Dynamic Eye Convergence for Head-mounted Displays Improves User Performance in Virtual Environments

Andrei Sherstyuk *
University of Hawaii

Arindam Dey †
Magic Vision Lab.
University of South Australia

Christian Sandor ‡
Magic Vision Lab.
University of South Australia

Andrei State §
InnerOptic Technology Inc.
and University of North Carolina at Chapel Hill

Figure 1: *Extreme close-ups in Virtual Environments are a challenge for parallel eyes. It is nearly impossible to fuse the images of the butterfly into a single stereo view, while the hand is relatively easy to fuse. Distance to butterfly 16 cm, distance to hand 70 cm.*

Abstract

In Virtual Environments (VE), users are often facing tasks that involve direct manipulation of virtual objects at close distances, such as touching, grabbing, placement. In immersive systems that employ head-mounted displays these tasks could be quite challenging, due to lack of convergence of virtual cameras.

We present a mechanism that dynamically converges left and right cameras on target objects in VE. This mechanism simulates the natural process that takes place in real life automatically. As a result, the rendering system maintains optimal conditions for stereoscopic viewing of target objects at varying depths, in real time.

Building on our previous work, which introduced the eye convergence algorithm [Sherstyuk and State 2010], we developed a Virtual Reality (VR) system and conducted an experimental study on effects of eye convergence in immersive VE. This paper gives the full description of the system, the study design and a detailed analysis of the results obtained.

CR Categories: I.3.7 [Computer Graphics]: Three-dimensional Graphics and Realism — Virtual Reality I.3.6 [Computer Graphics]: Methodology and Techniques — Interaction techniques

Keywords: stereoscopic vision, hand-eye coordination.

*e-mail: andreis@hawaii.edu
†e-mail: aridey@gmail.com
‡e-mail: chris.sandor@gmail.com
§e-mail: andrei@cs.unc.edu

I3D 2012, Costa Mesa, CA, March 9 – 11, 2012.
© 2012 ACM 978-1-4503-1194-6/12/0003 $10.00

1 Introduction

Multiple studies in experimental neurophysiology tell us that human eyes always focus and converge on objects or locations that are associated with the current task: a hand, a tool or a location of tool application [Biguer et al. 1982]. This automatic eye fixation on the object of interest brings that object into the center of the visual field. When projected into the eye, the object's image falls onto fovia, a special area of the retina, which has the largest density of photoreceptors, and, consequently, the highest spatial resolution. Thus, automatic eye convergence ensures that both left and right retinal images of the object will have the best possible quality.

In virtual environments, the human gaze was also shown to be task-oriented [Ballarda and Hayhoea 2005; Rothkopf et al. 2007]. Stereoscopic vision in VE, often implemented with head-mounted displays (HMDs), provides important clues about objects' position and orientation. HMD's left and right channels represent views as seen by virtual cameras, co-located with the user's real eyes. However, in most VR systems both cameras are attached to the virtual head objects which is controlled by a single motion sensor. The cameras are set to converge at some predefined distance, which can be only a few feet away from the viewer or, more often, at infinity. In addition, the focal distance is also fixed in most HMDs. Because of such rigid settings of the display hardware, the image pairs produced by the virtual cameras will significantly differ from images that would form in the real eyes, whenever the objects of interest are located farther or nearer than the HMD's native convergence distance. Figure 1 demonstrates the problem. In this stereo-pair, rendered with fixed parallel cameras, the closely located butterfly appears in the opposite sides of the view frames, making them very hard to fuse into a single stereo-view.

We can identify the following problems, related to lack of flexible convergence in HMD-based VR systems. When objects of interest are at close range to the viewer, use of parallel cameras:

- *Makes stereo imagery hard to fuse* because left and right object views appear on opposite sides of the viewing frames;

- *Breaks stereo vision easily*: if one object image is centered, the other easily goes off-screen;
- *Diminishes sense of presence*: when both images are on-screen, they appear close to the black display borders, which continuously reminds users that their visible field is restricted;
- *Affects all objects at close range* (within the hand's reach, including the hand itself), where most accurate rendering and most precise object control are needed.

The latter became a serious issue for using VR in automotive industry, where all objects inside a virtual car are in close range from the viewer [Moehring et al. 2009]. In general, use of parallel cameras (or cameras with otherwise fixed convergence distance) makes rendering scene- and gaze-independent, which never happens in real life: human vision is both gaze and context sensitive.

In this paper, we offer a solution how to improve the visual response of HMD-based VR systems, by simulating physical eye convergence in software. Our approach is based on the expectation that human eyes will mostly converge on the hand-object contact point while performing direct object access and manipulations; this hypothesis is well supported by experimental data. We present the dynamic eye convergence mechanism and evaluate it in a formal experimental study, demonstrating that our method improves user performance in tasks that require precise hand-eye coordination.

2 Previous work

The use of virtual convergence for see-through HMDs was first discussed and implemented by State and colleagues [State et al. 2001] for their augmented reality guidance system for medical procedures. The location of current user activity was guessed through heuristics that worked quite well for that application. There was no formal user study conducted. The authors of a similar approach to simulate convergence via camera rotation [Peli et al. 2001] built and evaluated a prototype system for a desktop CRT stereo display with shutter glasses. The point of convergence was forcefully moved around the scene during trials; the intention was expressed to use eye-tracking in order to locate the point of regard in real time. More recently, Sherstyuk and colleagues suggested using the virtual hand as a locator for the current fixation point in VR systems that use conventional, non-see-through HMDs [Sherstyuk et al. 2008]. In a user study, participants with hand-enhanced camera controls showed significant improvement in their use of a virtual hand and virtual tools, in the context of a medical simulator.

In this work, we aim to improve the stereoscopic imagery for interactive VR systems, by combining simulated eye convergence [State et al. 2001] with the idea of using the virtual hand to predict the location of the current point of user attention [Sherstyuk et al. 2008].

The method we present here focuses on stereoscopic imagery with known depth maps, that is computer-generated image pairs instead of photographic ones. Recent work by Didyk and colleagues [Didyk et al. 2011] introduced a novel image-based perceptual model and metric that applies to all types of stereoscopic imagery (and could be integrated with our method in the future, yielding a hybrid system). The paper also contains an excellent summary on depth perception in general, and on stereopsis in particular.

In addition to software solutions, there exists a separate body of research on improving certain aspects of stereoscopic rendering on hardware level. One of the recent results in that field [Liu et al. 2010] makes use of active optical elements for producing imagery at various focus depths. Alternative approaches involve translation of a relay lens inside the HMD [Shiwa et al. 1996] or microdisplays [Shibata et al. 2005]. Even with displays that are capable of accommodating for very close objects, the stereoscopic convergence problem remains. In order to support eye convergence on a physical level, an HMD needs to provide large nasal-side display areas, which is nearly impossible from engineering standpoint. Thus, we offer a software solution for that missing feature.

3 Dynamic eye convergence algorithm

In order to find the point of eye convergence, one needs to know the exact gaze direction, for each eye. In real life, the gaze direction is a combination of two rotations, the head's and each eye's. The similar situation exists in VR, especially if the display device has sufficiently large field of view (FOV) to allow wide eye rotations. Thus, for simulating camera convergence in real time, one would have to track eye gaze. Reliable eye tracking requires special hardware, such as temple-mounted electromagnetic sensors that detect rotation of the eyeballs or near-eye cameras tracking the pupil's location. Both approaches add significant complications to system configuration, and call for special calibration procedures. Even with gaze tracking in-place, an HMD supporting that would either have to provide an extremely large FOV, or possess displays that can rotate around the center of the user's eyeballs, such that they always present their images in line with the exit pupils, following the gaze direction.

We suggest an alternative approach to estimating gaze direction. Human visual field spans 180 degrees horizontally, for both eyes combined, with 60 degrees of stereo overlap. This allows enough room for active eye movements while the head direction remains fixed. To compare, the horizontal field of view of most commercially available HMDs ranges from 20 to 50 degrees. This effect is known as "tunnel vision" and is widely considered as one of the most objectionable features of HMD-based system. However, for our purposes, this visual impediment turns into an advantage. When viewing the scene through a narrow HMD frame, users are forced to rotate their head instead of and in addition to moving their eyes. Therefore, we propose to approximate the user's gaze direction by orientation of their head. (While previous work [Watson et al. 1997] has shown that head motion does not fully correlate with gaze direction, especially within the central $\pm$ 15° area of the field of view, only an approximate gaze direction is required for our technique.)

By excluding free eye rotations, that happen in all directions, from our implementation, we restrict the dynamic eye convergence mechanism to controlling the angle between the eyes' virtual cameras. The algorithm will rotate the cameras, maintaining the convergence distance as set by the location of the target object, in our case, the virtual hand. The target object will only move in the screen space horizontally, keeping its vertical position unchanged. This approach helps avoid possible perceptional conflicts caused by head rotations. In addition, users will not feel that they lost control over their virtual hand.

The algorithm. Dynamic eye convergence computations are executed in the main graphics loop, as listed below.

1. Check visibility of the target object (i.e., the hand) in cyclopic camera space. If the target is outside the viewable area, rotate both cameras to their default angles and return.
2. Find the target location in camera space (x, y, z), the azimuth $a_x = \arctan(x/z)$ and elevation $a_y = \arctan(y/z)$ angles.
3. Compute the convergence angle $A = \arctan(D/2z)$, where D is the camera separation distance, and z value is obtained in the previous step.

4. Compute the attenuation factor $f = (1 + s^2 d^2)^{-2}$, where d^2 is the distance to the target $d^2 = a_x^2 + a_y^2$ and s^2 is the parameter that controls the slope of the attenuation function, plotted in Figure 3. We used $s^2 = 0.28$, which practically nullifies convergence, when the target moves from the screen center farther than 10 degrees, in any direction.

5. Finally, rotate left and right cameras inwards by fA. The cameras must be facing in $-Z$-direction, separated by D.

Figure 2 demonstrates the results of the eye convergence mechanism, using a 1 cm sphere as a target object and outdoor settings. It is very hard, if possible, to fuse the spheres in the top stereo pair, rendered with parallel cameras. In the bottom pair, the left and right cameras converge on the target, placing it at the center of the screen. As a result, the spheres can be fused easily. Note, that the images of distant objects (here, the palm tree and its shadow) are separated further apart, resulting in diplopic appearance, in stereo view.

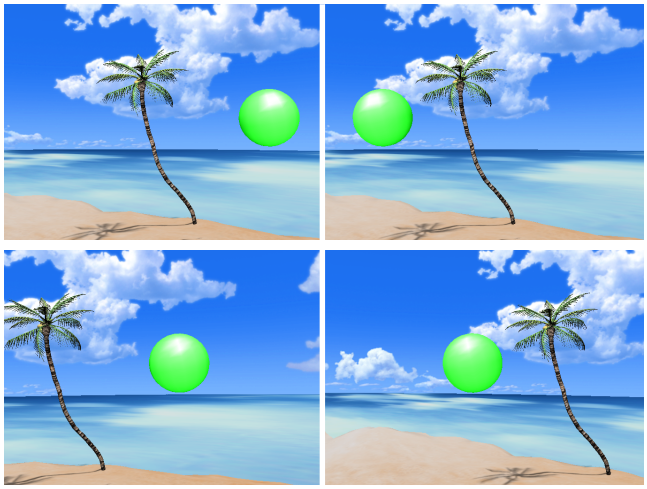


Figure 2: *The test scene, rendered with fixed (top) and converging cameras (bottom). Distance to target 13 cm, convergence 14.4 degrees. Sphere radius 1 cm, camera stereo separation 7 cm.*

The use of angular attenuation enforces our assumption on predicted gaze direction being equal to user head rotation. Also, it prevents the algorithm from taking full control over the horizontal position of the virtual hand in screen space. For example, if the user needs to place his or her hand outside the central viewing area, the attenuation factor will cancel the pending camera rotation and the hand will remain in its intended place.

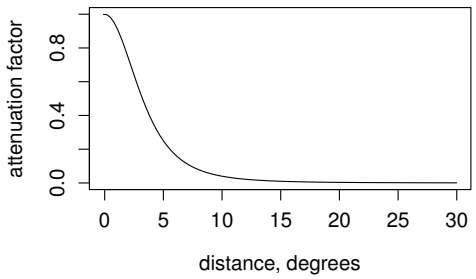


Figure 3: *Angular attenuation function. Convergence is strongest at the center and rapidly falls off towards the screen edges.*

The dynamic eye convergence is designed to remain active whenever users are performing near-field viewing tasks. For most operations that involve use of virtual hand or hand-held tools, the algorithm may be used in its original form. However, there are certain application-specific tasks and conditions, when the convergence mechanism may need additional modifications. Below we provide a few examples.

- For tasks that require bimanual operations, a point halfway between the two hands must be used as a new target location.
- When a hand, or a hand-held tool is operated as a pointer, or requires aiming, stereo rendering should be temporarily disabled by setting the camera separation distance to zero. Examples: selecting a distant destination for travel, shooting a hand-gun.
- A special case when the user is operating a tool that has visible effect on other objects. Examples: a fishing pole, a magic wand. If the object of interest is known, the system may converge on that object (for example, a bobber or a fish on the line). If the target is unknown or can not be localized at one point, convergence should be switched off.

4 Implementation and preliminary tests

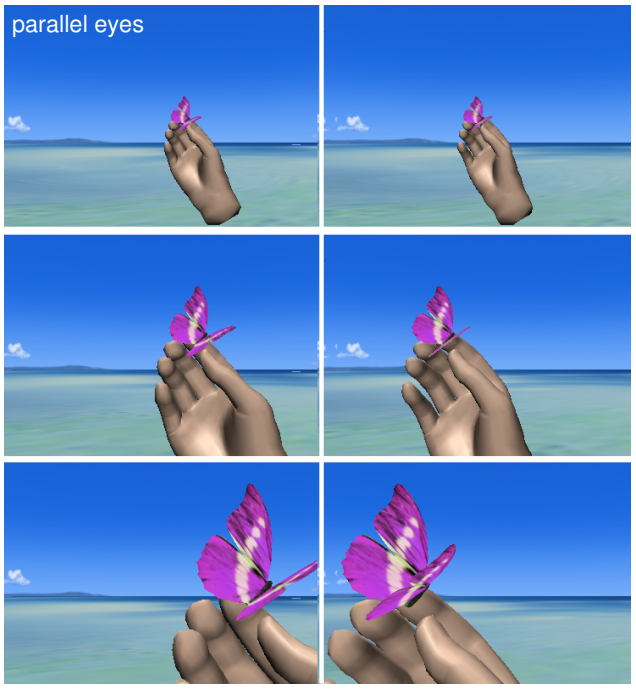
We implemented the eye convergence mechanism in *Flatland*, an open source 3D engine [AHPCC 2002], and tested it on a laptop, running in non-immersive mode. Both virtual cameras and the virtual hand were controlled with a mouse; the left and right views were displayed on a laptop screen, as a stereo pair. One of the authors, after some practice, learned to operate the system, while maintaining the stereo view by fusing the stereo pair, continuously.

In our system, users interact with virtual objects by pointing and touching them with the virtual hand. For these purposes, 1 cm invisible cubic shapes are attached to the tip of the index finger, on each hand, as shown on Figure 4. These cubes are used as probes for detecting and processing collisions; also, they serve as target objects for eye convergence. The virtual hands are implemented as deformable objects, driven by skeleton-based animations.

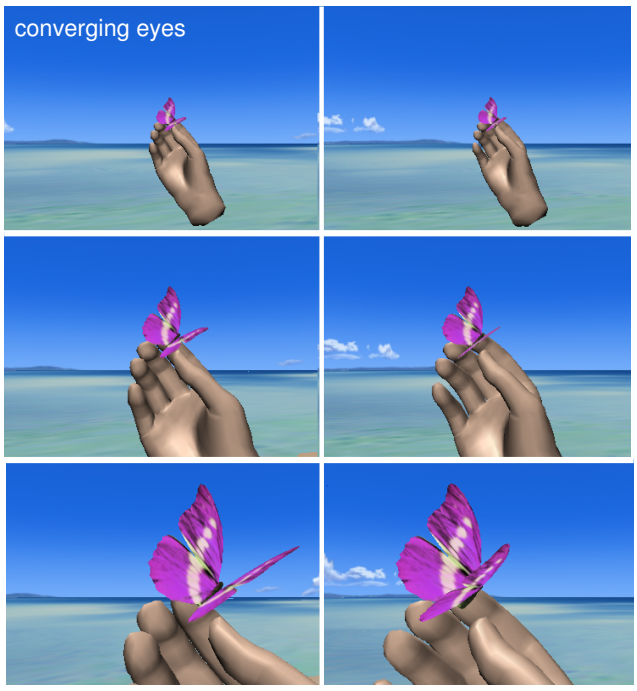
Figure 4: *The shape of the virtual hand is task- and context-sensitive. The hand assumes various poses using pre-recorded animation data, applied to its skeleton joints.*

In order to test the effects of dynamic eye convergence, a beach scene was used, with few static objects and a flock of butterflies flying around the user and making occasional stops. The task was to reach and capture butterflies by touching them with the index finger. Whenever the hand was in view, the cameras were converging on the index finger automatically. Upon a completed catch, the virtual hand assumed the "closed-hand" pose; the butterfly was attached to the hand, for close-up examination. After five seconds, the captured butterfly was released, the hand assumed the initial pose and the exercise continued. During repeated trials, we observed that

- Dynamic convergence does not produce disturbing or unpleasant sensations in non-immersive mode;
- Helps reduce diplopia (double vision) for close objects;
- Allows to reduce the effort required to fuse stereo views of objects positioned at varying depths.



(a) Fixed parallel cameras: as the hand moves closer to the viewer, its images slide towards the inner edge of the display frames.



(b) The cameras converge on the hand dynamically, according to the hand's position. As a result, the hand with the attached butterfly always remains centered.

Figure 5: *Pilot test: catching butterflies on a tropical beach, using (a) parallel and (b) converging cameras. Distance to hand: 60 cm, 30 cm, 15 cm.*

Figure 5 demonstrates the last result. Stereo pairs produced with parallel cameras require viewers to converge their eyes separately for each pair in order to achieve fusion. As a result, it is impossible to see the target objects in all three top images in stereo. On the contrary, the pre-converged bottom pairs can be viewed all together, because the objects of interest are centered and require the same convergence angle from the user. We leave it to the readers to perform this simple exercise.

5 Experimental study

In order to evaluate the proposed technique, we conducted an experimental study in immersive VE. The goal of the experiments was to collect and compare objective and subjective data on how dynamic eye convergence affects user performance in general and hand-eye coordination in particular.

System components. To create an immersive environment, Flatland was reconfigured to use a stereo HMD as an output device and a motion tracker for head and hand controls, in 6 degrees of freedom. For this study, we used a Canon VH-2007 HMD (resolution 1280×960 pixels, 60° horizontal, 47° vertical, 76° diagonal field of view). For tracking, a Flock of Birds system from Ascension was used, operating in standard 4 feet radius range. The system was installed on a single Ubuntu Linux PC. The content was rendered with the OpenGL and OpenAL APIs at 25 frames per second.

The participants. Fifteen healthy volunteers, with normal or corrected to normal vision, were invited to participate in this experiment. The experiment was organized as a within-subjects study, so each participant completed two sessions with eye convergence turned on and off. The order of sessions was counterbalanced. All participants were recruited among the students and faculty of University of South Australia. None of the participants had any previous experience with the experimental system.

The mission. Participants were asked to spend 10 minutes on a virtual beach, catching large tropical butterflies (wing span 7.2 cm). The objective of the exercise was to capture as many butterflies as possible, without "damaging" them. A completed capture was detected when the participant's index finger remained in continuous contact with a butterfly's wings for two seconds. Users were penalized for touching the butterflies bodies: such butterflies were marked as "damaged". Thus, to make a valid clean catch, users had to manipulate their hands very carefully. The butterflies were programmed to flutter around the user in 3 feet radius space in all possible directions, making occasional stops at randomized resting locations, for a few seconds. While in flight, the butterflies were oriented horizontally, with their heads pointing towards the next resting location. The butterflies' motion was a combination of linear movement towards the destination with added Perlin 3D noise, to simulate effects of the wind. Capture was allowed only while the butterfly was resting. During the trials, participants were seated on a chair. Such arrangement prevented them from wandering in VE and helped them focus on their task. Also, remaining seated helped the users avoid feeling constrained by the HMD cables (see Figure 6). Participants were not informed about the purpose of the experiment, but only the task they had to perform.

The procedure. Each participant completed two 10 minute sessions, with and without eye convergence, with a break of at least 10 minutes between the sessions to eliminate arm fatigue. Each new participant went through a brief calibration sequence. During calibration, the arm length and vertical position of the virtual hand were adjusted interactively by a VR operator. At this time, participants were fully immersed, wearing the HMD and a glove with

motion sensors attached. After the calibration, participants were given brief verbal instructions:

> Your goal is to capture as many butterflies as possible. To capture a butterfly, wait until it comes to a full stop, then touch its wings with your index finger, for a few seconds. Avoid touching the body, that will damage the butterfly.

Participants were not offered any additional time to practice the capture procedure; they had to learn it as they progressed with the mission. Immediately after calibration and instruction, the flock of virtual butterflies was set in motion, and the data collection started.

Data collection. For each session, the following information was collected and saved into a log file:

- Number of completed captures.
- Number of clean (undamaged) captures, as described above. Clean captures are a subset of completed captures.
- Number of failed captures. A failed capture was detected when the user's hand touched the target, but failed to maintain contact long enough to complete the catch. Failed captures are not part of any other set.
- Location of each capture event, in camera space.
- Hand jitter, defined as the length of the path traversed by the pointer object, while in contact with the butterfly. Low values of jitter indicate that (a) the hand fatigue is low and (b) the user has good control over their virtual hand.

From the collected data, we derived the success rate as the ratio of completed captures, including clean and damaged ones, to the sum of completed and failed captures. These characteristics of user activities constitute the objective metrics that we used in our analysis. All logged records were time-stamped at the frequency of the graphics loop (i.e., 0.04 sec), which allowed us to detect and analyze trends in user performance over time.

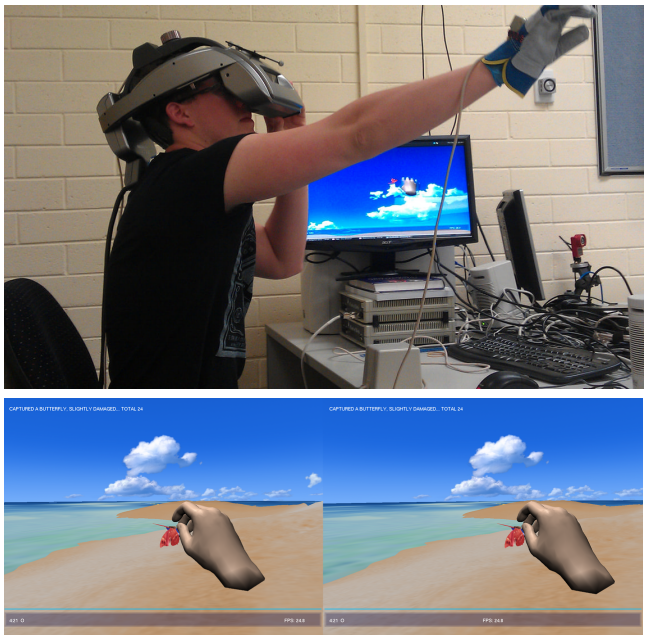


Figure 6: *Reenacted experimental session. Top: the participant is reaching out for a butterfly (photo used with permission). Bottom: the captured butterfly is attached to the hand.*

6 Subjective evaluation

In order to collect subjective feedback, participants were asked to complete a short survey about their experience, immediately after each session. The questions and the answers are listed in Table 1. We used a single table as no statistically significant differences were found between the groups. The answers were given on a 1-5 scale:

1 strongly disagree
2 disagree
3 neutral
4 agree
5 strongly agree

Table 1: *User evaluation results, combined for both conditions.*

Questions	Answers	
	mean	median
1. Catching butterflies was fun	3.9	4
2. The eye-hand coordination felt natural	3.8	4
3. It was easy to place the hand on the target	3.0	3
4. By the end of the session my eyes were tired	2.3	2
5. By the end of the session I felt dizzy	1.5	1

Because the mission implied and required active user participation, we offered the first question as a self-selection test to detect bored or frustrated users, that we might have had to exclude from the analysis. Fortunately, that turned out to be unnecessary: all participants were able to complete their missions and gave positive comments on the game-play (see Table 1, Question 1).

Another good outcome is that none of the 15 participants reported any discomfort nor dizziness (Questions 4 and 5). This result supports our preliminary findings that automatic eye convergence does not induce cyber-sickness, for specified experimental conditions and limited exposure time not exceeding 10 minutes.

Regarding questions 2 and 3 that addressed the subjective "feeling" of the eye convergence technique, the results came as a surprise. The users gave scores 3 and 4 (neutral or agree) for both conditions, with no statistically significant difference. The lack of preference towards any condition may be explained by the fact that we did not prime the participants to look for the differences between the conditions. That was done on purpose, to collect unbiased responses.

However, the collected objective data on user performance allowed us to observe significant differences between the two conditions. These results are discussed next.

7 Objective analysis of the effect of dynamic eye convergence on user performance

All user log files were processed by a custom parser, which extracted the numbers of complete captures, clean (undamaged) captures, failures, success rate and hand jitter values, for each participant. The recorded events were checked for bad data samples, produced by noise in the motion tracker system. For example, the parser discarded all events that involved hand-target contact, if the amount of hand jitter was excessively large at this moment. The parser also removed data samples recorded when users were playing with the butterflies, trying to "pet" and "slap" them. There were only a handful attempts of non-standard interactions with the targets, because the butterflies only responded to the correct capture procedure. After preprocessing, the output data arrays were formatted for the *R* statistical package [R Development Core Team 2009], which was used to process and display data in this work.

7.1 General outcomes

Table 2 shows the values of user performance, collected for all participants, over the whole duration of their sessions. The median and mean values are given per person; the value in the "total" line gives a sum over the whole group. Using Shapiro-Wilk tests, we confirmed that all our datasets are normally distributed, so we used paired-samples two-tailed t-tests to analyze the differences between the means of these two experimental conditions.

Table 2: *Summary of user performance values and standard error of means (SEM), under control and effect conditions.*

	Parallel N=15	Converging N=15	Significance
Complete captures			$t(14)$ = -0.901
median	39.0	40.00	p = 0.442
mean	38.4	41.27	
SEM	2.44	2.45	
total	576	619	
Clean captures			$t(14)$ = 0.188
median	23.00	21.00	p = 0.853
mean	23.33	22.73	
SEM	2.32	2.54	
total	350	341	
Failed captures			$t(14)$ = 3.944
median	24.00	11.00	p = 0.0014 $\star\star$
mean	23.93	13.07	
SEM	3.26	2.00	
total	359	196	
Success rate (%)			$t(14)$ = -4.481
median	62.82	79.66	p = 0.0005 $\star\star$
mean	62.92	75.79	
SEM	3.29	2.72	
Hand jitter (cm)			$t(14)$ = -1.335
median	6.62	6.82	p = 0.0.203
mean	6.28	6.68	
SEM	0.38	0.41	

The number of completed captures and the number of clean captures differ slightly between the groups, but the differences are not significant. Similarly, the hand jitter values appear almost identical.

The most significant difference was observed for the number of failed captures and the success rate. The success rate was calculated as the ratio of completed captures to the sum of completed and failed captures. The p-values are marked with a single asterisk ($\star$) for standard significance ($p < 0.05$) and a double asterisk ($\star\star$) for strong significance ($p < 0.01$). This notation is used in all tables in this paper.

Summary: users with converging cameras made only a little over half as many hand-eye coordination errors, compared with those using parallel cameras (196 vs 359). The reduced number of errors yielded a median success rate of 79.66%, compared to 62.82% for parallel cameras.

7.2 Analysis of user performance over time

To investigate how user performance changed over time, the log files were rescanned, collecting data samples into consecutive bins, with a time step set to 60 seconds. Each bin contained collective data for all participants under the same condition. For each time series obtained, linear models $y_i = a + bx_i + \epsilon$ were fit, using the least squares regression method.

The observed results for changes in collective performance are summarized in Table 3. The estimated start values (a) were obtained with very high probabilities, with all p-values less than 1e-05. The slopes (b), however, came out with various levels of significance, testing hypothesis $b \neq 1$. As Table 3 demonstrates, the numbers for clean captures, success rate, and hand stability did not show significant changes over time.

The number of complete captures per minute changed significantly for participants with parallel cameras, as indicated by its p_b value of 0.0084. They started at relatively low rate of 49.46 captures per minute and improved to 64.22, counting for all participants in the group. To compare, users with converging cameras started at 55.53 and ended at 67.11 captures per minute, which indicates that their capture rate remained higher for the whole duration of the exercise. The number of failures noticeably increased for the group with parallel cameras, from 26.53 to 43.51 per minute, compared with 16.80 to 21.84 for users with converging cameras. Figures 7 and 8 show plotted timelines for both groups. The linear models are also plotted, using solid lines for $p_b < 0.01$, dashed line for $p_b < 0.1$ and dotted line otherwise.

To summarize: user performance under the two experimental conditions was found to be significantly different. Dynamic eye convergence allowed participants to capture targets at a higher rate, by means of making fewer mistakes in positioning their virtual hands at the target objects. The relatively stable hand jitter values, observed under both conditions, suggest that the hand fatigue was not a factor in neither case. This suggestion is verified in the next section, where we examined individual performance of those users who learned to improve their hand stability over time.

Table 3: *Details on user performance on per-minute basis: estimated values (a), slopes (b) and trends. The end values in parentheses were estimated using a slope with insignificant p-value.*

	Parallel N=15	Converging N=15
Complete captures		
start value, a	49.46	55.53
end value	64.22	(67.11)
slope, b	0.0246	0.0193
slope, p_b	0.0084 $\star\star$	0.206
trend	increase	–
Clean captures, %		
start value, a	60.45	54.00
slope, b	0.0013	0.0034
slope, p_b	0.9120	0.7710
trend	–	–
Failed captures		
start value, a	26.53	16.80
end value	43.51	(21.84)
slope, b	0.0283	0.0084
slope, p_b	0.0704 .	0.3059
trend	increase	–
Success rate,%		
start value, a	64.95	76.38
slope, b	-0.0089	-0.0011
slope, p_b	0.37	0.876
trend	–	–
Hand jitter, cm		
start value, a	6.40	6.96
slope, b	-8.696e-05	-0.00059
slope, p_b	0.9210	0.4750
trend	–	–

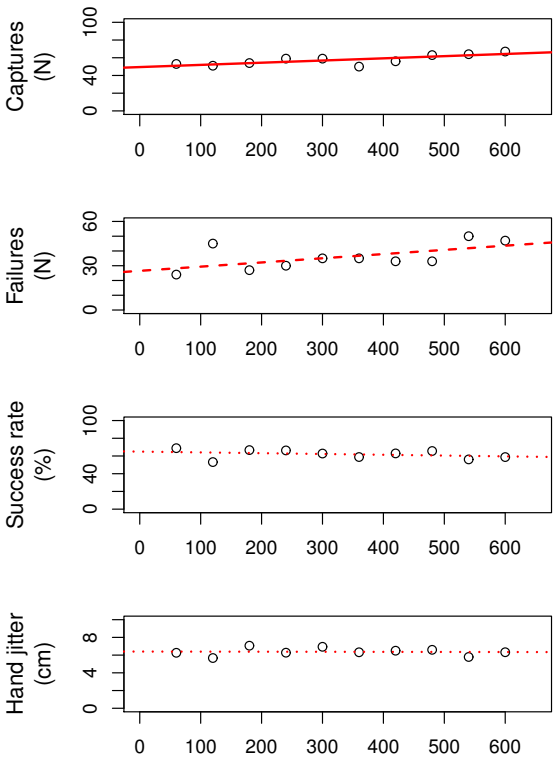


Figure 7: *Timeline plots for users with parallel cameras: simultaneous increase of number of captures and number of failures resulted in unchanging success rate (a = 64.95%). More details are provided in Table 3.*

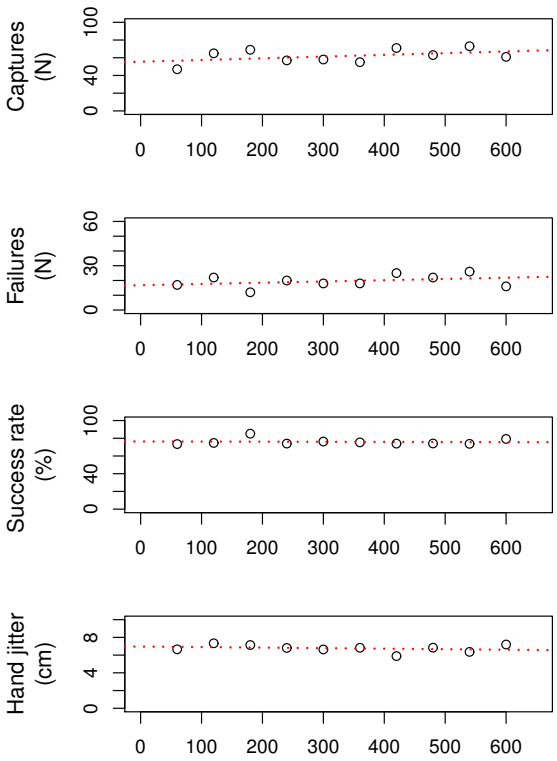


Figure 8: *User progress with converging cameras: both number of failures (a = 16.80) and success rate (a = 76.38%) remain steady. X-axis shows mission time, in seconds. Y-axis shows aggregate data, collected with 60 second time step.*

7.3 Hand stability and user performance

For direct hand manipulation tasks, such as picking objects, hand stability is of utmost importance. Thus, after conducting a collective group analysis of user performance, we examined user logs individually, fitting linear models for hand jitter values. In 18 out of 30 sessions, we observed a significant decline in hand jitter over time, split even between both conditions, as shown in Table 4. Evidently, these users realized the importance of hand stability for their task and were able to reduce jitter from 7.28 to 5.8 cm and 7.99 to 6.58 cm, for parallel and converging cameras, respectively.

However, improving hand stability did not result in better performance. Participants with parallel cameras showed a strong increase in number of errors, from 15 to 31 failures per minute. To compare, users with converging cameras proceeded at a stable rate of 13 failures per minute. Their group performance was also significantly higher, which is consistent with the finding discussed in Section 7.1. Because of smaller group sizes (N=9), we used non-parametric Wilcoxon rank sum test for comparison.

Basing on these results, we assert that user performance was not predicted by the hand stability values. Therefore, the superior hand-eye coordination demonstrated by participants with converging cameras can only be explained by the fact that these people had better viewing conditions than the control group.

7.4 On gaze direction approximation

As described in Section 5, the system recorded the coordinates of each completed capture, in camera space. We used this information to check our hypothesis that head rotation may be used to approximate gaze direction, discussed in Section 3. To do so, we calculated azimuthal and elevation angles for all 15 sessions with parallel (i.e., unaltered) cameras, shown in Figure 9. It turned out that most captures happened in the center of the viewing area. It is worth to note that the relatively wide field of view of our HMD (60° horizontal, 47° vertical), gave users a large room for eye movements within the visible area. Nevertheless, for precise hand-eye coordination, the users rotated their heads instead, keeping the target object at the center of view. This result confirms our assumption that under certain conditions, head rotation may sufficiently approximate gaze direction. This finding may be useful for those types of VR applications where fixation points are easily identifiable.

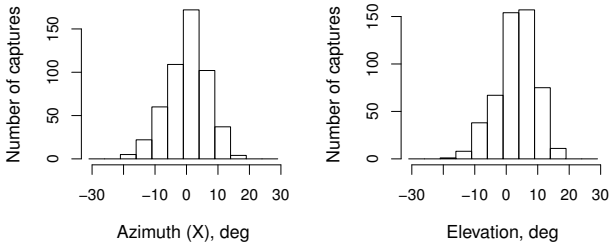


Figure 9: *Distribution of angles, from the (0,0,-1)-direction to all capture locations, collected from all users with parallel cameras. Total number of samples 514. The slight shift towards +X and +Y directions is likely due to the fact that all users were right-handed and approached the targets from the upper-right side.*

Table 4: *User performance in sessions with improving hand stability. Participants with parallel cameras show significant increase of errors over time (15 to 31), while the group with converging views remains stable (13). The right column shows Wilcoxon test results.*

	Parallel N=9	Converging N=9	Significance
Complete captures			p=0.199
median	33.00	40.00	
mean	36.67	42.67	
Failed captures			p=0.0467$\star$
median	28.00	11.00	
mean	26.78	13.22	
slope, b	0.02576	-0.00212	
slope, p_b	0.04382 $\star$	0.608	
start / end	15.6 / 31.1	–	
Success rate,%			p=0.0056$\star\star$
median	57.89	82.93	
mean	60.20	76.94	
Hand jitter, cm			p=0.2581
median	6.62	7.02	
mean	6.32	7.11	
slope, b	-0.00237	-0.00247	
slope, p_b	0.086 .	0.0346 $\star$	
start / end	7.28 / 5.84	7.99 / 6.58	

8 Discussion

We presented the dynamic eye convergence technique for head mounted displays and evaluated it in a formal experimental study. The proposed technique simulates the natural process of human eyes converging onto a current object of interest, which always happens in real life situations. Adding simulated convergence to virtual environments was implemented by dynamically rotating cameras towards the fixation point.

The experiment demonstrated that participants with dynamic eye convergence had significantly higher success rate in handling virtual objects, compared to participants with conventional parallel cameras. During the exercises, we monitored the level of user hand stability over time. The amount of hand jitter remained unchanged under both conditions for most people and improved for some of them. Therefore, we conclude that the difference in performance rates must be due to the visual, not motor, component of the hand-eye coordination process. This confirms that dynamic eye convergence has positive effects on the quality of viewing in immersive VE. Also, personal reports collected from the participants indicate that the proposed technique feels comfortable and does not promote cyber-sickness or eye strain, for specified experimental conditions.

We consider the obtained results as an evidence that dynamic eye convergence may become a helpful addition to HMD-based VR systems, where users are expected or required to manipulate objects at close range with their virtual hands. Medical simulators for training fine-motor skills, are one large class of such applications. Because the proposed technique does not require any special hardware, it can be easily implemented and evaluated for usability for most VR systems and applications, on a case-by-case basis.

9 Acknowledgments

The authors are thankful to Mary Whitton at the University of North Carolina at Chapel Hill for her help with experiment design and Canon Japan for the HMD used in this project. Also, many thanks to all the participants who volunteered their time and enthusiasm.

References

AHPCC. 2002. *Albuquerque High Performance Computing Center*. Flatland, http://www.hpc.unm.edu/homunculus.

BALLARDA, D., AND HAYHOEA, M. 2005. Modelling the role of task in the control of gaze. *Visual Cognition 17*, 6-7, 1185–1204.

BIGUER, B., JEANNEROD, M., AND PRABLANC, P. 1982. The coordination of eye, head and arm movements during reaching at a single visual target. *Experimental Brain Research 46*, 301–304.

DIDYK, P., RITSCHEL, T., EISEMANN, E., MYSZKOWSKI, K., AND SEIDEL, H.-P. 2011. A perceptual model for disparity. *ACM Transactions on Graphics (Proceedings SIGGRAPH 2011, Vancouver) 30*, 4.

LIU, S., HUA, H., AND CHENG, D. 2010. A novel prototype for an optical see-through head-mounted display with addressable focus cues. *IEEE Transactions on Visualization and Computer Graphics 16*, 381–393.

MOEHRING, M., GLOYSTEIN, A., AND DOERNER, R. 2009. Issues with virtual space perception within reaching distance: Mitigating adverse effects on applications using HMDs in the automotive industry. *Virtual Reality Conference, IEEE 0*, 223–226.

PELI, E., HEDGES, R., AND LANDMANN, D. 2001. A binocular stereoscopic display system with coupled convergence and accommodation demands. In *2001 SID International Symposium, Digest of Technical Papers*, SID '01, 1296–1299.

R DEVELOPMENT CORE TEAM. 2009. *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

ROTHKOPF, C. A., BALLARD, D. H., AND HAYHOE, M. M. 2007. Task and context determine where you look. *Journal of Vision 7*, 14, 1–20.

SHERSTYUK, A., AND STATE, A. 2010. Dynamic eye convergence for head-mounted displays. In *Proceedings of the 17th ACM Symposium on Virtual Reality Software and Technology*, ACM, New York, NY, USA, VRST '10, 43–46.

SHERSTYUK, A., VINCENT, D., AND JAY, C. 2008. Sliding viewport for interactive virtual environments. In *Proceedings of ICAT 2008: 18th International Conference on Artificial Reality and Telexistence*.

SHIBATA, T., KAWAI, T., OHTA, K., OTSUKI, M., MIYAKE, N., YOSHIHARA, Y., AND IWASAKI, T. 2005. Stereoscopic 3-D display with optical correction for the reduction of the discrepancy between accommodation and convergence. *Journal of the Society for Information Display 13*, 665–671.

SHIWA, S., OMURA, K., AND KISHINO, F. 1996. Proposal for a 3-d display with accommodative compensation: 3ddac. *Journal of the Society for Information Display 4*, 255–261.

STATE, A., ACKERMAN, J., HIROTA, G., LEE, J., AND FUCHS, H. 2001. Dynamic virtual convergence for video see-through head-mounted displays: Maintaining maximum stereo overlap throughout a close-range work space. In *Proceedings of the International Symposium on Augmented Reality (ISAR)*, 137–146.

WATSON, B., WALKER, N., AND HODGES, L. F. 1997. Managing level of detail through head-tracked peripheral degradation: a model and resulting design principles. In *Proceedings of the ACM symposium on Virtual reality software and technology*, ACM, New York, NY, USA, VRST '97, 59–63.