FUSION OF EYE-TRACKING DATA FROM MULTIPLE OBSERVERS FOR INCREASED 3D GAZE TRACKING PRECISION

Marianne Hanhela[1]*, Atanas Boev*[1]*, Atanas Gotchev*[1]*, Miska Hannuksela*[2]

[1]Department of Signal Processing, Tampere University of Technology, Tampere, Finland
[2]Nokia Research Center, P.O. Box 1000, Tampere 33721, Finland

ABSTRACT

In this paper we discuss an approach to extract 3D gaze data information from binocular eye-tracking data. Factors such as tracking noise, tracking precision and observation distance limit the resolution of gaze tracking in three dimensions. We have developed a methodology, which uses a model of the stereoscopic human visual system (HVS) to analyze per-eye gaze data and to convert it into a something we call stereoscopic volume-of-interest (SVOI). We have found that using data from multiple observers increases the tracking precision. We aim to find the link between number of observers and tracking precision. This would allow one to optimize the number of participants involved in 3D gaze-tracking experiment, in order to achieve certain level of 3D gaze tracking precision.

Index Terms**—** *stereoscopic vision, binocular eye-tracking, 3D region of interest*

1. INTRODUCTION

Increased consumption of 3D content has brought the need of detailed understanding of the human gaze behavior in three dimensions. Such knowledge can be used in content creation and region-of-interest based encoding. However, there are two major problems in 3D gaze tracking experiments: precision and number of participants. The tracking precision can be increased in two ways – intelligent handling of the data (e.g. fusion of binocular data and filtering of the outliers) and fusing gaze data from multiple observers.

We have developed a framework for extracting *stereoscopic volumes-of-interest* (SVOIs) from gaze data [1]. The method uses data obtained from multiple observers. However, large number of observers increases the duration and the cost of an experiment, and as well as the computational time for data analysis. In this work we aim uncover the link between number of observers and precision of gaze position estimation, and estimate the number of participants needed for sufficient estimation precision.

2. GAZE TRACKING IN THREE DIMENSIONS

Point-of-gaze (PoG) is the position where both eyes of an observer are fixated. Gaze tracking is the process of continuously measuring the PoG coordinated over time. Currently, the achieved precision in tracking the gaze onto a 2D plane is 0.5° to 1°, or 1.5 cm at viewing depth (VD) equal to 150 cm [2], 0.84 cm at VD = 60 cm [3], 0.73 cm at VD = 84 cm, 0.79 cm at VD = 50 cm [4], 0.66 cm at VD = 62.2 cm [5]. However, when measuring the *gaze depth,* i.e. distance between observer and PoG, the achieved precision is much lower. The accuracy is limited by two factors – the angular precision of the eye-tracker and the fact that for a constant angular error, the estimation error for absolute depth grows quickly with the distance [6].

For still 3D images, the reported gaze depth measurement accuracy is 3.93 cm for VD = 67.5 cm [7] and 2.6 cm at VD = 0.5 m [8]. By using a neural network method, the Euclidean error has been reduced to 2.78 cm with viewing depth of 50 cm [9].

The deviation of this magnitude is acceptable on the large screens and low resolutions, but it is especially problematic if portable 3D displays are used in eye-tracking experiments [10][11][12].

In [6] the authors claim that the best precision can be achieved at observation distance of 120 cm, with precision becoming worse with smaller and larger distances. The angular precision limits the possibility to estimate absolute coordinates of the gaze, e.g. to distinguish which pixel is being observed. On the other hand, the estimation of gaze depth is a function of the observation distance. For a given disparity estimation error, the error in estimating the distance to an object increases exponentially with the observation distance.

3. FUSION OF EYE-TRACKING DATA FROM MULTIPLE OBSERVERS

Our method for deriving SVOIs from eye-tracking data has five steps, as shown in Figure 1. First we collect eye-tracking data from multiple observers and create gaze maps, where each gaze map represents the gaze points of one observer over a given period of time. Knowing the angular size of the fovel vision, we convert the gaze maps to heat maps, where the "heat" of each pixel is proportional to the quantity and duration of gazes over the pixel. The heat maps are projected in 3D space to create stereoscopic volumes of interest. SVOIs from multiple observers are combined into *intersected SVOIs* (iSVOI).

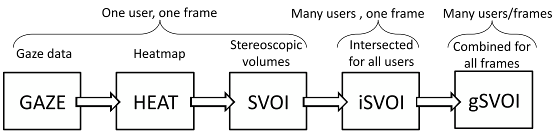


Figure 1. Flowchart of the procedure for building iSVOI

3.1. Gaze maps

One gaze map contains of the locations of the gazes directed to the display during one 15 ms timeframe. The locations are given in absolute coordinates on the screen, where point (0,0) is the lower left corner of the screen, as seen in Figure 2. Separate gaze maps are built for each eye of each observer. In the figure, gaze positions are marked with circles, where the circle area is proportional to the number of gazes onto the corresponding position of the screen.

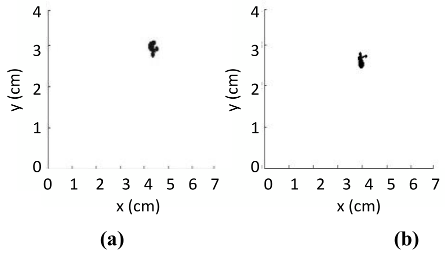


Figure 2. Gaze maps for each eye: a) left eye b) right eye

3.2. Heat maps

In order to build a heat map, each gaze point is used as a center of a heating window described by a two-dimensional Gaussian distribution, with parameters mean m = 1 and standard deviation σ = 0.37 cm. The standard deviation σ is selected so that the area of a circle of 2σ radius corresponds to the area of the sharpest gaze. The heats of all gazes are accumulated into a heat map, as shown in Figure 3.

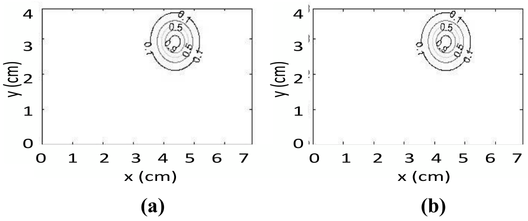


Figure 3. Heat maps for each eye: a) left eye b) right eye

3.3. Stereoscopic Volume-of-Interest (SVOI)

We consider SVOI as the volume formed by intersecting the projections of two heat maps in 3D space. Each heat map can have arbitrary shape. Each heat map is projected towards the retina of the eye, as shown in Figure 3. We estimate the volume of the intersection by scanning procedure. The procedure considers the 3D space as a collection of slices parallel to the screen plane as shown in Figure 4. At each step we derive the shape of the heat map in one slice by shifting and scaling it as a function of the distance between the slice and the display. We assume that each map has its original size on the display level, shrinks to a point on entering the observer's pupil, and grows behind the display.

The projection of each heat map is done towards the corresponding eye, and in each slice, the intersection between the maps is calculated. All calculations are done inside of confined 3D space, defined by the size of the display and the range of apparent object depths generated by the 3D display, as shown in Figure 4. For our experiments the dimensions of the space were 4 cm x 7 cm x 60 cm, stretching 20 cm in front and 40 cm behind the display surface. An example for rendered SVOI can be seen in Figure 5.

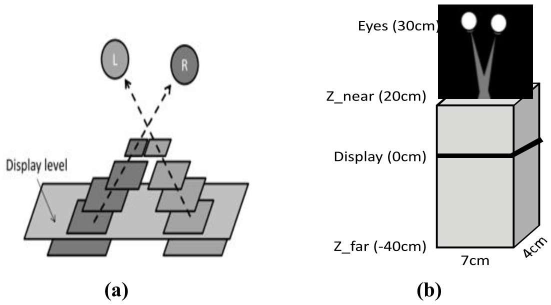


Figure 4. SVOI from complex shape: a) finding the intersecting areas scanning one slice at a time and b) absolute coordinates of the scanned volume

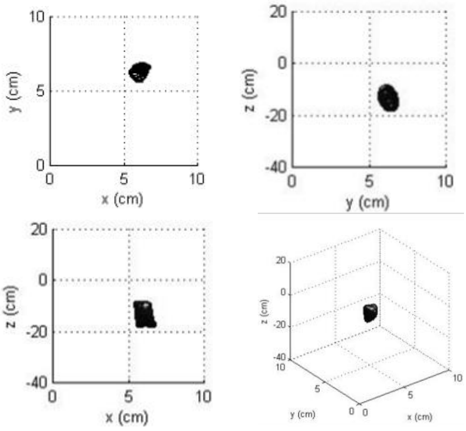


Figure 5. SVOI of one frame: view from top, view from left, view from front, and an orthogonal projection

3.4. Intersected SVOI

The objective of the intersected SVOI (iSVOI) is to find gaze volumes that are shared by many observers. The purpose is to further delineate the gaze volume, since it is likely that the volume-of-interest will lie inside a gaze volume shared by many observers. Intersection the gaze volumes of many observers also removes outliers created by incidental "distracted" gaze of a single observer outside of the volume-of-interest.

Since the amount of registered gaze points has big variance between frames (e.g. between 0 and 50) it is not feasible to mark a point of space as "interesting" by using a fixed threshold value. Instead, we use a dynamic threshold procedure, where we threshold points in the SVOI with a heat greater than certain percentage of the maximum heat value for the SVOI.

We selected the threshold percentage as a value which satisfies two conditions. First, we wanted the difference between estimated position and the ground truth to be low, and secondly, the resulting iSVOI to be a single confined volume in the 3D space. The second condition was assessed by comparing the estimated standard deviation of the iSVOI to the σ as predefined in section 3.2. The sum of the error from the ground truth and error of the standard deviation from 1000 randomly selected frames, with all 13 observers, is represented in Figure 7 as a function of bounding percentage. The minimum seems to be located at approximately 75 %, which is the value we selected for the further study.

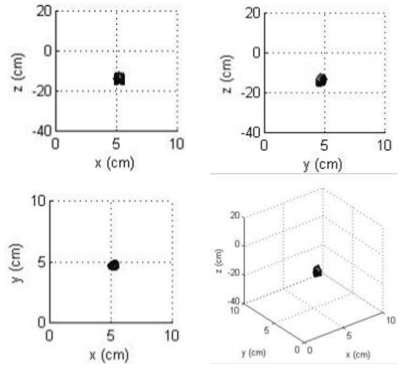


Figure 6. iSVOI of one frame: view from top, view from left, view from front, and an orthogonal projection

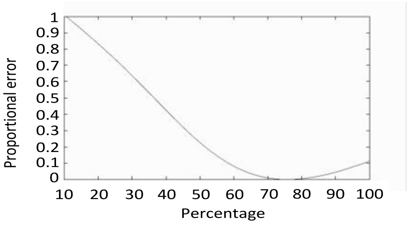


Figure 7. Sum of error from the ground truth and error of the standard deviation in function of the bounding percentage

4. EXPERIMENTAL DATA

4.1. Eye-tracking experiments

For the purpose of our tests we created a number of synthetic stereoscopic movies. The movies were visualized on a portable 3.1" autostereoscopic display created by NEC Technologies. The display has identical resolution in 2D and 3D modes due to its horizontal double-density (HDDP) pixel arrangement as described in [13]. The display input is in side-by-side stereo format. The resolution of each view is 427 x 240 px, and each view covers the entire screen area of 6.9cm x 3.88 cm. The pixel density of the display is 157DPI.

We had 13 participants in the subjective tests. The test group consisted of both professionals and non-professionals between ages 18 and 45 some of which had prior 3D experience. The distance between the observers and the display was 30 cm. The disparity range in the synthetic content was limited between -25 and +25 px, which corresponds to apparent depth of the objects between z_{near} = 28.24 cm and z_{far} = 31.99 cm. The test content featured a single object moving in 3D space. We had calculated the apparent 3D position of the object for each frame, and used it as a ground truth.

4.2. Synthetic content

We have created a number of synthetic stereoscopic movies where we know the ground truth of the depth (Table 1). In our eye-tracking experiments, we presented the video streams listed in Table 1. Each movie contains a single moving white ball with 20 px diameter on a black or textured background as shown in figure X. Movement in x and y directions goes from one edge of the screen to the other, while the ball is always fully visible. The length of each movie is 10 seconds and the frame rate is 30 frames per second.

Filename	Direction	Speed	Initial disp.	Final disp.
s10	z	slow	-25px	0
s12	z	fast	-25px	0
s17	x	slow	0	0
s23	y	sudden	0	0
s37	xyz	slow	0	-25px
s39	xyz	fast	0	-25px
s49	xyz	slow	25px	-25px
s51	xyz	fast	25px	-25px

Table 1. List of the test movies

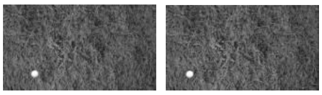


Figure 8. Left and right frames from movie s39 at 2 seconds.

5. ANALYSIS

To find the optimal number of the observers for gaze tracking experiments, we compared iSVOIs resulting from different number of observers to the known ground truth of the test material clips. Before comparison, the means and standard deviations of the iSVOIs were estimated using the assumption of Gaussian distributions. The parameters were computed for x, y and z directions separately for getting knowledge of the behavior of the iSVOI in different dimensions.

In Figure 9 one can see the comparison between the iSVOI data and the ground truth for x, y, and z directions correspondingly. Fat solid lines represent the ground truth and thin solid lines give the estimated mean. The range between -2σ and 2σ, where σ is the estimated standard deviation, is marked with thin dashed lines.

For investigating the optimal number of observers, we selected random groups of 1 to 13 observers for each movie and built iSVOIs from their gaze data. Then we computed average RMSE and the total estimation error from the ground truth for groups with different sizes. As one can see in Figure 10, the estimation error in z direction is significantly larger than the errors in x and y direction. We assume this is due to the small observation distance used in a mobile display setup. Still, the overall estimation error is comparable with the values reported in other studies [2][3][4][5][6][7].

With increasing the number of observers, the error is approximately halved when data of four observers is processed. After that, the error starts to stabilize in x and y directions, and the decreasing of the error in depth also slows down.

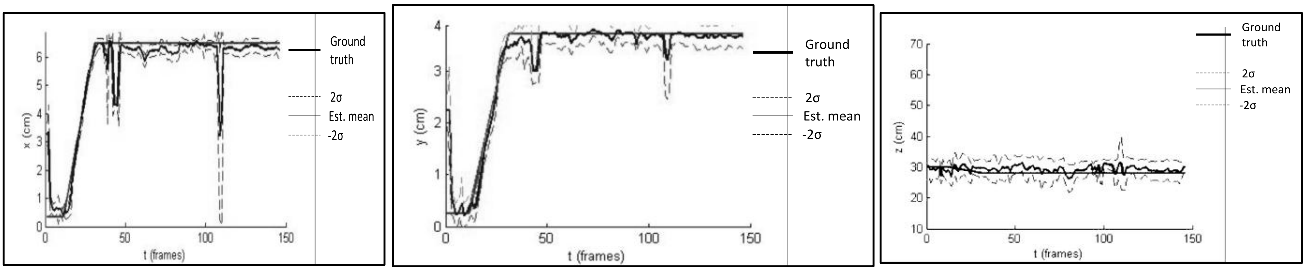
6. RESULTS

As it was shown, the deviation of the observations from the ground truth decreases as the number of the observers grows. However, the error quickly saturates so that adding more observers does not yield to considerably better accuracy.

In x and y direction, results start to saturate as soon as with 4 observers, but as the largest error source is the depth estimation, adding more observers is reasonable if higher accuracy is needed. The lowest average RMSE we got with this set of experiments was 3.55 cm, with all 13 possible observers. This is well comparable with the earlier results. Even though a larger group of observers might further improve the estimation accuracy, we suggest that a group size of 4 to 5 observers is good trade-off between experiment cost and 3D POG estimation precision.

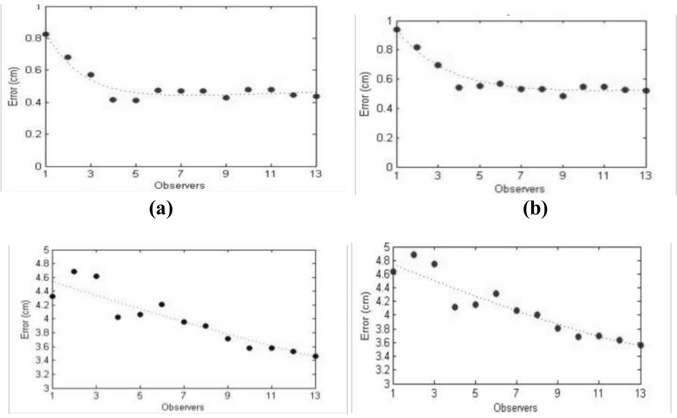
We identify multiple sources of measurement noise which limit the POG estimation precision for a single observer. The measurement error of the eye-tracker has the biggest influence. In our experiments the measurement precision is 0.1º which for the observation distance of 30 cm is equivalent to 4 pixels. Large visual targets and microsaccadic oscillations add a degree of uncertainty about the position of the POG in respect to the object. Visual fatigue produces POG data which falls outside of the targeted region. In x direction, the tendency of having a dominant eye also affected the results as the x coordinate is assumed to be mean of the left and right eye coordinates.

Intersecting the gaze from multiple observers reduces the error in many ways. Number of outliers diminishes, as they are results of single observers blinking, looking away or having anomalies in vision. Also, the smaller errors from microsaccades and the center of object being near to the margin of the sharp gaze volume cancel out as the most probable common volume for many observers lies in the center of the target object.



(a) **(b)** **(c)**

Figure 9. s39 – Ground truth and observed location in a) x direction b) y direction c) z direction



(a) (b)

(c) (d)

Figure 10. Mean squared error (cm) in a) x coordinates b) y coordinates c) z coordinates d) distance from the ground truth

7. CONCLUSION

We have developed a methodology that allows us to obtain precise estimation of the gaze position in 3D. The methodology uses eye-tracking data from multiple observers. In this work, we have discussed the connection between number of observers and gaze-tracking precision, and the minimum amount of observers (4 to 5) needed for sufficient precision. In our experiments, fusing data from multiple observers increases the 3D POG estimation, and even for in a mobile 3D display setup it becomes comparable to the measurements reported for desktop display setups. Precise data about the position and depth of the binocular gaze allow better understanding of the gaze behavior in 3D. Outcomes of such understanding include stereo-video quality estimation and region-of-interest based 3D video compression.

8. REFERENCES

[1] A. Boev, M. Hanhela, A. Gotchev, T. Utriainen, S. Jumisko Pyykkö, ”Parameters of the human 3D gaze while observing portable autostereoscopic display: a model and measurement results”, Multimedia on Mobile Devices, part of Electronic Imaging, Symposium 2012.

[2] Wang, J.-G.; Sung, E.; Ronda Venkateswarlu; , "Eye gaze estimation from a single image of one eye," Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on , vol., no., pp.136-143 vol.1, 13-16 Oct. 2003

[3] Brolly, X.L.C.; Mulligan, J.B.; , "Implicit Calibration of a Remote Gaze Tracker," Computer Vision and Pattern Recognition Workshop, 2004. CVPRW '04. Conference on , vol., no., pp. 134, 27-02 June 2004

[4] Jeongseok Ki; Yong-Moo Kwon; "3D Gaze Estimation and Interaction," 3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video, 2008 , vol., no., pp.373-376, 28-30 May 2008

[5] Sheng-Wen Shih; Jin Liu; , "A novel approach to 3-D gaze tracking using stereo cameras," Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on , vol.34, no.1, pp.234-245, Feb. 2004

[6] Beymer, D.; Flickner, M.; , "Eye gaze tracking using an active stereo head," Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on , vol.2, no., pp. II- 451-8 vol.2, 18-20 June 2003

[7] Mitsugami, I.; Ukita, N.; Kidode, M.; , "Estimation of 3D gazed position using view lines," Image Analysis and Processing, 2003.Proceedings. 12th International Conference on , vol., no., pp. 466- 471, 17-19 Sept. 2003

[8] Hennessey, C.; Lawrence, P.; , "Noncontact Binocular Eye-Gaze Tracking for Point-of-Gaze Estimation in Three Dimensions," Biomedical Engineering, IEEE Transactions on , vol.56, no.3, pp.790-799, March 2009

[9] Ke Zhang; Xinbo Zhao; Zhong Ma; Yi Man; , "A Simplified 3D Gaze Tracking Technology with Stereo Vision," Optoelectronics and Image Processing (ICOIP), 2010 International Conference on , vol.1, no., pp.131-134, 11-12 Nov. 2010

[10] E. Miluzzo, T. Wang, and A. T. Campbell, “EyePhone: Activating Mobile Phones With Your Eyes,” Design, pp. 15-20, 2010.

[11] T. Nagamatsu, M. Yamamoto, and H. Sato. 2010. “MobiGaze: development of a gaze interface for handheld mobile devices”. In Proc. 28th Int. Conf. Extended abstracts on Human factors in computing systems (CHI EA '10). ACM, New York, NY, USA, 3349-3354.

[12] Drewes, H., Luca, A.D. and Schmidt, A. “Eye-gaze interaction for mobile phones”. in Proceedings of the 4th international conference on mobile technology, applications, and systems (2007), 364-371.

[13] S. Uehara et al, “1-inch Diagonal Transflective 2D and 3D LCD with HDDP arrangement”. in Proceedings of EI 2008 Volume 6803, San Jose, USA, January 2008. SPIE-IS&T Electronic Imaging 2008, Stereoscopic Displays and Applications XIX.