Using natural versus artificial stimuli to perform calibration for 3D gaze tracking

Christophe Maggia, Nathalie Guyader, Anne Guérin-Dugué

Using natural vs. artificial stimuli to perform calibration For 3D gaze tracking

Christophe Maggia, Nathalie Guyader, Anne Guérin Dugué
Laboratory of Images Speech Signal Automatism, 11 rue des mathématiques, Grenoble Campus, FRANCE

ABSTRACT

The presented study tests which type of stereoscopic image, natural or artificial, is more adapted to perform efficient and reliable calibration in order to track the gaze of observers in 3D space using classical 2D eye tracker. We measured the horizontal disparities, i.e. the difference between the x coordinates of the two eyes obtained using a 2D eye tracker. This disparity was recorded for each observer and for several target positions he had to fixate. Target positions were equally distributed in the 3D space, some on the screen (with a null disparity), some behind the screen (uncrossed disparity) and others in front of the screen (crossed disparity). We tested different regression models (linear and non linear) to explain either the true disparity or the depth with the measured disparity. Models were tested and compared on their prediction error for new targets at new positions. First of all, we found that we obtained more reliable disparities measures when using natural stereoscopic images rather than artificial. Second, we found that overall a non-linear model was more efficient. Finally, we discuss the fact that our results were observer dependent, with variability's between the observer's behavior when looking at 3D stimuli. Because of this variability, we proposed to compute observer specific model to accurately predict their gaze position when exploring 3D stimuli.

Keywords: vision, binocular disparity, eye tracking, calibration, natural scenes, artificial images

1. INTRODUCTION

3D perception arises from multiple depth cues. Monocular cues need only one eye to be perceived and include occlusion, perspective, relative size, texture.[1], whereas binocular cues need the two eyes and include binocular disparity and vergence. In this research, we are interested in binocular cues. *Binocular disparity* is mainly due to the horizontal separation of the two eyes, getting two shifted versions of the same visual input in the two retinas[2,3]. The horizontal disparity between the two retina images is responsible for the depth perception in stereo vision.

A stereo image is composed of two versions of the same scene, a left image and a right image that corresponds to a shifted version of the left image. 3D devices use the same principle since Wheatstone[2]. To ensure that one eye only sees one image of the stereo pair, glasses are usually used: either active or passive glasses. When studying binocular disparity, two types of disparity are distinguished: the crossed disparity to perceive objects that are in front of the reference plane and the uncrossed disparity to perceive objects behind the reference plane. Zero disparity occurs when fixating the reference plane. *Vergence eye movements* correctly align the fixation point on the centers of the retinas, which allows the brain to fuse the two images into a single percept [4,5]. Consequently, vergence eye movements are crucial to an accurate stereoscopic perception of depth and were showed to be proportional to the retinal disparity amplitude. Due to fusional vergence, the fixation point in the reference plan has zero disparity whereas all the other objects have specific retinal disparity according to their position in depth from the fixation point. There are two types of vergence eye movements, convergent and divergent movements that occur respectively when fixating objects nearer or further than the precedent fixation.

Since recent developments of easy-to-use 3D visualization devices such as 3D TVs or 3D compatible screens with 3D glasses, some research on visual perception use these displays to work with more natural stimuli. In parallel, eye tracking is often used in several research fields like computer science, psychology, neurosciences. By now, eye tracker predict eye positions on a 2D screen but are not able to accurately recover the point of gaze in a 3D space. Consequently, there is an attempt to develop methodology to predict, using 2D eye tracker, the gaze of observer looking at 3D visual stimuli. We called this methodology 3D calibration. Without a reliable 3D calibration, it is not possible to predict gaze position in the 3D space. However, some research did study gaze exploration during 3D stimuli viewing. But they did not retrieve the gaze of observers in the 3D space. They used a “depth map” associated to each visual stimulus and projected the 2D gaze monocular coordinates onto this “depth map” acquired with a laser camera to estimate the depth plane gazed at. This constitutes a technological constraint to study large natural image database. Moreover, the laser camera is quite expensive and we think that using only one eye data is not reliable to determine the 3D gaze location.

Hence, some studies were interested in recovering 3D gaze positions using artificial stereoscopic stimuli from vergence eye movements or from binocular disparity using 2D eyetrackers[5,6]. The goal of these authors was to compute a model specific to the observer’s gaze characteristics and to compensate for camera-related image distortions. One study developed a geometrical model (fitting model[5]) that uses the same polynomial interpolation that the one used for 2D calibration. They tested their model using artificial visual stimuli with a rotating target. This rotation was to test the influence of motion parallax on stereopsis but they did not found any significant effect of motion parallax. In another study, they developed a neural network[6] that showed good performances on artificial stimuli. Another way was also explored[7] by an algorithm using the concept of Pupil Center Distance to estimate 3D gaze position on artificial stimuli. This measure is dependent of vergence since these ocular movements tend to increase or decrease this distance according to the depth plan fixated. Their system was coupled to a game application where observer had to select with gaze a dart displayed at a given depth and gaze on a target at another depth to throw the dart. Their system seemed to have good accuracy but was purely geometric so it did not adapt to the observer 3D fixation pattern. Vergence eye movements were showed to be correlated with disparity changes displayed on a 3D device[5]. Vergence eye movements might be measured to further retrieve disparity.

In our study, we propose a method to predict, using a 2D eye tracker, the observer’s gaze in a 3D space when he is viewing 3D stimuli. Knowing that natural images are complex and richer in depth cues than artificial images, it is interesting to discuss the best type of stimuli to use in a 3D calibration. In fact for the moment, as far as we know, all the papers that worked on 3D calibration used artificial stereo stimuli. The method which we proposed was used with either natural or artificial stereoscopic stimuli. Observers were asked to gaze at nine positions at different depths for both kinds of stimuli. We think it is important to test each calibration point with all disparity conditions to take into account the alteration that eccentricity would have on 2D eye tracking and by extension on 3D gaze tracking. For each position, the binocular disparity recorded through a 2D eye tracker was measured, this constitutes the calibration part. This allowed us to compute and to test different types of models, linear and non-linear, to predict the true disparity from the recorded disparity (and hence, the depth of gaze). In order to test the performance of this 3D calibration for the two types of stimuli (natural and artificial)and the efficiency of the different models, the experiment involved a test part in which new positions and new depth planes were presented using both types of stimuli. Models, computed with the positions displayed during the calibration part, were tested to predict 3D gaze for the new positions displayed during the testing part. In the following, we present the design of the proposed experiment, the calibration and the test parts. We develop the data analysis and the comparison between the different models.

2. MATERIALS AND METHODS

The eye tracker is calibrated with both eyes on a 2D reference plane (the point P on the screen surface, Figure 1). During a fixation (P), the fusional vergence adjusts each eye’s view on the same point resulting in a null disparity. When gazing at a point (F, C) in front or behind the 2D plan, a disparity between the gazes of the two eyes is detected on the screen, since each eye gaze at a different positions (C_L, F_L for the left eye and C_R, F_R for the right eye). By simply measuring the disparity between the right and the left horizontal gaze coordinates ($\delta x = C_L - C_R$, or $F_L - F_R$) obtained through an eye tracker, we might deduct the horizontal disparity type (crossed or uncrossed) with the sign of the disparity and its amplitude. Finally, depth (Z coordinate) might be deducted from the real disparity Δx. Indeed, knowing the geometric configuration (D and L) and the real disparity Δx displayed for each point (F, or C), the theoretical Z (Figure 1) can be estimated such as :

$$Z = \frac{-D.\Delta x}{L + \Delta x}$$

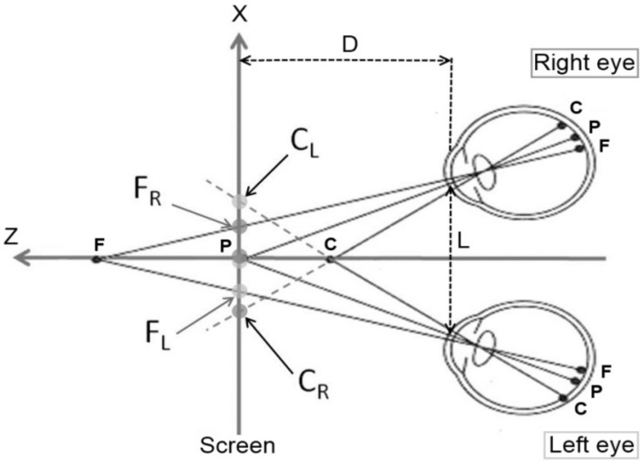

Figure 1: The point P represents the basis of the calibration. C and F elicit gaze positions which reflect each point disparity. C has a crossed disparity, while F has an uncrossed disparity. D represents the viewing distance. L is the inter-ocular distance.

2.1. Participants

All participants (n=7, 4 F, 3 M, ages 30.5 on average) had a normal or corrected to normal vision and were pre-screened for depth perception by binocular disparity thanks to the test for stereoscopic vision TNO (Lameris Ootech BV Nieuwegen). This test uses red/green anaglyphic stereogram to determinate each participant perceptive thresholds. Only the participants who had a stereo threshold of 60 arc sec or less in both crossed and uncrossed binocular disparity were accepted to participate in our experiment.

Visual impairments like strabismus and color blindness were also avoided although we did not test them. All participants had already seen stereoscopic movies and we made sure they understood the full task and got familiar with the material before the experiment. The interpupillary distance (L, Figure 1) was measured for each participant.

2.2. Apparatus

We used a remote binocular eye tracker EyeLink 1000 (SR Research) to track the gaze of each eye while the observer is looking at stereo stimuli. The EyeLink system was used in the Pupil-Corneal Reflection tracking mode sampling at 1000Hz. In order to combine eye tracking and 3DTV experience, 3D glasses must be compatible with the Infrared camera of the eye tracker. Only polarized glasses (passive system) should be used because most of the shutter glasses (active system) use infrared signal to synchronize with the TV, which results in interferences between the two devices. The TV used is an LG 32LW4500 (70cm wide and 39.3cm high). According to the manufacturer, the 3d brightness is 150 cd/m² and the contrast ratio measured is 1240:1. The resolution of 1920*1080 pixels allows at a viewing distance (D) of 132cm a minimum disparity of 57.6 arc sec for a total angular size of 30.13°. This is much bigger than the finest disparity that the human eye can see (up to few arc sec[8,9]). The experiment took place in a dark room to increase 3D immersion.

2.3. Stimuli

Two types of stimuli were tested in this study: artificial images and natural pictures. Two images were created for each eye (the left and the right image) with the same information and a horizontal shift between both images (the shift that corresponds to the true disparity is given in pixels). Natural pictures were taken with a stereoscopic photo camera (Fujifilm W3), whereas artificial images were created using Matlab software. All images were converted to grayscale then equalized to an average luminance of 127 (on a luminance scale [0,255]). Stimuli were resized to 12.8° of visual angle (800*450pixels). Images were centered and the rest of the screen was black. The black border enhances the 3D perception and immersion. Both types of stimuli were constructed according to the same principle as the eye tracker calibration stimuli. Each stereo image contained a target located on one of nine possible locations equally distributed on the 2D screen (similar to the well-known 9 points calibration used in classical 2D eye tracking[10] Figure 2). Then, each stereo image was modified to display the nine positions on five distinct depth planes. One depth corresponds to the null disparity, two planes to crossed disparity (disparities equal to 32 and 16 pixels) and two to uncrossed disparity (Δx equal to -32 and -16 pixels). 45 targets (9 positions x 5 depth planes) referred as "learned positions and learned disparities" were displayed during the calibration part and further used to compute the models.

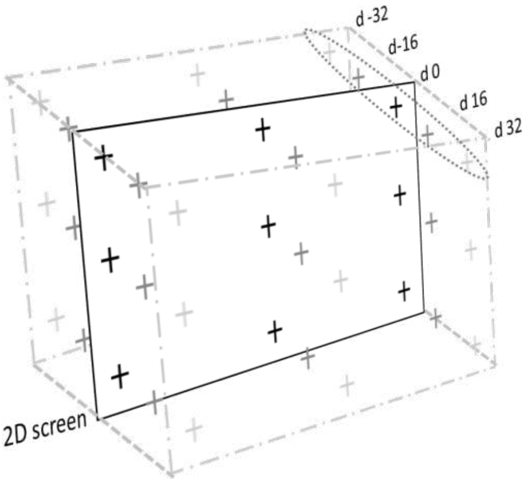

Figure 2. Schematic representation of the 3D space used in the calibration. Nine positions are used at five different disparities. d is given in pixels

In the test part, we designed images with four new targets located at intermediate positions “New positions”. In addition to the five learned disparities “New positions and learned disparities”, we tested four new disparities “New positions and new disparities” (two crossed Δx= 40, 24pixels and two uncrossed Δx=-24, -40 pixels). This testing part was constituted of 36 new targets (4 positions * 9 depth planes). These positions allow us to test models efficiency to retrieve new positions (interpolation and extrapolation).

2.3.1. Artificial stimuli

Artificial stimuli can be easily adapted to both crossed and uncrossed disparities by reversing right and left images. When testing these artificial stimuli with only the target that had a different plane, it was very difficult to perceive clearly the depth of the target. To make the task easier for the observer, other squares were added both in and out of the target depth plane to allow relative disparity judgment (Figure 3). In fact, it is known that human depth perception is more based on relative disparity than absolute disparity making it easier and more accurate[3,9]. The target was a white square (21*21 pixels, 20,1 arc min of visual angle with a black dot on the center) included in a 5*5 grid of purple squares (Figure 3). The nine positions described earlier moved in the same depth plane as the target whereas the others stayed in 2D as a reference plan.

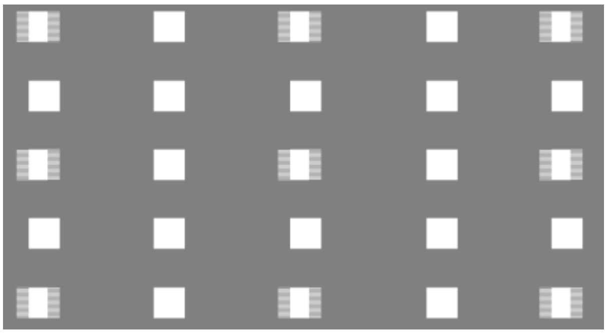

Figure 3 Stereoscopic view of an artificial stimuli. 9 positions are displayed in 3D whereas all the other squares are on the screen plane (null disparity)

2.3.2. Natural stimuli

It is almost impossible to find natural scenes where both disparities would be present and localized at 9 different positions. Hence, we used two different scenes for crossed and uncrossed disparities. In the Figure 4 we only represented the scenes with the target in the central position. We tried to avoid monocular depth cues (perspective, occlusions, and relative size) as it has been shown that monocular cues can influence vergence eye movements even on 2D images[11]. Observer knew beforehand the target he had to gaze at in the scene. Moreover a small dot was added on the center of each target to help the observer to be more precise when fixating.

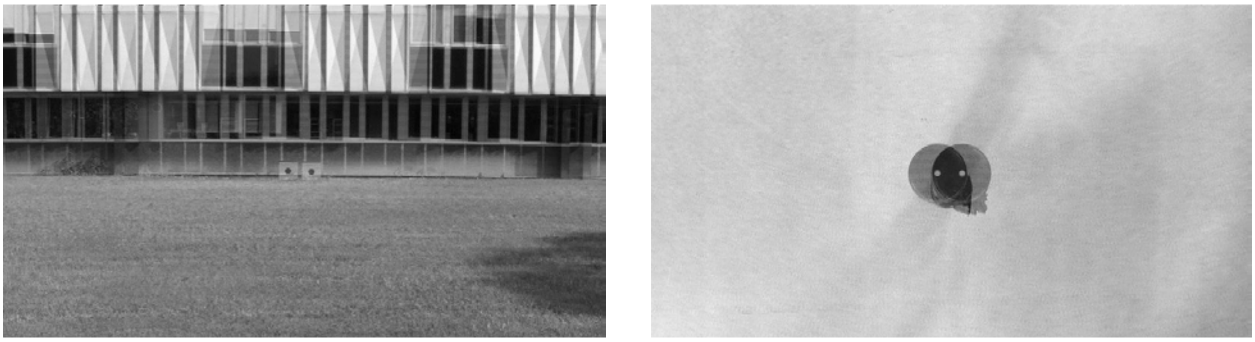

Figure 4 Stereoscopic view of the central position of natural stimuli for crossed (Right) and uncrossed (Left) disparity. Target angular size is 24 arc min for uncrossed disparities and 73 arc min for crossed disparities

2.4. Experimental design

All participants underwent the experiment twice, one with natural and one with artificial images in a random order in separated days to avoid visual fatigue. The experiment was divided in two parts: a calibration and a test.

One trial was made up of three steps. In the first one, a 2D image displayed the target in one of the nine spatial positions for 1.5sec. Then, the image with the same target in one of the five depth planes was displayed for 6sec maximum. Observers were told to gaze at the target as precisely as possible during this delay. This succession of two images permitted to clearly see the difference between 3D and 2D images and make the task easier. For the targets in the reference plane (null disparity), there was a succession of two 2D images. We made the images switched from one image to the other of the stereo pair. The image used for the target was selected according to the dominant eye of the observer. The target image was displayed for 6 sec which might be long, but vergence is a slow eye movement and this duration allowed the observer to explore the stimulus and the different depth planes before fixating the target. Finally, a mean gray screen remained for 1sec before the next trial. Target positions were randomized.

Visual fatigue is a crucial factor when observing vergence eye movements. Some authors[12,13] noticed a decrease in oculomotor performance even before observer did subjective complaints. To fixate on a 3D point can alter vergence eye movements if it is repeated during a long period. To make the task more pleasant and to shorten the experiment, we used a contingent gaze display to remove the target image before the 6 sec. If the gaze of both eyes were fixated within a restricted area around the target for at least 600ms, the image was removed and the next trial started. The size of the area varied according to the dominant eye of the observer. In fact, we noticed that the non-dominant eye gaze was less precise than the dominant one. We defined an area of 75 pixels centered on the target for the dominant eye and an area of 100 pixels centered on the target for the non dominant eye. If the two eyes did not gaze in this area, the stimulus was displayed for 6 sec and the recorded data were not further analyzed. When eyes were fixed in the defined area around the target, we kept all the eye positions recorded.

We further need to select during this period which eye positions to analyze. Actually, during this period we recorded several events: a saccade toward the target, then several saccades still might be recorded on the target area but also around the target area. An offline selection was realized on the fixations detected by the Eyelink system for the dominant eye. We kept the fixations that were not farther than 50 pixels from the target. If multiple fixations fitted this criterion, the longest one was kept.

We chose to repeat every 45 targets twice. This repetition decreases the risk to lack data for some positions. Every 9 images, a 2D calibration was done. A verification step was made to ensure that the mean error between the eye tracker estimation and observer's fixations was lower than 1° for each eye. This allowed to keep the maximum tracking accuracy all along the experiment, but also to allow breaks for observers.

2.5. Data extraction

As already noticed by authors [5], ocular data stability is an important issue when tracking gaze in depth (Figure 5). It has been shown that vergence eye movements in response to stepwise change in disparity need time before stabilization[14]. However, during the 3D fixations, it happened that the stabilization did not last. We noticed some oscillations or drifts of the recorded eye positions around the target position without a conscious perception of this phenomenon by observers

(Figure 5). The fixation detected by the EyeLink system is based on the mean of raw eye positions. As we can see in the example showed in Figure 5, the δx measured stabilize at a lower disparity than the Δx displayed (-32 pixels). The mean δx of such a fixation is influenced by data outside the stabilized period. So the fixation summary given by the Eyelink system do not represent accurately 3D fixation. On the contrary, by taking the most frequent δx (distribution mode) as the resulting fixation summary, we free ourselves from the vergence adjustments movements to only keep the real stabilized period even if it stray away from the Δx displayed .

Each fixation corresponded to several raw eye positions, and in function of the fixation duration we might have more or less raw eye positions. For each fixation, we extracted 80 eye positions closest to the mode of the fixation. . Two third of them were used to compute the different models whereas the last third was kept to test the efficiency of models.

When no fixation occurred in the target area (defined 2.4) the corresponding trial was referred as "Missing data". Because we did two repetitions for each target there were in most of the cases two correct fixations for each target. If the two repetitions were consistent (a 30 pixel thresholds was empirically chosen after preliminary testing), the closest to the target position was kept. Sometimes, it happened that two fixations for the same target correspond to different disparity signs, crossed and uncrossed; these trials was referred as "Uncertain data".

Finally, it sometimes happened that an observer for one spatial position showed disparities larger than the true disparities. Hence, we also considered for one spatial position the evolution of the measured disparities for the five depth planes. Due to our experimental design with true disparities linearly distributed, we should also recorded linearly distributed disparities. Hence a linear regression was computed from the selected data and all the possible combinations of the uncertain ones. The combination, for which the mean square error (MSE) between the linear regression and the data recorded was minimal, was kept. This way, we preserved the observer tendency even if his data were far from the theoretical ones but aligned. However despite this procedure, aberrant values could remain. We manually checked every model to remove outliers impairing a reliable regression (see 2.5.1).

Some trials were considered as "outlier" if the MSE error of the regression significantly decreased after removal.

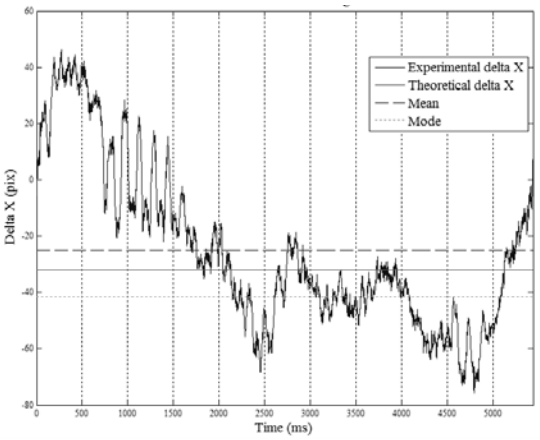

Figure 5 Time course of a delta x measure during a 3D fixation. Ocular data are noisy. The observer does not always follow the theoretical value, in this case he stabilizes himself at farther depth than the presented one. The mean as a fixation resume is too influenced by drift and noise, whereas the mode only keep the best stablilized period.

2.5.1. Models

We decided to divide the screen in nine parts to model the ocular behavior in front of 3D stimuli. The accuracy of the eye tracking tend to be poorer a high eccentricity [10], so to create a mapping that includes this loss of accuracy to retrieve the disparity displayed, we made models on each of the nine positions.

Theoretically, the measured δx has to match the displayed disparity Δx (Figure 1 C_L-C_R and F_L-F_R). This describe a linear evolution whereas the computed depth coordinate Z has a non-linear evolution close from a hyperbole (Figure 6, equation in 2.) with two asymptotic lines ($\Delta x= -L$, $Z= -D$). Consequently, we designed different types of model based either on δx (linear: model *ab*) or Z (nonlinear: model *dO* and *DLO*). The relationship between Z and Δx is dependent of the inter-ocular distance L and the viewing distance D. D is constant depending of the experimental setup. L is measured for each subject. Let us notice δx be the experimental disparity measured by the eye tracker, Z the theoretical depth, $\hat{Z}$ the estimated depth and $\delta\hat{x}$ the estimated disparity. Concerning the parameters for the models *dO* and *DLO*, let us notice

D and L the physical parameters and **D** and **L**, the estimated parameters thanks to the nonlinear regression. And then for all the models, all the estimated parameters are noticed in bold. To compare them on the disparity values, we have to convert the Z estimation obtained with the non linear models in δx estimation by using the inverse of the theoretical equation:

$$\delta\hat{x} = -\frac{\hat{Z}\times\mathrm{L}}{\mathrm{D}+\hat{Z}}$$

The tree models are detailed hereafter (Figure 6b):

- Model *ab* : Linear regression applied on the δx data with two adjustable parameters:

$$\delta\hat{x} = \boldsymbol{a} \times \delta x + \boldsymbol{b}$$

- Model *dO*: Non-linear model with two adjustable parameters, in which additional noise on δx (**d**) and a general offset **O** are adaptable to obtain $\hat{Z}$. The initial value for d and O are 0 for both. The model is described by:

$$\hat{Z} = \frac{-D \times (\delta x + \boldsymbol{d})}{L + (\delta x + \boldsymbol{d})} + \boldsymbol{O}$$

- Model *DLO*: Non-linear model with three adjustable parameters which are the initial parameters **L** and **D** as well as the offset **O**. The initial value for D, L and O are respectively D, L and 0. The model is described by:

$$\hat{Z} = \frac{-\boldsymbol{D} \times \delta x}{\boldsymbol{L} + \delta x} + \boldsymbol{O}$$

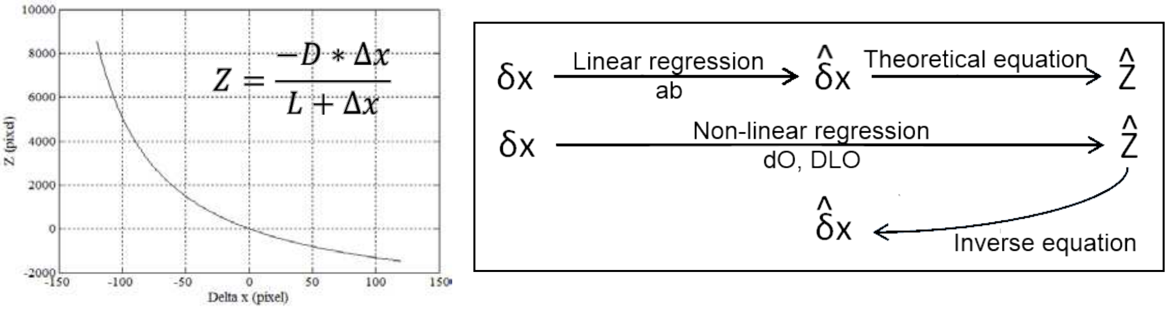

Figure 6 (Left) hyperbolic evolution of Z in function of Δx. There is an asymptote in –L and D. b. (Right) Schema of the application of the linear and non-linear model on δx and Z

We noticed that the theoretical equation with the real values of D and L could not always fit the evolution of $\hat{Z}$ described by the data. That's why we realized a regression on D, L and O (model *DLO*) so we could consider potential noise in all dimensions and adjust the parameters to fit the data. For the two nonlinear models, the parameters are estimated thanks to the conjugate gradients method for the penalized least squared error optimization. The penalty term provides interpretable values for the estimated parameters (**D** and **L** close to D and L respectively and **d** and **O** close to zero)

Figure 7 shows an example of regression on δx or Z by the different models. Here, experimental data drift clearly from Δx and Z towards the crossed disparities. The absolute disparity perception is biased, however relative disparity relationship between targets at different depth are for the most part respected. In this case the observer's fixation for the highest disparity stray away from the Δx the most. The three models differ mostly for high disparity estimation. We can therefore choose which model is best adapted to each observer. Such regression gave us one set of parameters per model and per position. Deleted positions were interpolated from the nearby positions. With these sets coupled to positions coordinates maps were built using cubic spline interpolation between the 9 calculated parameters values. Each model was then represented by the spatial maps of its parameters (Figure 8).

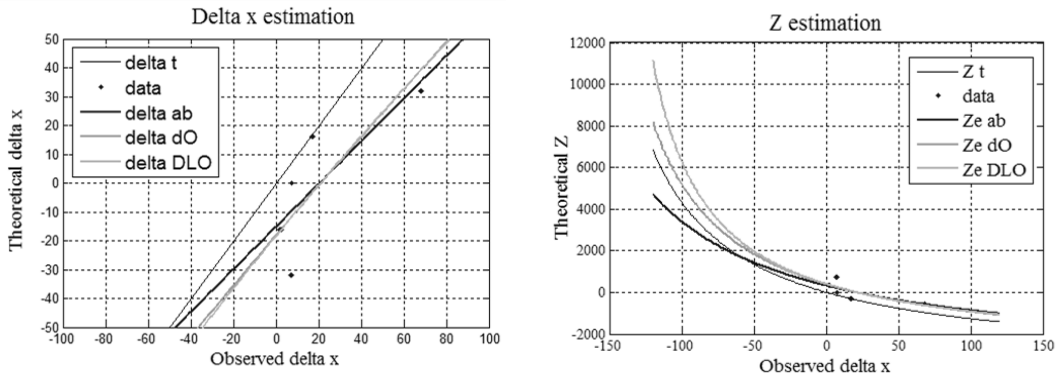

Figure 8. Representation of the different models on both δx and Z estimation in pixels for one position. The black line represents the theoretical evolution of δx and Z. The different gray scaled lines represent the different models following the raw data (black dots).

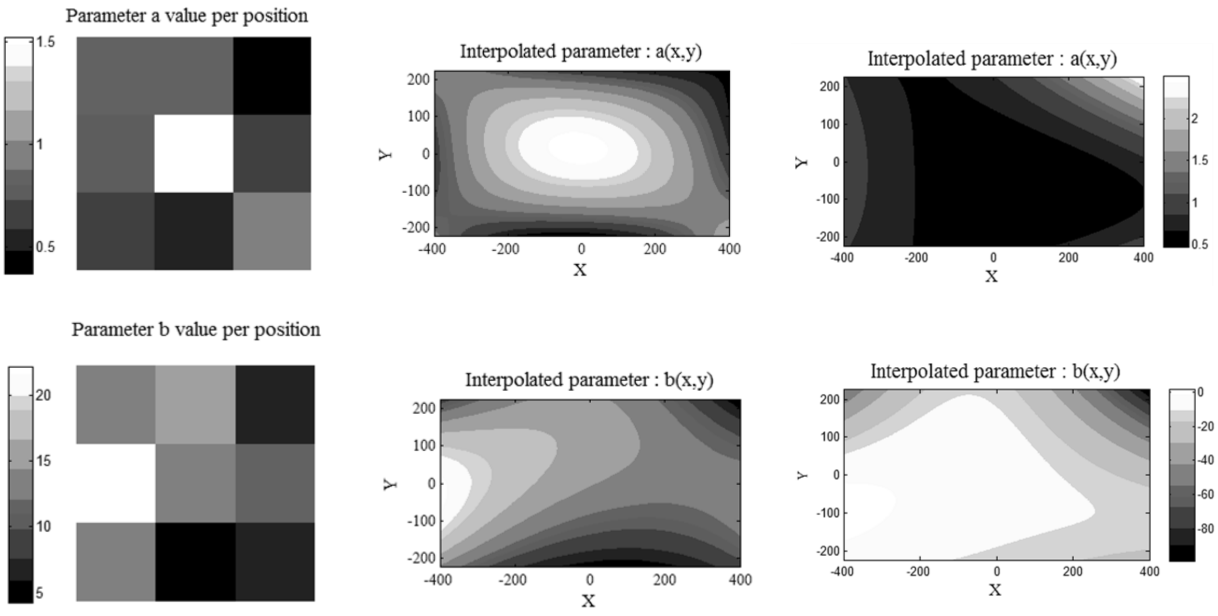

Figure 7. (Left,center) Example of the two interpolated parameters maps of the *ab* model built from the calibration of one subject. (Right) The same parameters map for another observer.

Each observer had his own map of parameter for each model determined according to his own ocular behavior (Figure 8). We found large variability between observers and the variability of the parameters across the 2D space showed that to take into account the spatial position is determinant for 3D gaze tracking.

We applied the parameters corresponding to the gaze coordinates of the last third of the calibration data(*Learned positions and learned disparities*) and we calculated its RMS error to check the models accuracy to predict the data. Then we used these same parameters maps on the *new positions and learned disparities* data from the test part and calculated the RMS error of the estimation. This result represents the efficiency of the spatial interpolation since disparities are the same for the two target positions. Afterwards, as for calibration, we built new models based on the test part (*Learned new positions, learned disparities*) and calculated the RMS error for the learned and the new disparities. Positions being the same between learned and new disparities, we tested this time the 3D interpolation efficiency of our models. Finally, we used the models based on calibration on the *new positions and new disparities* data to check the spatial and depth crossed interpolation efficiency of our models (Figure 9).

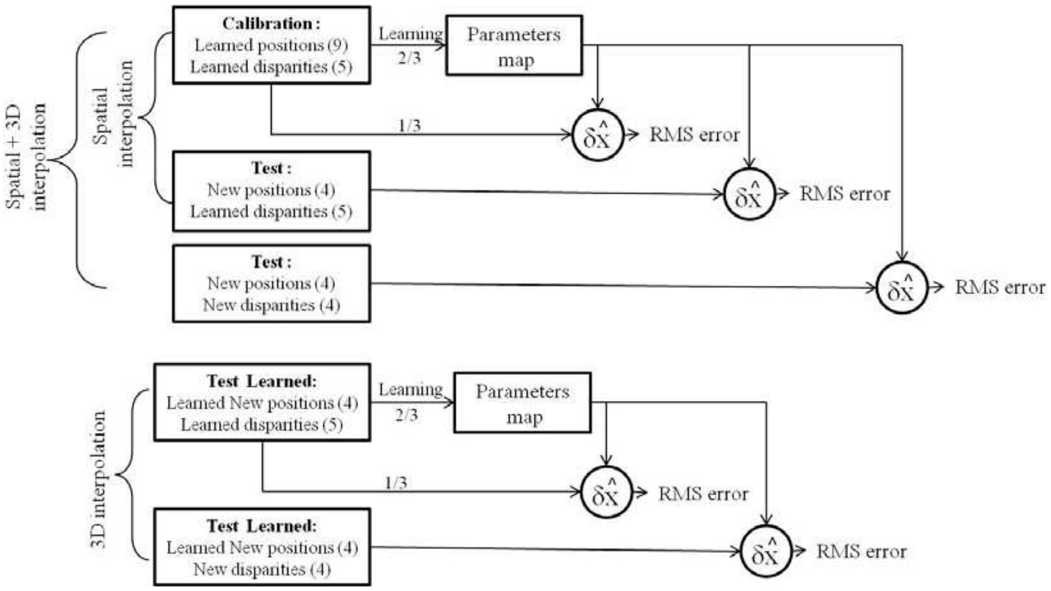

Figure 9. Schematic representation of the treatments realized on the different conditions. We compared the different conditions and models with the calculated RMS error.

3. RESULTS

The first aim of this research was to test which type of stimulus, artificial or natural, was more adapted to perform efficient 3D calibration using 2D eye tracking. We compared the reliability of the disparity recorded when viewing natural and artificial stimuli that displays target at several positions in the 3D space.

3.1. Natural vs. Artificial: Data extraction

First, it was important to note that most observers (5 out of 7) reported after the experiment that it was easier for them to fixate the target positions for the natural stimuli rather than the artificial stimuli. Moreover, some of them reported that although they perceived 3D artificial stimuli, it was in some cases difficult for them to perceive the depth plane of the target.

Table 1 reports several errors reflecting the fixation instability (see §2.5 for the procedure explanation). "Missing data" represents the percentage of missing data if we take into account one or two iterations. As expected to repeat twice each target position reduces the probability of missing data. It is important to note that the proportion of missing data varies between observers despite the fact that observers were checked for similar stereo acuity thresholds. "Uncertain" data correspond to the percentage of trials for which we obtained different δx for the two iterations for the same target position (difference above 30 pixels). "Outliers" data represent the percentage of data that was removed after the calibration.

Stimuli Observer	Natural stimuli				Artificial stimuli			
	Missing (%) 1 iteration	Missing (%) 2 iterations	Uncertain (%)	Outliers (%)	Missing (%) 1 iteration	Missing (%) 2 iterations	Uncertain (%)	Outliers (%)
AN001	29	9	28.9	6.7	40	23	35.6	6.7
AN002	12	5	17.8	2.2	12	5	37.8	4.4
RA003	25	0	15.6	2.2	9	3	37.8	2.2
NA004	0	0	48.9	0	0	0	48.9	6.7
FR005	3	0	44.4	8.9	7	0	44.4	17.8
AL006	0	0	6.7	0	18	7	17.8	2.2
AL007	12	0	24.4	0	25	7	20	2.2
Mean	11.6	2	26.7	2.85	15.9	6.4	34.6	6

Table 1 Calibration data quality. For the two types of stimuli, Natural and Artificial, different measures are reported for each observer: the missing data (considering the first iteration), the missing data (considering the two iterations), uncertain data and outliers.

We ran an ANOVA with two within –subject factors: the type of stimulus (Natural / Artificial) and the type of error (missing 1 iteration / missing 2 iterations / uncertain / outliers). We observed a significant effect of the type of stimulus (F (1, 18) =20.01; p<0.005): there were less data errors when using natural stimuli (10.8%) compared to artificial stimuli (15.7%). We also observe a significant effect of the type of error (F (3, 18) =13.45; p<0.001). As expected, there were less missing data when taking into account two iterations (F (1, 6) =13.53; p=0.01). Moreover, the proportion of uncertain data was higher than the proportion of missing data 2 (F (1, 6) =22.25; p<0.005) and the proportion of outliers (F (1, 6) =43.9; p<0.001). We did not observe any significant effect of the interaction (F (3, 18) =0.38; ns).

Hence, the conclusion of this first analysis was that using natural stimuli observers better manage to correctly fixate the different target positions. Moreover, it might be important to obtain more data by repeating twice (2 iterations) each target position. In fact, if we had used only one repetition, we would have lost 11.6% of data for natural stimuli whereas by using two repetitions, we only lost 2% of the data. For the further analyzes, the outliers were removed.

3.2. Natural vs. Artificial: Raw data analysis

We calculated, for all models, the RMS error between each recorded δx during correct fixations and their corresponding true disparity Δx for different testing target positions (Figure 9).

Model performances were highly observer dependent. For some observers, the measured disparities were very close to the true disparity. Hence, for these observers, when plotting the measured disparity as a function of the theoretical disparity, we obtained a line (AL006, AL007 and NA004). For other observers, there was a global drift toward the crossed or uncrossed disparities that was stable for the different positions (AN001, FR005, RA003 and AN002). For 3 observers, this bias is toward the uncrossed disparities, observers tend to fixate the target farther than its real depth. This effect is accentuated by visual fatigue which causes a progressive release toward the relaxed state where the two lines of sight are parallel. Only one observer (RA003) exhibits the inverse bias towards the crossed disparities.

Before analyzing model performance, we focused on the raw data for all observers according to the target positions. We ran an ANOVA on the RMS error measure with two within subject factors: the stimulus type (Natural / Artificial) and the target positions (*Learned pos Learned disp / New pos Learned disp / New pos new disp*). A significant effect of the target position was found on raw data (F (2, 12) =4.66; p<0.04). The RMS error was higher for the new positions and new disparities compared to the learned positions and learned disparities (F (1.6) =5.96; p=0.05). No significant effect of the stimulus type was observed (F (1, 12) = 1.72; ns), however on average natural stimuli always showed smaller RMS error than artificial stimuli. Data tended to be closer from theoretical values and more stable (standard error smaller) for natural stimuli than for artificial ones (Figure 10).

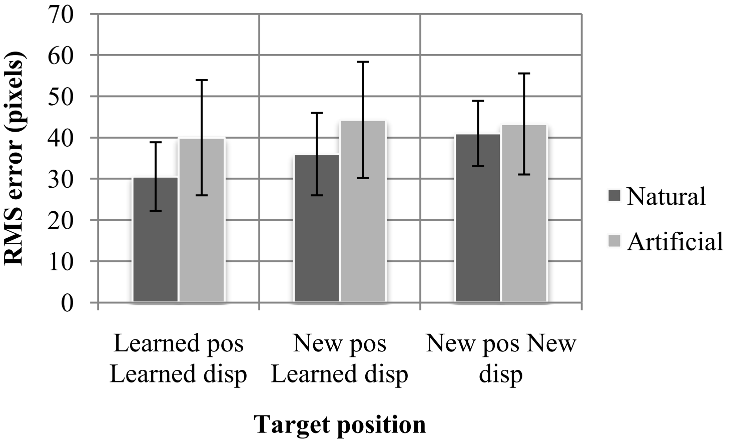

Figure 10. RMS error calculated on raw data according to the type of stimulus and the target positions. Error bars are standard errors

Although no significant difference was found between the stimulus types, the criterions of stability and consistency and the raw data analysis clearly favor the natural stimuli. The following analyses are presented only on the natural stimuli.

3.3. Natural: Models evaluation

We ran an ANOVA analysis with two within subject factors: the target position (*Learned pos Learned disp / New pos Learned disp / New pos new disp*) and the type of model (*Raw / ab / dO / DLO*) (Figure 11). We observed a significant effect of the target position (F (2, 36) = 20.8; p<0.001). As expected, models perform clearly better on learned positions and learned disparities. Moreover RMS error was higher for the new disparities than for the learned ones (F (1.6) =13.23; p<0.01). The type of model did not have any significant effect (F (3, 36) = 2.12; ns). Both factors interact (F (6, 36) = 2.5; p<0.05). The linear model *ab* was more efficient on the learned position and learned disparities. This might be explained by the fact that this model was computed using the same set of data and our data extraction used a linear regression (see 2.5). However we observed on the "*Learned new positions and learned disparities*" that this advantage of the linear model seemed to not be related to the selection process (see 2.4); in fact, the testing part had only one iteration and did not endured such a selection. We also noticed that the spatial interpolation was the best for the nonlinear *DLO* model (RMS=20.15). In fact if there was a god spatial interpolation, models should be efficient to predict target positions at *new positions and learned disparities*. Moreover, it produced the best results for the new disparities tested both with the calibration model and the model built on the testing part.

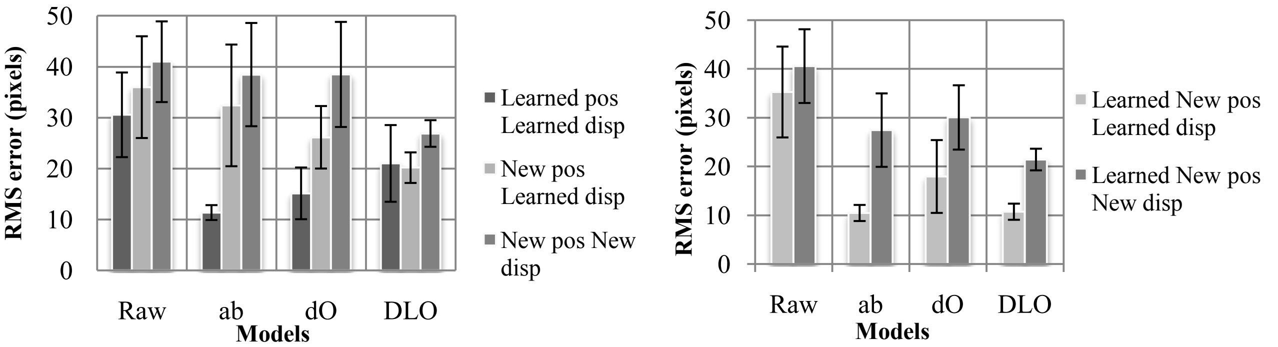

Figure 11. Left: Representation of global RMS error according to target position and models in the natural condition. Pos = Positions. Disp= Disparities.Error bars show standard errors. Right : Global RMS error of two new target positions where models are built directly from the new positions and learned disparity.

Actually, when we look at the individual results, the models efficiency compared with raw data is more observers dependent. Three subjects show an improvement of the error with model parameters in comparison with raw data.

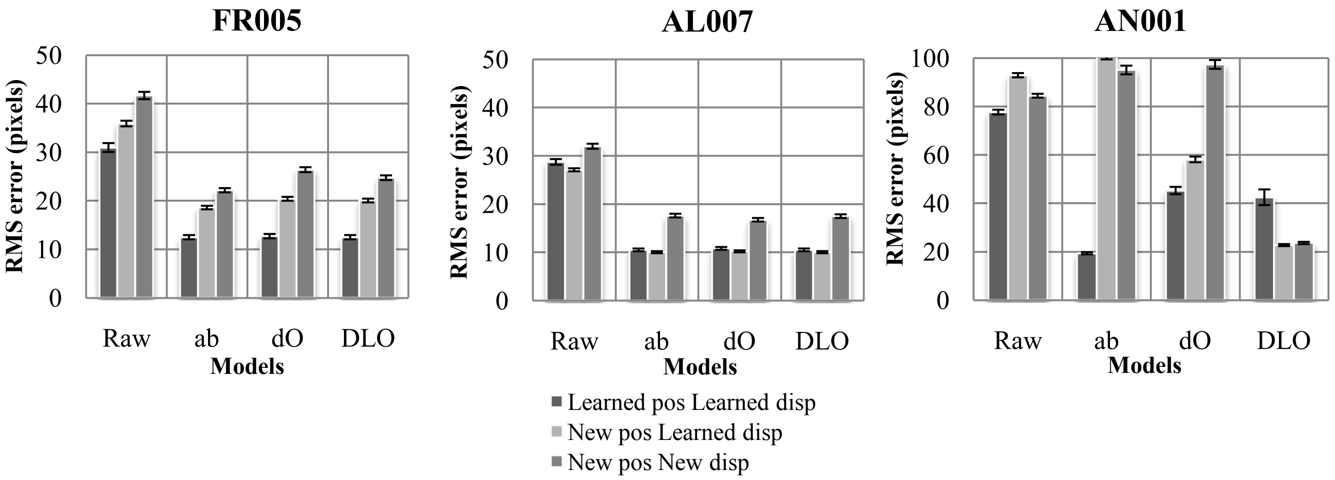

Figure 12. Representation of RMS error of data according to the target position for raw data and the three tested models. Three observers results are described here. Models show an improvement on these participants data in comparison with raw data.

For two of them (FR005 and AL007), the three models are equally efficient on new positions (Figure 12). They permit the decrease of the error on the *new positions and learned disparities* of 63% for the first and 48% for the second. The improvement is slightly smaller for *the new disparities on the new positions* (46% and 44% respectively). AN001 data were biased such as most of the models were not efficient to predict the new positions, however the non-linear model DLO was able to decrease the error of 75% and 73% for *learned and new disparities of new positions* compared to raw data.

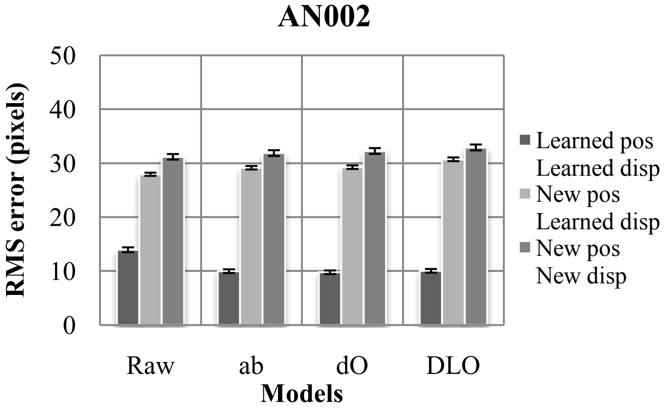

Figure 13 RMS error of one observer where models do not decrease the error compared to raw data.

On the other hand, there are cases where models do not manage to decrease the error from raw data. Figure 13 shows an example of this case.

4. DISCUSSION

In this paper, our main issue was to determine what type of stimuli might be the most appropriated to perform a 3D calibration using a 2D eye tracking. Our approach is similar to the concept used during 2D calibration and also it depends on the 2D tracking accuracy. While the 2D calibration assure the correspondence between the screen coordinates and the gaze coordinate, the 3D calibration attempt to follow the correspondence between the disparity displayed and the vergence pattern to track accurately the 3D gaze. We know that monocular depth cues can affect the vergence state REF. Hence, vergence eye movements in an experiment using natural images are necessarily biased. Although we chose our natural images with limited monocular depth cues, the scene architecture still influence the vergence eye movements of the observer. The use of two different images to represent close and far object could better consider this bias. Although, no significant difference was observed on the model efficiency, both subjective evaluation of observers and quality criterions confirm that natural stimuli should be used to calibrate the depth of gaze when viewing 3D stimuli. Natural stimuli elicited more repeatable fixation as we obtained significantly lesser uncertain values than for the artificial stimuli. Besides, we lost more data for the artificial stimuli reflecting a higher difficulty to gaze at 3D targets. The data consistency also favors the natural stimuli. The higher proportion of outliers makes the artificial stimuli condition data unreliable and increases the risks to have to discard data. And models need as much valid data as possible to follow the observer behavior. Consequently, natural stimuli might be used to efficiently perform 3D calibration. However because perception quality of stereoscopic stimuli through time varies from one observer to another, side effect such as headache, ocular dryness and visual fatigue[12] may appear during the viewing time and lead to a disturbed vergence state. So, in an experiment where the observer has to perform a 3D task after a 3D calibration, two iterations increase the risk of biases caused by these symptoms. Consequently, only one iteration of each target should be used during the calibration part.

For some observers, we did not expect that the absolute disparity extracted from vergence would be so far from the theoretical gaze position. So, in a 3D exploration context, we might wonder if the recent researches about 3D saliency based on monocular gaze position are valid. These studies [15,16] used only the guiding eye data coupled to a disparity map computed beforehand. To determine the depth gazed at, they looked at the gaze location on the disparity map. However, we saw in our experiment that even the guiding eye gaze location can be influenced by vergence. In some case where vergence state is shifted from the theoretical state, the gaze location extracted from only one eye might be wrong and the corresponding disparity on the map too. We think, that in a 3D experiment, in order to retrieve the depth and the object fixated, the two eyes should be considered since vergence eye movements influence the gaze location recorded.

According to Ogle[17–19], the stereoscopic depth perception can be divided into two categories based on the amplitude of the disparity and the perception of the observer : the quantitative and the qualitative stereopsis. While in the quantitative stereopsis, the observer's perception is directly proportional to the binocular disparity displayed, in the qualitative stereopsis range of disparity, perception reach a plateau where it does not follow the disparity anymore and where the observer cannot estimate reliably the disparity amplitude. Consequently, it is important that the disparity displayed in the calibration and the future 3D task remains in the quantitative range where vergence eye movement still manages to fusion the stereo images and for which its relation with binocular disparity is still linear. That is why we put the largest disparity recorded in the calibration at approximately half a degree. Visual comfort is equally a factor since observers complained at higher disparities especially for the crossed disparity at d=40.

We can notice that the RMS error for new disparities is significantly higher than for learned disparities on new positions. Besides, when we made the models directly on new positions and learned disparities we noticed that the extrapolation of models to new disparities have some issues since the error is significantly higher for new disparities than for the learned ones on the same position. We could explain this higher error by the fact that stereo threshold increases with absolute disparity amplitude ([20]for review). The vergence eye movements would be prone to more variation.

On average, the nonlinear model *DLO* is more efficient to retrieve the depth gazed at both the spatial and 3D interpolation level. Our mean error level is similar to other studies having used close paradigm (19.3arcmin of error for our model on learned disparities to be compared with 25.6arcmin for Duchowky's fitting model[5] and 28.8arcmin of error for our model on new disparities), despite the fact that they did not tried to generalize their model to others disparities than the learned ones. With such an error level we cannot track accurately the absolute disparity level perceived during a free viewing experiment, however the 3D scan path will be available since we observed that relative disparity interaction were preserved in the vergence extraction. This will allow us to better understand how humans explore a 3D image and for example, confirm if features salient in 2D image stay as much salient in 3D[16].

However when we analyze each subject individually, we notice that model efficiency depends highly on the observer. Some people tend to be extremely precise and consistent within their 3D fixations whereas others have an ocular behavior more erratic, making estimation more difficult. By building several distinct models, we can choose the best adapted model to each observer's behavior.

The perception ability seems to not be the source of this inter-subject variability since the observers' stereo acuity was similar. So this inconsistency necessarily comes from the vergence eye movements. The fact that some observers found quite difficult and tiring to look at 3D stimuli becomes an important issue to track accurately the gaze in 3D from the vergence eye movement.

During an experiment using 3D images such a 3D calibration will permit to track the 3D gaze providing that models can be kept up to date with the vergence state of the observer. Hence, multiple calibration session have to be considered all along the experiment. So it's important to consider as a whole, the 3D calibration procedure and the 3D target experiment when designing this 3D target experiment. Having an efficient, flexible and fast 3D calibration procedure adapted for each subject is then essential for the reliability of the 3D target experiment.

ACKNOWLEDGMENTS

The work presented in this paper was supported by the MOOV3D project, via the MINALOGIC and IMAGINOVE competitive clusters, and Rhône-Alpes region in France.

REFERENCES

[1]. Cutting, JE., Vishton, PM., [Perception of Space and Motion], Academic Press, San Diego, Chapter 3, 69-117 (1995).

[2] Wheatstone, C., "Contributions to the Physiology of Vision.--Part the First. On Some Remarkable, and Hitherto Unobserved, Phenomena of Binocular Vision", Philosophical Transactions of the Royal Society of London, 128, 371-394 (1838).

[3] Gonzalez, F., Perez, R., "Neural mechanisms underlying stereoscopic vision", Progress in Neurobiology, 55(3), 191-224 (1998).

[4] Fincham, EF., Walton, J., "The reciprocal actions of accommodation and convergence", J Physiol. 137(3), 488-508 (1957).

[5] Duchowski, AT., Pelfrey, B., House, DH., Wang, R., "Measuring gaze depth with an eye tracker during stereoscopic display", Proceedings of the ACM SIGGRAPH Symposium on Applied Perception in Graphics and Visualization,. 15–22 (2011).

[6] Essig, K., Pomplun, M., Ritter, H., "A neural network for 3D gaze recording with binocular eye trackers", International Journal of Parallel, Emergent and Distributed Systems, 21(2), 79-95 (2006).

[7] Kwon, Y-M., Jeon, K-W., Ki, J., Shahab, Q. M., Jo, S., Kim, S-K., "3D gaze estimation and interaction to stereo display", The international Journal of Virtual Reality, 5(3), 41-45 (2006)

[8] Schor, CM., Wood, I., "Disparity range for local stereopsis as a function of luminance spatial frequency", Vis Res,.23(12), 1649-1654 (1983).

[9] Westheimer, G.,"Cooperative neural processes involved in stereoscopic acuity", Exp Brain Res., 36(3), 585-597 (1979).

[10] Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., Weijer, J., [Eye tracking: A comprehensive guide to methods and measures], Oxford University Press. (2011).

[11] Masson, GS., Busettini, C., Miles, FA., "Vergence eye movements in response to binocular disparity without depth perception", Nature., 389(6648), 283-286 (1997).

[12] Lambooij, M., Fortuin, M., Heynderickx, I., IJsselsteijn, W., "Visual Discomfort and Visual Fatigue of Stereoscopic Displays: A Review", Journal of Imaging Science and Technology., 53(3), 30201-1-30201-14 (2009).

[13] Neveu, P., "Vergence tracking: a tool to assess oculomotor performance in stereoscopic displays", J. of Eye Mov. Res., 51(1), 1-8 (2012).

[14] Pobuda, M., Erkelens, CJ., "The relationship between absolute disparity and ocular vergence", Biol. Cybern., 68(3), 221-228 (1993).

[15] Liu, Y., Cormack, LK., Bovik, AC., "Natural scene statistics at stereo fixations", Proc. ETRA ’10, 161–164 (2010).

[16] Jansen, L., Onat, S., König, P., "Influence of disparity on fixation and saccades in free viewing of natural scenes", J Vis., 9(1) (2009).

[17] Ogle, KN., "On the limits of stereoscopic vision", J. of Exp. Psy., 44(4), 253-259 (1952).

[18] Ogle, KN., "Disparity limits of stereopsis", Arch Ophthalmol, 48(1), 50-60 (1952).

[19] Ogle, KN., "Precision and Validity of Stereoscopic Depth Perception from Double Images", J. Opt. Soc. Am., 43(10), 906-913 (1953).

[20] Wilcox, LM., Allison, RS., "Coarse-fine dichotomies in human stereopsis", Vis Res, 49(22), 2653-2665 (2009).