Available online at www.sciencedirect.com

ScienceDirect

Procedia Manufacturing 3 (2015) 837 – 844

Procedia MANUFACTURING

6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015) and the Affiliated Conferences, AHFE 2015

Gaze estimation technique for directing assistive robotics

C. Cole Drawdy, Paul M. Yanik

Western Carolina University, Cullowhee, NC, USA

Abstract

Assistive robotics may extend capabilities for individuals with reduced mobility or dexterity. However, effective use of robotic agents typically requires the user to issue control commands in the form of speech, gesture, or text. Thus, for unskilled or impaired users, the need for a paradigm of intuitive Human-Robot Interaction (HRI) is prevalent. It can be inferred that the most productive interactions are those in which the assistive agent is able to ascertain the intention of the user. Also, to perform a task, the agent must know the user's area of attention in three-dimensional space. Eye gaze tracking can be used as a method to determine a specific Volume of Interest (VOI). However, gaze tracking has heretofore been under-utilized as a means of interaction and control in 3D space. This research aims to determine a practical volume of interest in which an individual's eyes are focused by combining past methods in order to achieve greater effectiveness. The proposed method makes use of eye vergence as a useful depth discriminant to generate a tool for improved robot path planning. This research investigates the accuracy of the Vector Intersection (VI) model when applied to a usably large workspace volume. A neural network is also used in tandem with the VI model to create a combined model. The output of the combined model is a VOI that can be used as an aid in a number of applications including robot path planning, entertainment, ubiquitous computing, and others.

Keywords: Eye tracking; Gaze; Human-Robot Interaction; Neural networks

1. Introduction

1.1. Motivation

Assistive robotics may extend capabilities for individuals with reduced mobility or dexterity. However, effective use of robotic agents typically requires the user to issue control commands in the form of speech, gesture or text. Thus, for unskilled or impaired users, the need for a paradigm of intuitive Human-Robot Interaction (HRI) is prevalent. It can be inferred that the most productive interactions are those in which the assistive agent is able to

doi:10.1016/j.promfg.2015.07.339

ascertain the intention of the user. To this end, the assistive device could make use of current information on the user's area of attention. It is this aspect of HRI that is explored in this paper. There are many ways of representing a user's area of attention to an assistive device. One intuitive method is to track their *line of sight* (LoS) or *point of gaze* (PoG).

Where an individual is looking is a clear indicator of where their attention is focused [1]. For example, one area of research using gaze tracking has involved navigation of graphical user interfaces [2]. Two types of systems emerge when investigating gaze tracking: the 2-Dimensional system and the 3-Dimensional system. Here, a 2D gaze tracking system does not refer to how the eye is physically modeled, but to what the system uses as a display to the user. Such 2D systems estimate where on a plane (e.g. a video screen) the user is looking. That is, the system calibrates to a particular display at a certain distance from the user. These types of systems are used for GUI navigation which includes the use of the eye-mouse and eye-typing. Currently, there are several accurate systems that have been commercialized to navigate GUI interfaces [3-5].

Although 2D gaze tracking systems are useful when working with 2D devices, they limit the ability of a user from interacting with a device that moves in, or utilizes 3D space. The extension of gaze tracking from 2D to 3D has been the focus of research by several reseachers [1, 6-10]. Unlike 2D gaze tracking systems, 3D systems aim to use a stereoscopic environment (virtual reality) or 3D physical space as the display to an individual. However, the added dimension introduces inherent complexity in depth estimation for 3D systems.

For an individual to interact with a device in 3D space, the gaze tracking system must know the fixation point or area of interest of the individual in 3D space. These systems differ from gesture recognition systems or systems that track the 3D coordinates of a stylus because there is no item to physically track. Instead they use features that characterize the eye or the physical appearance of eyes to estimate the PoG. These include, fixation, saccadic rhythms, and vergence, among other eye characteristics [11]. Fixation is the act of keeping the eye in a fixed position. Saccades are quick rotations of the eye between points of fixation. Vergence is a measure of how two eyes converge when focusing on objects at particular distances [12].

Two-dimensional tracking techniques use data collected from the characteristics of eye movements and classify these movements into commands. Generally, these systems include sensory hardware that collect a user's eye movement data, and software to correlate it to a PoG on a monitor. A system of this type is used for the research in this paper and is extended to use in 3D gaze estimation.

The need for convenient methods of directing assistive robotics suggests the use of 3D gaze tracking. The purpose of the reported research is to use eye tracking to extend the recent 2D advances in eye tracking into a 3D environment. Instead of collecting data only on eye movement characteristics that correspond to point of gaze (PoG) at one depth plane, it is proposed that vergence data can be used to indicate other depth planes. Key objectives of this research are first, to apply a previously established non-contact, minimal-calibration 3D gaze estimation model to determine a volume of interest in a usable workspace. Second, the study combines techniques as it aims to boost the performance. Prior work in 2D and 3D gaze tracking is reviewed in section 1.2. Extending past approaches, two methods to determine the efficacy of using vergence data collected in a 2D setting for finding 3D PoG are proposed and investigated in sections 2 and 3.

1.2. Related work

Toward detecting PoG, eye movements must be classified according to their characteristic features. Blythe et al. [12] compare eye movements in three separate viewing scenarios: 2D representation, stereoscopic, and 3D representation (physical space). Vergence magnitude is shown to be very similar in both stereoscopic settings and in physical settings. This finding supports the motivation of this paper.

Several methods have been developed to track PoG in both 2D applications and 3D applications. However, in both cases, the approaches used are similar. Two prominent methods for tracking PoG in 3D space include the Pupil-Corneal Reaction technique and the model-based technique. Pupil-Corneal Reaction (PCR) utilizes the glint of light off of the cornea and pupil to determine the orientation of the eye and is thus, highly sensitive to participant movement. Model-based techniques are less sensitive to participant position and use models use eye contours to track the movement. Appearance-based models predict the appearance of the eye under varying movements. The predictions and actual images are compared using a similarity measure to indicate gaze direction [13]. Some

methods incorporate several techniques for eye tracking and are referred to as *hybrid* models. Such systems could be separated into categories depending on (1) display type, (2) hardware requirements, and (3) the use of *a priori* environment data [1].

Duchowski et al. [6] developed a system to determine 3D PoG by using stereoscopic display. An electromagnetic tracker was used to track head pose and a head-mounted eye tracking system determined the PoG on each 2D screen presented to the respective eyes. Using both head pose information and two separate 2D data sets for each eye, the 3D PoG was estimated.

Mitsugami et al. [7] developed a head-mounted system to estimate 3D PoG from 2D PoG estimates using a view camera. The display was a physical 3D setting. Intersection of view lines along with the known position and orientation of the head were used to find 3D PoG. The authors took samples of view lines from divergent angles of head positions. The viewing distances of the gaze targets were set at impractically far viewing distances (300 - 500 cm) and results were poor. This research uses a similar technique to estimate 3D PoG, without the use of head gear and within a practical but usable volume.

Essig et al. [8] introduced a tracking system that utilized a neural network to estimate the 3D PoG based on 2D binocular data. Because of their usefulness in statistical pattern recognition as applied to the problem of gaze estimation by these authors, a multilayer neural network [17] was be implemented as part of the research described in this paper. Participants viewed dot stereograms displayed on a computer screen. Data was collected on each eye's characteristics using a head-mounted gaze tracker. Kwon et al. [9] displayed a stereoscopic setting using a 2D parallax barrier to create the effect of depth by displaying slightly different images to each eye without the use of headgear.

Hennessey and Lawrence [1] used a hybrid model to estimate the absolute (x,y,z) coordinates of an individual's gaze. No head gear is required for their approach to calculate PoG in 3D space. Although highly accurate, the need for an extensive multi-part apparatus to extract gaze features along with a small 30 × 23 × 25 cm workspace would need to be circumvented for practical application with an assistive agent.

Hanhela et al. [14] attempted to determine the link between the number of participants and the precision of gaze estimation. Stereoscopic volumes of interest (SVOI) are extracted for each participant and intersected to estimate an overall volume of interest (VOI). It was concluded that accurate 2D estimation could be achieved using as few as 4 participants. As many as 13 participants were used to achieve desired accuracy in 3D. Wang et al. [15] also use stereoscopic displays to achieve accurate 3D gaze estimation through the addition of an online filtering technique and a computed triangulation disparity model.

Duchowski et al. [10] built upon [15] by comparing depth estimation when viewing stereoscopic displays and viewing a similar physical scene composed of four Snellen charts [16]. For depth estimation in the physical scene, the binocular Dikablis system was used. The hardware for the system was comprised of two individual cameras that track eye movements and produce monocular gaze estimates for each eye. Once calibrated the system could extract gaze vectors for each eye. Intersection of these vectors produced a gaze point that indicates where the user is looking. This method is referred to as the *vector intersection* (VI) model. Because, unlike the disparity model, the VI model is useful in physical spaces, it will be implemented as part of the research reported in this paper.

Drawing on these past efforts, the system described below addresses a physical space using non-contact hardware, and *a priori* knowledge of gaze targets within the environment.

2. Method

2.1. Experimental setup and design

A Tobii TX300 eye tracker was used to collect gaze data for the experiment. The Tobii device measures several eye movement characteristics at high precision on a 2D platform [4]. Gaze data samples were collected at 300 Hz. The eye-tracking device was set facing the user at a distance 60 cm. A series of 5 planes (or *windows*) printed with dots were placed in view of the user at distances of 60, 80, 100, 120 and 180 cm. The planes increased proportionally in size as their distance from the user increased. This arrangement (Fig. 1) created a pyramidal frustum of gaze targets. Calibration was conducted using the window nearest to the user (window 1).

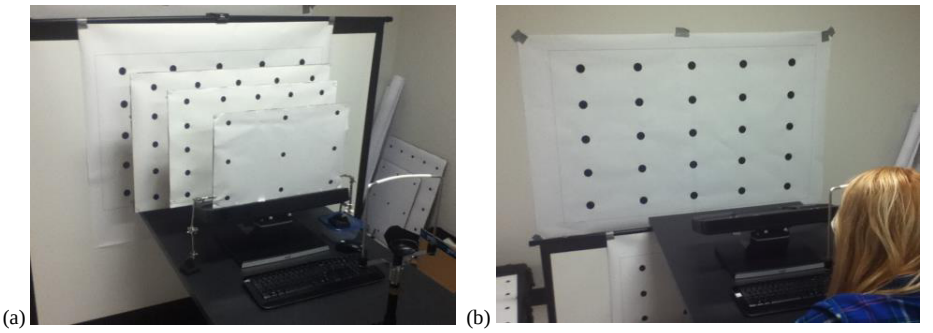

Fig. 1. (a) The data collection fixture; (b) data collection at window 5.

The Tobii device calibrates to a 2D monitor. The device is meant for tracking eye movements with respect to one plane of depth. Thus, all data is used to approximate the PoG of a user on the 2D screen rather than in 3D space. Similar to the Dikablis system used in [10], the user calibrates the eye-tracker to a 2D plane before any data is taken. Window 1 was used as the calibration plane.

2.2. Data collection and processing

Gaze data was collected from 8 participants. Each participant first underwent a 2D calibration step using the nine dots of window 1. Next, participants were asked to fixate upon each dot in turn, progressing through the 5 windows. For each dot, fixations were recorded for 1 to 3 seconds to obtain 300 – 900 samples at 300 Hz. The data for each sample consisted of the (*x*,*y*) coordinates of locations where the left and right eyes' line of sight intersected window 1. Thus, each sample consisted of (*x-left, y-left*) and (*x-right, y-right*) points. Outlier samples that fell outside of two standard deviations from the mean were removed. Smoothing was performed in both *x* and *y* directions using a moving average over 10 samples.

2.3. Vector intersection approach

The vector intersection model uses the position of each eye to draw vector to their associated gaze points on the calibration plane. The two vectors are then extended through the calibration plane to a point of intersection (or the closest point at which they cross) producing a specific point (*x*,*y*,*z*) for each sample. Samples for each dot per participant produces a cluster. The centroid of this cluster was assigned as the estimated gaze point for each dot. Thus, 109 centroids (equal to the total number of dots) per participant were computed. Fig. 2 illustrates this concept for the case of a single participant fixating on a given dot. Error for the VI model is calculated as the Euclidean distance from the centroid to the actual dot location.

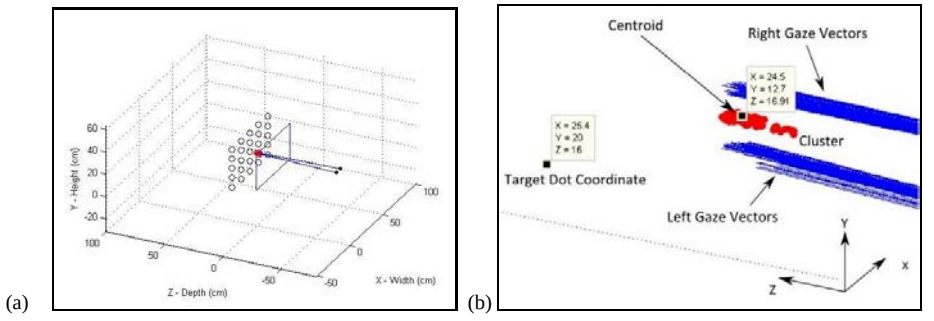

Fig. 2. Vector intersection: (a) gaze vectors for a participant fixating on a specific dot; (b) gaze vectors (blue lines) for all samples associated with the dot, their intersections (green dots), and the actual coordinates of the actual stimulus dot (empty black circle).

2.4. Neural network approach

A neural network was implemented for this research using the Matlab® Neural Networks Toolbox. The network topology consisted of three fully-connected layers with four input units, 50 hidden units and five output units. Input patterns consisted of the left and right eye gaze coordinates at the calibration window: (*x-left, y-left*) and (*x-right, y-right*). Training data consisted of 120 samples for each dot per participant or 52320 (120 × 109 × 4) total samples. Output patterns (for training) were encoded as +1 for the desired window, and -1 for all others. For example, if the input pattern was known to correspond to a participant fixating on a dot of window 4, the output pattern would be [-1,-1,-1,+1,-1].

The network was trained using 4 participants and tested using the remaining 4 participants. Thus, it was possible to generate 70 different combinations (*8 choose 4*) of training and test data. Patterns from the test data set were used to determine the accuracy of the network. The maximum value of the outputs for each sample was interpreted as the depth plane corresponding to a given input vector. With each sample assigned to a discrete depth plane (*z* value), an average *vector of interest* (VecOI) was computed. Each dot would have a cluster of points and, unlike the vector intersection model, the cluster was restricted to discrete *z* values (windows). The centroid of each cluster was found for each dot, representing the estimated PoG for each of the 109 dots. The Euclidean error was then computed using the actual coordinates of the dots and the estimated centroids.

2.5. Combined approach

A Bounded Vector of Interest (BVecOI) may be generated using PoG centroid estimates from both the VI and Neural Network (NN) approaches. This was done by connecting the centroids from the two approaches by a vector and extending a radius from the vector to create a volume. Thus, for each participant, a *volume of interest* (VOI) was generated. Accuracy was found by determining whether a VOI encompasses the coordinates of its respective dot. Fig. 3 shows the BVecOI for an example group of dots.

2.6. Searching region approach

The VOI in the Combined Model is determined by two centroids that may not be accurate under high resolution. Understanding that vergence, and in general, the human eye, is prone to variations depending on factors such as focus, human mechanics, or fatigue, it is warranted that larger *searching regions* (SR) be defined. Therefore, the frustum of view used here may be divided into three volumes, or *searching regions* as shown in Fig. 4. Both the VI approach and the NN approach yield centroid estimates for every dot which fall into one of the defined regions. Using these regions the VecOI may be bounded. Unlike the combined approach, bounds are fixed and the models are treated separately. The SR approach consists of (1) generating an average VecOI, and (2) bounding the average VecOI by the searching region (the *Search Region Bounded Vector of Interest* or SBVecOI). Accuracy of this approach is the Euclidean error between each point on the SBVecOI and the dot coordinates.

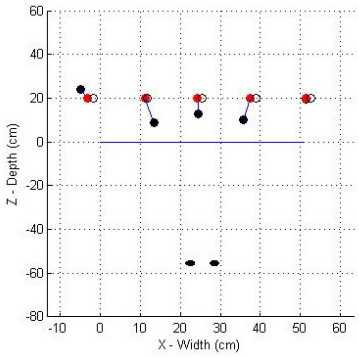

Fig. 3. Combined approach: overhead view of vector-connected centroids from the VI approach (black) and NN approach (red).

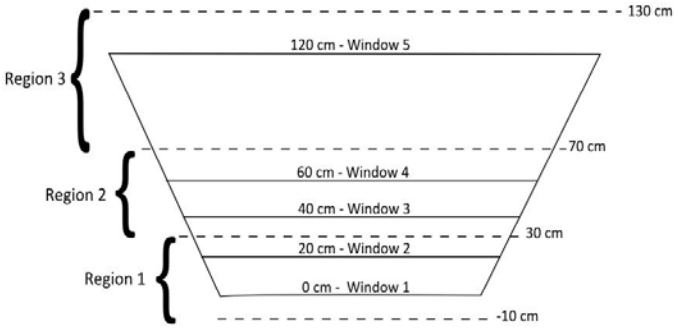

Fig. 4. Search regions.

3. Results and discussion

3.1. Vector intersection approach

Accuracy for the VI approach was computed as the Euclidean distance from the estimated centroid to the respective dot under scrutiny. For each window an accuracy radius (tolerance) was defined as half the minimum distance between adjacent dots. Table 1 gives results for the VI approach. It can be seen that the accuracy is poor (less than 25% across most participants). The large variation in results over the pool of participants could occur due to what was actually focused upon. The black round dots used as stimuli provided little interesting information to the user. Being 2-dimensional and one solid color, it may be that focus was not easily maintained.

Table 1. Accuracy of the VI approach.

Participant	Accuracy (%)
1	18.3
2	34.8
3	9.1
4	3.6
5	9.1
6	23.8
7	16.5
8	20.1

3.2. Neural networks approach

The neural network was trained using 70 different combinations of training data and test data. Results from all combinations are not reported here. Combinations producing the most accurate, median, and poorest results are given by Table 2. Accuracy was calculated in the same manner as accuracy for the VI model. A radius drawn from the each estimated centroid was used to create a spherical volume of interest. If the corresponding dot coordinates fell within the volume, then the estimate deemed accurate. The percentage accuracy is higher than that of the VI model, owing to the fact that the neural network was trained to discrete values; effectively *snapping* to the windows. Nonetheless, the accuracy is not sufficient for robot guidance in HRI scenarios. Taking into consideration the poor performance of the VI and NN approaches to accurately predict where one is focused in 3D space, the motivation to investigate the combined approach is justified.

Table 2. Accuracy of the NN approach.

NN index (of 70)	Training Participant #s	MSE	Test Participant #s	Accuracy per participant (%)
6	1, 2, 4, 5	0.3169	3, 6, 7, 8	65.1, 60.5, 59.6, 70.6
54	2, 5, 7, 8	0.4765	1, 3, 4, 6	44.9, 49.5, 39.4, 44.9
29	1, 4, 6, 7	0.7385	2, 3, 5, 8	10.0, 9.1, 10.0, 9.2

3.3. Combined approach

The combined VI-NN approach uses the previous steps to produce a BVecOI based on estimated centroids. The centroids are connected by a vector and a radius is extended from each point on the vector. If the actual dot's coordinates lie within this VOI, then the approach is deemed accurate. Results of this approach are given by Table 3. The table shows results for best, median and worst NN training/test scenarios. It can be seen that accuracy is boosted when this approach is applied. However, the length of the BVecOI is inconsistent across the sample set. In fact, it may be quite long in cases where the VI model is inaccurate. This could result in a large uncertainty in path planning for a robotic agent whose goal is to travel to the PoG.

Table 3. Accuracy of the combined approach.

NN index (of 70)	Training Participant #s	Test Participant #s	Accuracy per participant (%)
6	1, 2, 4, 5	3, 6, 7, 8	73.7, 80.3, 69.9, 88.8
54	2, 5, 7, 8	1, 3, 4, 6	78.4, 74.4, 55.5, 66.1
29	1, 4, 6, 7	2, 3, 5, 8	59.2, 34.4, 36.0, 55.2

3.4. Searching region results

Unlike the combined approach, the SR approach is advantageous in that the length of a BVecOI is constrained to the depth of a region: 40 cm or 60 cm. The accuracy of the SR approach is given by Table 4 (VI approach) and Table 5 (NN approach). These results reflect higher accuracy than those of the VI or NN approaches alone. However, it is noted that, for other NN indices than those shown, the effective decrease in resolution results in significantly reduced accuracy.

Table 4. Accuracy of the searching regions approach with VI.

Participant	Accuracy (%)
1	71.4
2	83.4
3	59.6
4	42.2
5	50.4
6	73.3
7	69.7
8	77.9

Table 5. Accuracy of the searching regions approach with NN.

NN index (of 70)	Training Participant #s	Test Participant #s	Accuracy per participant (%)
6	1, 2, 4, 5	3, 6, 7, 8	76.1, 70.6, 68.8, 81.6
54	2, 5, 7, 8	1, 3, 4, 6	68.8, 71.5, 67.8, 63.3
29	1, 4, 6, 7	2, 3, 5, 8	35.7, 32.1, 33.0, 34.8

4. Conclusions

In this paper, methods for using vergence data to determine depth of gaze have been explored. The common vector intersection approach has been shown to perform poorly over our data set. A neural network approach showed improved results over VI, but with the constraint that depth computations were constrained to discrete steps. A combined model which bounded the region of a gaze stimulus between those of the VI and NN approaches further improved our results while adding uncertainty in the form of region volume. Finally, best results were found by dividing the viewing frustum into defined search regions. However, it was seen that gaze data vary significantly over our sample set. Thus, even in the best of cases, estimation depth using vergence should be considered as one among a larger set of depth search criteria.

References

[1] C. Hennessey and P. Lawrence, "Noncontact binocular eye-gaze tracking for point-of-gaze estimation in three dimensions", IEEE Transactions on Biomedical Engineering, 56:3 (2009) 790-799.
[2] Y. Zhang, "Improvements to the accuracy of eye tracking data based on probable fixations," Department of CIS Technical Report 2010-04, University of Oregon, 2010.
[3] SensoMotoric Instruments, "Eye tracking technology by sensomotoric instruments," http://www.smivision.com/oem-eye-tracking/index.htm, 2014.
[4] Tobii Technology, "Tobii eye tracking research," http://www.tobii.com/en/eye-tracking-research/global/, 2014.
[5] iMotions, "Remote eye trackers," http://imotionsglobal.com/hardware/remote-eye-trackers/, 2014.
[6] A.T. Duchowski, V. Shivashankaraiah, T. Rawls, A.K. Gramopadhye, B.J. Melloy, and B. Kanki, "Binocular eye tracking in virtual reality for inspection training," in Proc. of 2000 Symposium on Eye Tracking Research and Applications, (2000) 89-96.
[7] I. Mitsugami, N. Ukita and M. Kidode, "Estimation of 3D gazed position using view lines," in Proc. of the 12th Intl. Conf. on Image Analysis Processes, (2003) 466-71.
[8] K. Essig, M. Pomplun and H. Ritter, "Application of a novel neural approach to 3D gaze tracking: Vergence eye-movements in autostereograms," in Proc. of the 26th Annual Meeting of the Cognitive Science Society, (2004) 357-362.
[9] Y. Kwon and K. Jeon, "Gaze computer interaction on stereo display," in Proc. of the 2006 ACM SIGCHI Intl. Conf. on Advances in Computer Entertainment Technology, (2006) 526-531.
[10] A.T. Duchowski, D.H. House, J. Gestring, R. Congdon, L. Swirski, N.A. Dodgson, K. Krejtz and I. Krejtz, "Comparing estimated gaze depth in virtual and physical environments," in Proc. of the Symposium on Eye Tracking Research and Applications, (2014) 103-110.
[11] Tobii Technology, "Accuracy and precision test method for remote eye trackers," Test Specification Version 2.1.1, (2011) 6-7.
[12] H.I. Blythe, N.S. Holliman, S. Jainta, L.W. Tbaily and S.P. Liversedge, "Binocular coordination in response to two-dimensional, three-dimensional and stereoscopic visual stimuli," J. College of Optometrists, 32 (2012) 397-411.
[13] 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, 32:3 (2010) 478-500.
[14] M. Hanhela, A. Boev, A. Gotchev and M. Hannuteela, "Fusion of eye-tracking data from multiple observers for increased 3D gaze tracking precision," in Proc. of the 2012 20th European Signal Processing Conference (EUSIPCO), (2012) 420-24.
[15] R.I. Wang, B. Pelfrey, A.T. Duchowski and D.H. House, "Online gaze disparity via binocular eye tracking on stereoscopic displays," in Proc. of the 2012 Second Intl. Conf. on 3D Modeling, Processing, Visualization and Transmission, (2012), 184-191.
[16] G.D. Love, D.M. Hoffman, P.J.W. Hands, J. Gao, A.K Kirby and M.S. Banks, "High-speed switchable lens enables development of a volumetric stereoscopic display," Optics Express, 17:18 (2009) 15716-15725.
[17] D.G. Stork, R.O. Duda and P.E. Hart, Pattern Classification, second ed., John Wiley & Sons, New York, 1996.
[18] C.C. Drawdy, "A Technique for Estimating a Three-Dimensional Volume of Interest Using Eye Gaze," Master's thesis, Department of Engineering and Technology, Western Carolina University, Cullowhee, NC, USA, 2015.