r+1} * I \\ D_r = \frac{I'_{r+1}}{I_r} \\ argmax_{(r,x,y)}(D_r) \\ r\epsilon[r_{min}, r_{max}] \end{cases} \tag{1}$$

where K_r is the mask with a center weight and r stands for the radius of the circle inside the mask. The mask without a center weight is represented as K'_{r+1} whose radius is $r+1$. I_r and I'_{r+1} are the results of convolution of the different masks with eye image I. D_r stands for the ratio derivative calculated by the division of the convolution result image. r_{min} and r_{max} are set according to the size of eye image representing the minimum and maximum of the radius r. The weights of the points around the circular arcs are of equal value and normalized to 1, and the weight of the center point is set to a valid value. The final location of the eye center is determined by searching the maximum of different radius of D_r.

C. Gaze estimation

We propose a rotation invariant gaze estimation method by using the located eye centers, eye corners and eyelids. The gaze direction can be divided into two directions refer to the head pose, namely, horizontal direction and vertical direction. Fig.3 shows an illustrated example of the eye image. The ellipse and circle in the image represent the eye and the iris, respectively. The center of two eye corners is assumed to be the eye center (x_c, y_c). Because of the low resolution in the eye area, the iris center is assumed to be the pupil center (x_p, y_p).

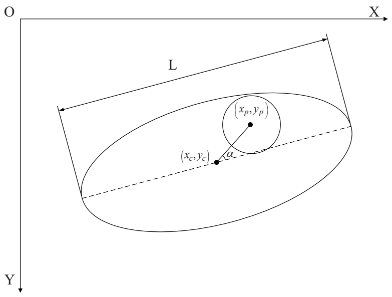


Fig. 3. An illustrated example of the eye image.

The gaze direction in the head coordinate system can be estimated by the following equations.

$$\begin{cases} \vartheta = tan^{-1}(a_0 * (D_{cp} * \frac{cos\alpha}{L} - a_1)) \\ \delta = tan^{-1}(b_0 * (D_{cp} * \frac{sin\alpha}{L} - b_1)) \\ D_{cp} = \sqrt{(x_p - x_c)^2 + (y_p - y_c)^2} \end{cases} \quad (2)$$

where ϑ, δ stands for horizontal and vertical axial components of the gaze direction respectively. L is the Euclidean distance of the two eye corners. D_{cp} is the Euclidean distance of the eye center and pupil center. α is the intersection angle of the eye-pupil line and corner-corner line. a_0, a_1, b_0, b_1 are parameters that can be determined in the calibration stage.

In order to determine where the child is looking at, we also need to know the 3D start point of the gaze direction which is calculated through the following equation.

$$\frac{u_s - u_0}{X_s} = \frac{v_s - v_0}{Y_s} = \frac{f}{Z_s} \quad (3)$$

where (X_s, Y_s, Z_s) is the 3D position of the start point in the relative camera coordinate system. (u_s, v_s) is the located facial feature point on the bridge of the nose in the captured image. f is the focus length. Firstly we find the 3D head center position and transform it to the relative camera coordinate system. The Z_s is considered to be equal to the depth of the head center position in the relative camera coordinate system. Then, the 3D coordinate of the start point can be solved using the equation 3.

D. Multi-sensor fusion strategy

We propose a multi-sensor fusion strategy to adaptively select the optimal camera to estimate the gaze as well as to fuse the depth information of the Kinect with the web camera. All the image data is synchronously captured and the whole system can work in real time under this strategy. To this end, we divide the strategy into two stages, namely detection stage and tracking stage. The detection stage is used to select the optimal camera for gaze estimation and also is prime for the tracking stage. Then the selected camera operates in the tracking stage until loses the face of the child.

Fig.4a and Fig.4b shows the flowchart of the detection and tracking stage respectively. The primary step is to calibrate the three cameras and two Kinects. The details of calibration step can be found in [14]. The image data is captured synchronously by using muti-thread programming. Each sensor is belong to one thread and one more thread is used to control the synchronous starting of the other thread. The first Kinect is used to detect the 3D location of child head whose center is then used as the start point of the gaze. The second Kinect is for the use of therapy for child. It detects the locations of either the toys on the desktop or the head of the robot. By combining the gaze direction and object positions, it is possible to estimation the visual focus of child's attention and thus to determine whether the child is looking at the object.

In the detection stage, the face detection is performed on each camera frame. The detection of facial features and head pose are processed only if the face is detected. The camera with the best view of the face is selected for the gaze estimation in the camera selection section. If there is no face captured by any of the three cameras, the system will start to capture new frames synchronously for all the sensors. Once the detection stage is finished, the system enter into tracking stage which operates until the head pose overstep the limited value of that camera.

III. Evaluation

In the experiments, the boosted cascade face detector proposed by Viola and Jones [12] as default parameters is used to locate the face position in the image. Then the eye center localization method is used in the rough eye regions extracted through anthropometric relations with the face. For facial feature detections and head pose estimation, we employ the supervised decent method (SDM) [13]. We evaluate the eye center localization method on public available database and

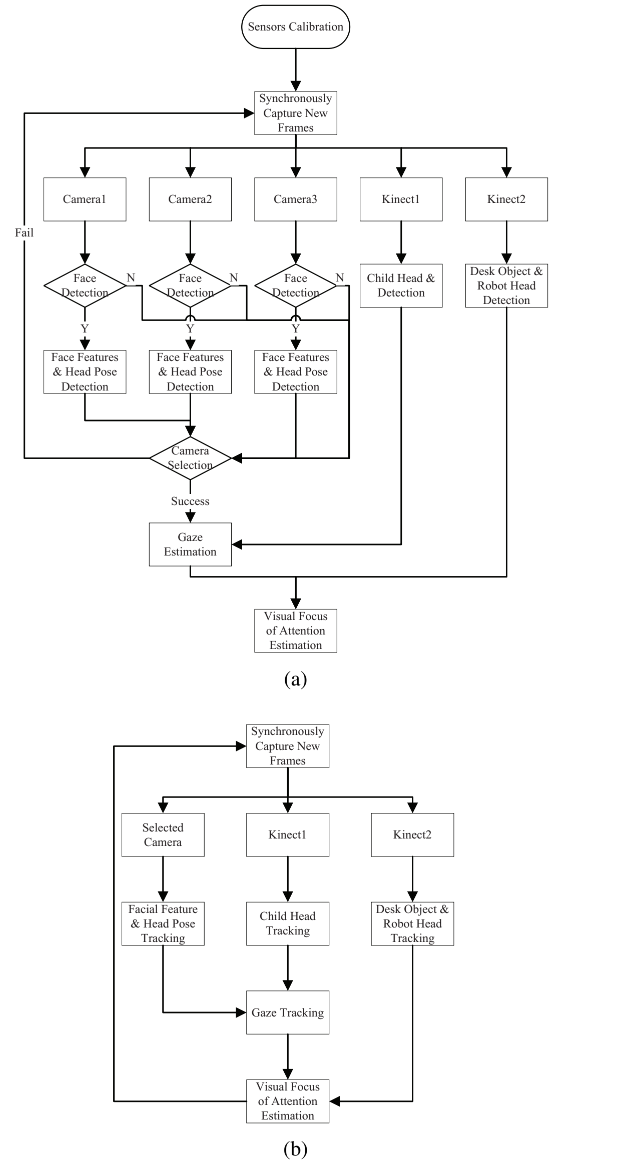

Sensors Calibration
Synchronously Capture New Frames
Fail
Camera1
Camera2
Camera3
Kinect1
Kinect2
Face Detection
N
Face Detection
N
Face Detection
N
Child Head & Detection
Desk Object & Robot Head Detection
Y
Y
Y
Face Features & Head Pose Detection
Face Features & Head Pose Detection
Face Features & Head Pose Detection
Camera Selection
Success
Gaze Estimation
Visual Focus of Attention Estimation
(a)
Synchronously Capture New Frames
Selected Camera
Kinect1
Kinect2
Facial Feature & Head Pose Tracking
Child Head Tracking
Desk Object & Robot Head Tracking
Gaze Tracking
Visual Focus of Attention Estimation
(b)

Fig. 4. The flowchart of detection and tracking stage. (a) Detection stage. (b) Tracking stage.

conduct two experiments to test the gaze estimation method. To protect the privacy of the children with asd, we conduct the gaze estimation experiments on our researchers instead of the children.

A. Evaluation of eye center localization

The proposed eye center localization method is validated on the BioID face database [15] which consists of 1521 grayscale images of 23 different people with a resolution of 384 × 288 pixels. The images in the database are head and shoulder frontal view images with a large variety of illumination, background, scale and pose. Some people in the database are wearing glasses. In some images the eyes are closed or completely hidden by reflections on the glasses. Because of these issues, the BioID database is considered to be challenging and realistic.

The accuracy measure of eye location is calculated in normalized error which records the maximum error of both eye points. The measure was introduced by Jesorsky et al. [16] and is defined as follows:

$$e = \frac{\max(d_l, d_r)}{d} \tag{4}$$

where d_l and d_r are the Euclidean distances between the detected left and right eye centers and the ones in the ground truth and d is the Euclidean distance between the left and right eyes in the ground truth. Herein, a relative error of $e \leq 0.25$ equals a distance of half an eye width, $e \leq 0.1$ means the diameter of iris and $e \leq 0.05$ corresponds to the length of pupil. Table I shows the comparison of accuracy of start-of-the-art method tested on the BioID database.

TABLE I
COMPARISON OF EYE LOCATION ACCURACY OF STATE-OF-THE-ART METHODS TESTED ON THE BIOID DATABASE.

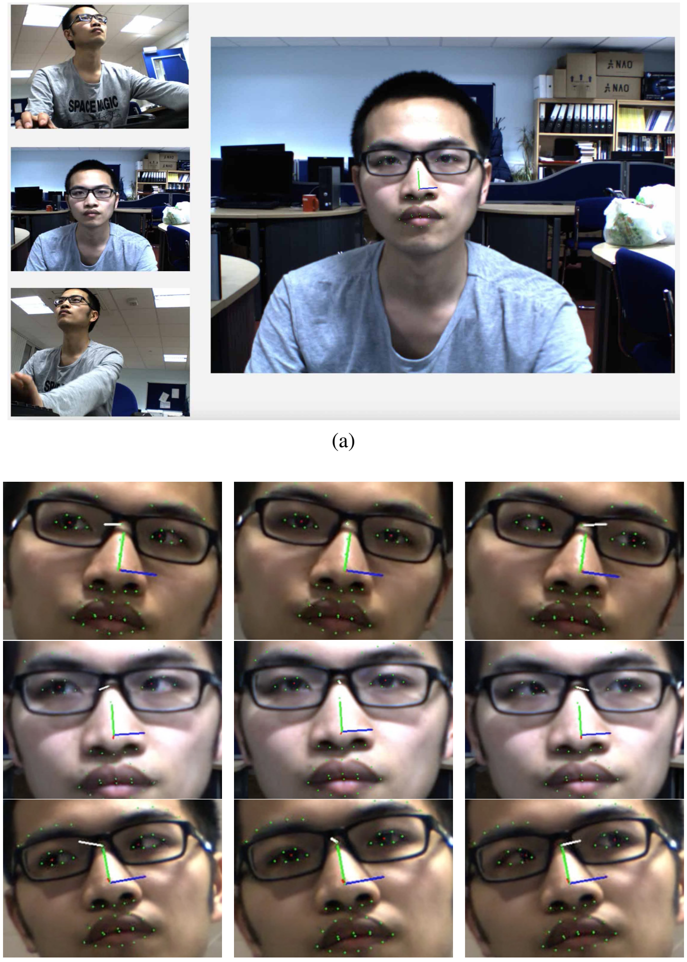
Method	$e \leq 0.05$	$e \leq 0.10$	$e \leq 0.25$
Hamouz et al.[17]	59.0%	77.0%	93.0%
Timm et al.[18]	82.5%	93.4%	98.0%
Valenti et al.[19]	86.1%	91.7%	97.9%
Markus et al.[20]	89.9%	97.1%	99.7%
Proposed method	86.8%	96.6%	99.9%

B. Evaluation of gaze estimation

This section reports two experiments on gaze direction: one is gaze estimation experiment from the whole system and the other is through a single web camera. Fig.5a shows the captured frames of three cameras and selected frame for gaze estimation on the interaction table. Fig.5b shows the gaze direction of the captured faces of three cameras. The first row, second and third row images correspond to the first, second and third cameras respectively. The total running time of the whole system is around 19fps.

In order to fully test the gaze estimate accuracy, another gaze estimation experiment from a single web camera and screen is also conducted. The hardware configuration is shown in Fig.6a. A web camera with resolution of 1920 × 1080 is attached to the 24-inch LCD monitor. The distance between the experimenter and screen is about 65cm. During the calibration stage, the camera collects 9 images of the subject who gazes at 9 different buttons on the screen. After applying facial features, eye center and head pose detection methods, the parameters (a_0, a_1, b_0, b_1) in the equation 2 can be solved using least square method.

Fig. 6b and Fig. 6c shows the gaze estimation results with frontal and nonfrontal head pose respectively. Table II shows the gaze estimation accuracy under different head poses. The gaze estimation results are obtained by looking at random points on the screen. Those points that can not been seen when the head pose is too large are abandoned. For example


(a)

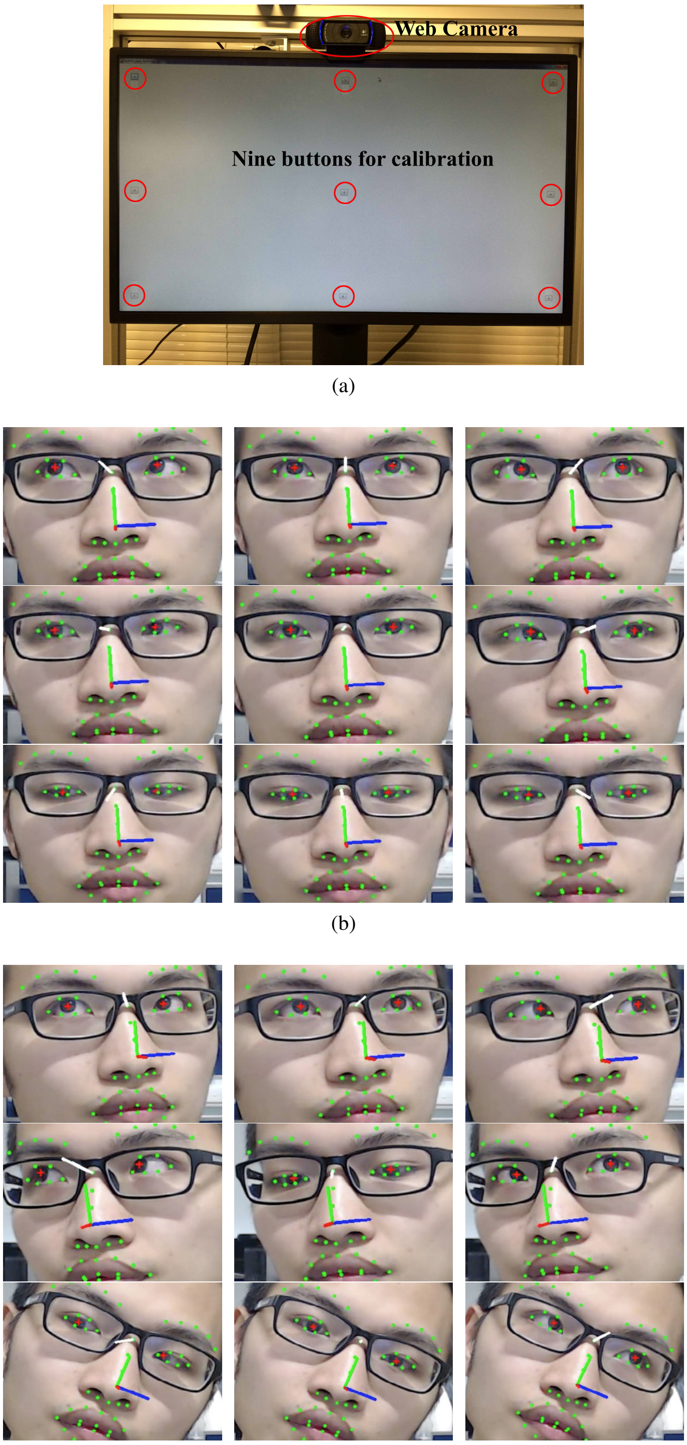
(b)

Fig. 5. Gaze estimation result on the intervention table. (a) The captured frames of three cameras and selected frame for gaze estimation. (b) The face images captured by the first, second and third cameras.

the right top points of the screen can not be seen when the head pose is at $yaw = 30^\circ$. The mean error of the visible points is chosen to be the final gaze estimation error. The average accuracy is about 6.0° horizontally and 9.4° vertically.

TABLE II
GAZE ESTIMATION ACCURACY UNDER DIFFERENT HEAD POSES

Head pose yaw	Horizontal	Vertical
0°	4.21°	5.90°
15°	5.49°	8.72°
30°	8.28°	13.48°
Average	6.0°	9.4°



(a)

(b)

(c)

Fig. 6. Gaze estimation on a web camera. (a) The setup of the web camera based gaze estimation system, the nine buttons is used for calibration. (b) Gaze estimation result on a web camera with frontal head. (c) Gaze estimation result on a web camera with non frontal head

IV. Conclusion

This paper explored gaze estimation solutions for Interacting children with ASD. Compared with the conventional gaze estimation methods, the proposed method can efficiently estimate the gaze of child with tolerance of LHM by designing an intervention table equipped with multiple sensors. We presented a multi-sensor fusion strategy to adaptively select the optimal camera to estimate the gaze as well as to fuse the depth information of the Kinect with the web camera. We employed a convolution based integer-differential eye localization approach to locate the eye center efficiently and accurately. Moreover, we proposed a new rotation invariant gaze estimation model by combining the located facial features, eye center, head pose and the depth data captured from Kinect. Experimental results show that our gaze estimation method can achieve acceptable accuracy even in LHM and can be applied in therapy for interacting children with ASD.

Our future work will focus on improving the accuracy of eye center localization, gaze estimation and also to interpret information behind eye movement with cognitive, psychological, human-like knowledge in the context of human-robot interaction and human-robot skill transfer [21, 22].

Acknowledgment

This work was supported by EU seventh framework programme under grant agreement No. 611391, Development of Robot-Enhanced Therapy for Children with Autism Spectrum Disorders (DREAM).

References

[1] D. W. Hansen and Q. Ji, "In the eye of the beholder: A survey of models for eyes and gaze," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 3, pp. 478–500, 2010.

[2] Y. Sugano, Y. Matsushita, and Y. Sato, "Appearance-based gaze estimation using visual saliency," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 2, pp. 329–341, 2013.

[3] O. Williams, A. Blake, and R. Cipolla, "Sparse and semi-supervised visual mapping with the sˆ 3gp," *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 230–237, 2006.

[4] F. Lu, Y. Sugano, T. Okabe, and Y. Sato, "Adaptive linear regression for appearance-based gaze estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 10, pp. 2033–2046, 2014.

[5] Y. Sugano, Y. Matsushita, and Y. Sato, "Learning-by-synthesis for appearance-based 3d gaze estimation," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1821–1828, 2014.

[6] K. A. Funes Mora and J.-M. Odobez, "Geometric generative gaze estimation (g3e) for remote rgb-d cameras," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1773–1780, 2014.

[7] Z. Zhu and Q. Ji, "Eye gaze tracking under natural head movements," *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 918–923, 2005.

[8] H. Lu, G. Fang, C. Wang, and Y. Chen, "A novel method for gaze tracking by local pattern model and support vector regressor," *Signal Processing*, vol. 90, no. 4, pp. 1290–1299, 2010.

[9] R. Valenti, N. Sebe, and T. Gevers, "Combining head pose and eye location information for gaze estimation," *IEEE Transactions on Image Processing*, vol. 21, no. 2, pp. 802–815, 2012.

[10] X. Xiong, Q. Cai, Z. Liu, and Z. Zhang, "Eye gaze tracking using an rgbd camera: a comparison with a rgb solution," *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, pp. 1113–1121, 2014.

[11] H. Cai, B. Liu, J. Zhang, S. Chen, and H. Liu, "Visual focus of attention estimation using eye center localization," *Systems Journal, IEEE*, vol. 99, pp. 1–6, 2015.

[12] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[13] X. Xiong and F. De la Torre, "Supervised descent method for solving nonlinear least squares problems in computer vision," *arXiv:1405.0601*, 2014.

[14] S. Zhang, H. Yu, J. Dong, T. Wang, L. Qi, and H. Liu, "Combining kinect and pnp for camera pose estimation," *Proceeding of the 8th International Conference on Human System Interactions*, pp. 357–361, 2015.

[15] The BioID Face Database. [Online]. Available: http://www.bioid.com/downloads/facedb/index.php

[16] O. Jesorsky, K. J. Kirchberg, and R. W. Frischholz, "Robust face detection using the hausdorff distance," *Proceeding of the 3rd International Audio- and Video-based Biometric Person Authentication*, pp. 90–95, 2001.

[17] M. Hamouz, J. Kittler, J.-K. Kamarainen, P. Paalanen, H. Kalviainen, and J. Matas, "Feature-based affine-invariant localization of faces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 9, pp. 1490–1495, 2005.

[18] F. Timm and E. Barth, "Accurate eye centre localisation by means of gradients." *Proceedings of the 6th International Conference on Computer Vision Theory and Applications*, pp. 125–130, 2011.

[19] R. Valenti and T. Gevers, "Accurate eye center location through invariant isocentric patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 9, pp. 1785–1798, 2012.

[20] N. Markuš, M. Frljak, I. S. Pandžić, J. Ahlberg, and R. Forchheimer, "Eye pupil localization with an ensemble of randomized trees," *Pattern Recognition*, vol. 47, no. 2, pp. 578–587, 2014.

[21] H. Liu, "Exploring human hand capabilities into embedded multifingered object manipulation," *IEEE Transactions Industrial Informatics*, vol. 7, no. 3, pp. 389–398, 2011.

[22] Z. Ju and H. Liu, "Human hand motion analysis with multisensory information," *IEEE/ASME Transactions on Mechatronics*, vol. 19, no. 2, pp. 456–466, 2014.