Modeling and Optimizing Eye Vergence Response to Stereoscopic Cuts

Krzysztof Templin[1,2]   Piotr Didyk[1]   Karol Myszkowski[2]
Mohamed M. Hefeeda[3]   Hans-Peter Seidel[2]   Wojciech Matusik[1]

[1]MIT CSAIL   [2]MPI Informatik   [3]Qatar Computing Research Institute
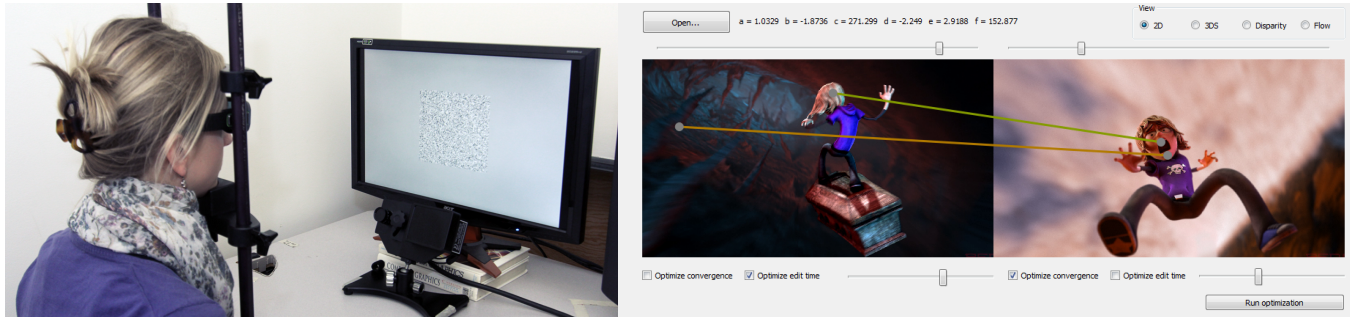
Figure 1: *Unexpected changes of disparity in stereoscopic 3D imagery, such as those introduced by cuts in S3D films, are challenging for the audience. Using an eye-tracker, we recorded eye vergence responses of 16 subjects to step-like changes in disparity* (left). *Then, a model of adaptation time was derived, and parameters for the average observer estimated. Having such a predictor enables, for example, optimization of film editing operations to best match the human depth adaptation abilities* (right). *The colors of the lines connecting points of interest before and after the cut visualize the corresponding vergence adaptation times. Pictures from* Dracula 4D *courtesy of Red Star 3D, www.redstar3d.com*

CR Categories: I.3.3 [Computer Graphics]: Picture/Image generation—display algorithms,viewing algorithms;

Keywords: binocular, eye-tracking, S3D

Links: ◉DL 🄿PDF ⬛WEB

Abstract

Sudden temporal depth changes, such as cuts that are introduced by video edits, can significantly degrade the quality of stereoscopic content. Since usually not encountered in the real world, they are very challenging for the audience. This is because the eye vergence has to constantly adapt to new disparities in spite of conflicting accommodation requirements. Such rapid disparity changes may lead to confusion, reduced understanding of the scene, and overall attractiveness of the content. In most cases the problem cannot be solved by simply matching the depth around the transition, as this would require flattening the scene completely. To better understand this limitation of the human visual system, we conducted a series of eye-tracking experiments. The data obtained allowed us to derive and evaluate a model describing adaptation of vergence to disparity changes on a stereoscopic display. Besides computing user-specific models, we also estimated parameters of an average observer model. This enables a range of strategies for minimizing the adaptation time in the audience.

1 Introduction

Over the past few years, stereoscopic 3D (S3D) technology has been constantly developing, and by now it has become ubiquitous. However, despite the significant improvements, not only in display devices, but also in image generation, capture and post-processing techniques, many consumers are still skeptical about the quality of current S3D content and the future of the technology itself. These concerns are usually related to naturalness, effortlessness, and overall appearance: S3D effect should not be a distraction.

The difficulty in S3D production is that it is not sufficient to produce two good images in place of one to arrive at a good stereoscopic effect [Zilly et al. 2011]. S3D is a strong illusion, since it isolates only one real-world phenomenon, failing to reproduce many others, a prominent example being the accommodation cue. This imposes numerous restrictions on the production process: the depth range and variation must not be too large, view-dependent effects need to be handled correctly, images carefully registered, and so on.

In this work, we are concerned with rapid temporal changes of disparity. Humans have a good understanding of the environment they observe and move through, a so-called "mental image", which enhances their capabilities in focusing on different objects [Finke 1989]. However, when the scene is merely a sequence of shots shown on a flat screen, it is easy to get confused or lose track of the point of interest, due to, among other things, unexpected changes of the location or the camera angle. Although less problematic in 2D, this can be challenging in stereoscopic 3D. In this context, an unpredictable and large change in disparity means that binocular fusion is lost, and a confusing double image is seen (diplopia). Moreover, the vergence system needs to quickly adapt to new conditions, in spite of the conflicting goal of the interconnected accommodation system. This has been identified as one of the sources of discomfort in stereoscopic viewing [Hoffman et al. 2008; Lambooij et al. 2009].

The Hollywood style of combining shots developed into a set of formal conventions that obey the dynamics of visual attention and control the continuity of space, time, and action. In modern movies cuts play the most important role (99% of all edits), while dissolves

and wipes have vanished almost completely. An extensive analysis by Cutting et al. [2011] shows that average shot duration over past 75 years has declined from ca. 15 s to ca. 3.5 s! Clearly, short shots increase the viewer engagement by forcing eyes to quickly follow newly appearing content. However, such accumulation of sharp cuts challenges the visual system by requiring seamless adjustment of vergence between many shots over a possibly wide range of depths. This requires a different approach to editing, e. g., ultra-short "MTV-style" shots need to be replaced by more slow-paced edits.

Nevertheless, modern movies are often simultaneously released in 2D and S3D, and one should not expect that directors, cinematographers, and editors will entirely give up on their artistic visions and style merely for the sake of S3D medium limitations. Instead, they apply different S3D post-production techniques to make depth transitions natural and effortless for viewers. Such manipulations range from simple depth manipulations and cross-dissolve types of cuts to more sophisticated transitions, where multiple sequences with gradually changing depth are combined [Owens 2013]. All these manipulations are time-consuming and expensive, as they are performed manually. For example, Owens [2013] pointed out that the editing of transitions was one of the most challenging steps in the post-production of the U2 concert recorded in stereoscopic 3D. Abrupt depth changes, well beyond the real-world experience, should be also expected in action computer games.

To address the problem of rapid depth changes, we propose to relate the transition quality to vergence adaptation time, instead of simpler disparity difference. We present a series of experiments with human observers, in which vergence responses were measured using consumer S3D equipment and a high-framerate eye-tracker. This leads to a simple model describing the vergence adaptation curve, given the initial and target disparities. The model allows for prediction of adaptation time after cuts, which facilitates its visualization and minimization. Impact of the optimization on the visual quality of S3D content is demonstrated in a separate experiment. To our knowledge, we are the first to apply such a principled approach to this problem.

In summary, we make the following contributions:

- measurements of vergence response to rapid disparity changes defined by initial and target disparities;

- derivation and evaluation of a model relating disparity change to vergence adaptation curve, along with average observer parameters;

- interactive tool for visualization and minimization of adaptation time.

2 Related Work

Here we overview basic findings on the eye vergence mechanisms, with the main focus on S3D display conditions. We refer the reader to a survey by Meesters et al. [2004] for an in-depth discussion of other aspects of S3D display perception.

Vergence as a Dynamic Process The eye vergence is driven by the depth changes of a target object, and can be performed with high accuracy (error below 10 arcmin) both in the real world and S3D display observation conditions [Okuyama 1998]. Other factors, such as blur, proximity, target size, and luminance might affect vergence, but to a lesser extent [Campbell and Westheimer 1959]. Vergence is a relatively slow process when compared to other eye movements, e. g., saccades (below 60 ms), and requires about 195–750 ms for convergence and 240–1000 ms for divergence. Vergence latency also demonstrates a similar asymmetric behavior with 180–250 ms for convergence and 200–210 ms for divergence [Krishnan et al. 1973; Krishnan et al. 1977; Semmlow and Wetzel 1979].

Vergence is a two-stage process, where at first the fast transient (a.k.a. phasic) mechanism (reacts even for brief 200 ms flashes) brings the vergence in the proximity of the target depth, and then the slower sustained (a.k.a. tonic) mechanism is responsible for the precise verging on the target, as well as further tracking of slower depth changes. Semmlow et al. [1986] found that for less dynamic depth changes, with the ramp velocity below 2 deg/s, only the sustained mechanism is active, above 9 deg/s the transient mechanism dominates, and otherwise both mechanisms are active. Vergence adaptation (similar to luminance adaptation) has been observed in which the sustained mechanism supports a given eye vergence angle, and comfort state is achieved during binocular vision [Hung 1992]. For small depth changes within Panum's fusional area, the motoric vergence is not activated, and sensoric fusion of images on the retina is sufficient.

Vergence vs. Accommodation While vergence is driven by depth, and accommodation is driven mostly by retinal blur, both systems are reflexively coupled, and they interact with each other through accommodative vergence and vergence accommodation [Hung 2001]. The accommodative vergence is quantified by the AC/A ratio, which relates the change of vergence caused by the change of accommodation in the absence of disparity. In an analogous way, the vergence accommodation is quantified by the CA/C ratio in the absence of blur. Since the range of accommodation while viewing the S3D display is determined by the distance to the screen, unnatural decoupling of vergence and accommodation is required, which may cause visual discomfort and increase binocular fusion times [Hoffman et al. 2008; Lambooij et al. 2009]. When the screen disparity increases beyond Panum's fusional area, vergence eye movements bring the disparity back to this area, which shifts accommodation away from the screen. When such a shift is beyond the depth of focus (DOF) zone, the accommodative-vergence feedback is activated to counteract the loss of sharp vision, which in turn directs vergence back towards the display [Lambooij et al. 2009].

The range of vergence angles that assure clear and single binocular vision is known as the "comfort zone" [Shibata et al. 2011; Zilly et al. 2011]. In the real world, objects away from the fixation point are perceived as blurred, which reduces the visibility of diplopia, because the limits of fusion are higher for low spatial frequencies. Thus, both accommodation and vergence response can be improved by manipulation of convergence and local image defocus, respectively [Ukai and Kato 2002; Zwicker et al. 2006].

In many practical S3D applications, the comfort zone is determined by the disparity range of 70 arcmin [Lambooij et al. 2009; Zilly et al. 2011]. Since it is a rather conservative bound, in this work we assume a wider range of ±2.5 deg. Out-of-screen effects even beyond this range are used in cinematography, but the object of interest typically moves steadily off the screen in such cases, so that the viewer can adapt to its extreme position [Zilly et al. 2011]. Achieving such extreme disparities would not be possible through sudden jumps as in the case of scene cuts.

Vergence Measurements There is a large body of research on measurements of vergence dynamics in response to pulse, step, ramp, and sinusoidal disparity changes. For us, the step-like changes are the most relevant. Most experiments used physical targets or passively-shifted screens [Erkelens et al. 1989; Hung et al. 1994]. Simple stimuli, such as vertical lines, were used to eliminate other cues that could affect vergence. Special care was taken to suppress accommodation by using pinhole apertures for blur-free viewing. A wide range of disparities ±35 deg have been considered [Erkelens et al. 1989], but a typical range was below ±10 deg with relatively large step amplitudes, typically larger than 2 deg [Hung 2001].

In this work we focus on the disparity range ±2.5 deg and lower

disparity step amplitudes, which are important for comfortable experience while viewing S3D displays. The assumed disparity range corresponds approximately to the comfort zone in desktop viewing conditions given by Shibata et al. [2011, Fig. 23]. By using an off-the-shelf S3D display in our experiments, and dealing with real-world images in the validation step, we ensure that the conditions are possibly similar to the ones in expected applications, where accommodation and pictorial cues may affect the vergence. Also, the initial disparity magnitude is important in our measurements, both for the convergence and divergence case.

Vergence Modeling Schor [1979] and Hung [1998] proposed sophisticated models of the eye vergence dynamics, which employ the concepts of control engineering to simulate the transient and sustained (a negative feedback loop) mechanisms. The models have been extended to handle accommodation as well as the AC/A and CA/C cross-link gains [Schor 1992; Schor 1999; Hung 2001]. An extensive validation of such models against measurement data has been performed; however, disparity steps interesting for us have been treated marginally. Furthermore, the viewing conditions did not force decoupling of accommodation and vergence.

While S3D displays have been considered in some computational models, the main goal was to artificially alter the link between the accommodation and vergence systems to study the change in pre-task and post-task measures of AC/A and CA/C [Eadie et al. 2000], or to investigate developmental plasticity in children exposed to S3D games [Rushton and Riddell 1999]. Alvarez et al. [2005] experimented with constant-sized, 4 deg steps, and found that in case of divergent steps, vergence dynamics are dependent on the initial disparity. In this work, we propose a simple data-driven model of eye vergence that is tuned to step-like disparity changes. We emphasize here on vergence dynamics as a function of the initial *and* target disparities, and our goal is minimization of the vergence adaptation time at scene cuts through disparity editing.

Temporal Changes vs. Comfort Yano et al. [2004] report that visual discomfort was induced if images were moved in depth according to a step pulse function, even if the images were displayed within the depth of focus. In a related work by Tam et al. [2012], influence of disparity and velocity on visual comfort was investigated, and a significant interaction between velocity and disparity was shown. The negative effect of object velocity on visual comfort was apparent even when the objects were displayed within the generally accepted visual comfort zone of less than 1 deg of horizontal disparity. Results obtained by Lambooij et al. [2011] show that rapidly moving objects and changing screen disparity indeed have a significant effect on visual comfort; however, their dominant role was not confirmed.

S3D Content Processing The problem of scene transitions is challenging in the context of stereoscopic content, since scene transitions often create large temporal disparity discontinuities leading to visual discomfort. To solve this problem, disparity adjustment techniques are required. They are performed either during the acquisition step by modifying camera parameters or in the post-processing step using, for example, horizontal image translation [Mendiburu 2009]. However, only few techniques can deal with temporal effects.

Disparity velocity was considered one of the important factors for disparity adjustment [Lang et al. 2010]. The authors proposed to interpolate between different disparity ranges at scene cuts to reduce large discontinuities in disparity. To this end, different disparity mapping operators can be used to make this adjustment; however, the decision of how this interpolation is defined was left to the user. A simpler technique has been proposed by Koppal [2011]. He

suggested to solve the problem of transitions by cross-fading the horizontal image translation to zero at the cut.

Bernhard et al. [2014] showed how binocular fusion times can be reduced by means of active manipulation of the convergence plane. The object of interest is brought back to the zero-disparity plane once the change in gaze has been detected, but before the vergence adaptation is complete. In contrast to Bernhard et al.'s active approach, we propose a cut optimization process that keeps the disparities constant during the vergence adaptation. The improvement in our case comes from more informed choice of the initial and target disparities. Nevertheless, both approaches could be potentially combined.

Heinzle et al. [2011] proposed a computational camera rig, which enables intuitive control over camera parameters. The artist's involvement is still needed, though, to design the transitions manually, without any feedback on human abilities to adapt to rapid disparity changes. Automatic control over camera parameters was proposed in the context of real-time systems (e. g., games) [Oskam et al. 2011]. However, their primary goal was to maintain the scene disparity range within given limits. This is not equivalent to minimizing the vergence adaptation time, which depends not only on disparity difference but also the initial disparity value. In our approach, we take those two factors into account.

More recently, a metric of visual comfort has been proposed [Du et al. 2013], which directly addresses the problem of temporal disparity changes. The authors also suggest that it can be used for optimizing stereoscopic parameters. However, their metric deals with motion, and it is unclear how to apply their technique in the context of rapid disparity changes such as those created during scene transitions.

3 Model Derivation

In this section, we experimentally derive and evaluate a model of eye vergence response to step-like changes in disparity. We also estimate model parameters for an average observer. The collected data is useful in a number of applications, as shown in Sec. 4.

Participants Sixteen subjects (8 F, 8 M) took part in our experiment. They were members of computer graphics and computer vision groups, between 21 and 35 years old. All had normal or corrected-to-normal vision, and all passed a test for stereo-blindness.

Equipment Stimuli were presented using an Nvidia 3D Vision 2 kit and an Acer GD235HZ 23.6-inch screen with native resolution of 1920×1080. In order to measure the vergence responses, both eyes were tracked using an EyeLink 1000 Plus eye tracker with a desktop mount. The tracker records 1000 samples per second (500 per eye), allowing for fine-scale analysis of the vergence response. The spatial accuracy according to the eye-tracker manufacturer is up to 0.25–0.5 deg. A chin-rest was used to stabilize the subject's head, and the viewing distance was fixed to 55 cm.

Stimulus The stimulus in our experiment was a low-pass filtered white-noise patch changing its disparity in discrete steps over time. The patch was presented centrally on the screen, on a neutral grey background, and it subtended ca. 11 degrees of visual angle. A single trial consisted of a sequence of disparities $d_1, d_2, \ldots, d_n$, chosen from a fixed set D. The ordering of the disparities was randomized to avoid learning effect, but only Eulerian paths were used, i. e., $d_1 = d_n$, and every possible transition appeared exactly once. Since prediction has been shown to have influence on vergence response (periodic disparity changes can be followed by vergence without typical latency [Hung 1998]), the time between the onsets of consecutive stimuli was set randomly between 1.25 s and 2.5 s.

Task Each session of the experiment started with a calibration procedure, as described in the eye tracker manual. Next, every participant had to perform m trials, and the task was to simply observe the patch. The participants were encouraged to take breaks whenever they felt tired, and after each break the eye tracker was re-calibrated. The entire session took approximately 40 minutes.

Data Analysis After each session, binary output of the eye tracker was converted to a plain-text version using the converter tool provided by the manufacturer. Next, the data was processed using a custom parser to extract gaze coordinates and times of disparity changes, and read into MATLAB R2012a. The times of stimulus onsets were marked in the output files with timestamps – a functionality provided by the tracker's API, which enabled easy synchronization of the gaze data with stimuli. For each transition, we extracted the 1-second segment following it, smoothed using a small box filter, and converted it to vergence values. Vergence was calculated as the difference between the x-coordinates of the two gaze positions expressed in pixels. Missing or unreliable samples (due to, e. g., blinks, saccades, or tracking errors) were interpolated linearly, and the segments that required interpolation of more than 50% samples were excluded. Data for transitions of one type was grouped, and a curve was fitted to the average. Next, for each type of a transition, the time to reach 95% of the required vergence change was determined, and two surfaces were fitted to the obtained data points. Since we were interested in relative gaze positions, the significance of drift was low. Moreover, adaptation times were determined by the 95%-of-change position, which is not very sensitive to shifts, scaling, etc. Based on these premises, we believe the precision was sufficient for our purposes.

3.1 Pilot Experiment

In order to gain insight into the relation of vergence response to the initial and end disparities, as well as to estimate the number of trials m necessary for the response curves to converge, we conducted a pilot study. In it, one subject (S7) performed $m = 30$ trials, with $d_i = 0, \pm 30, \pm 60, \pm 90$ px, and the cut-off frequency of the low-pass filter $f = 20$ cpd. This gave $30 \cdot 7 \cdot 6 = 1260$ transitions measured. The results are presented in Fig. 2.

Discussion The signal converged quickly, giving relatively smooth data after ca. 5 repetitions, and little could be gained after ca. 10 repetitions. The vergence response can be modeled very well by sigmoid functions of the form $v = ae^{be^{ct}} + d$, known as the Gompertz curves. The 95%-point does not depend on parameters a and d, and can be obtained using the following formula: $p_{95} = \ln(\ln(0.95)/b)/c$. The obtained data points can be modeled almost perfectly using two planes, with mean error close to 0, and standard deviation of ca. 27 ms. In light of these findings we decided to limit the disparity values used in the main experiment to $d_i = \pm 30, \pm 90$ px, and the number of repetitions m to 10.

3.2 Main Experiment

The aim of the main experiment was twofold: to confirm that vergence times can be well modeled using two planes, as suggested by the pilot experiment, and, if so, to estimate parameters of the average-observer model, useful in practical applications. In this experiment $n = 16$ subjects performed $m = 10$ trials (except subjects S6, S9, and S10 for whom $m = 5$), with the cut-off frequency $f = 10$. The range of disparities for subject S9 was reduced to 2/3, due to reported problems with fusion. The results are presented in Fig. 3.

Discussion The average standard deviation of error after fitting the planes to the obtained data equals 36 ms. This indicates a very good fit, and justifies our assumption that the vergence adaptation time can be modeled using planes. In particular, this means that the data from subject S9, who saw rescaled disparities, could be easily included in the average model.

As expected, our measurements show that given the initial disparity and direction, steps with larger magnitude lead to longer vergence adaptation times. An interesting finding is that the adaptation time depends also on the step direction and initial disparity. Given the initial disparity (Fig. 3, right, abscissae) and step magnitude (one yellow and one green line per magnitude), steps *towards* the screen are generally faster: To the right of the graph, yellow lines (convergent steps) have lower times than the corresponding green lines (divergent steps). To the left, this is reversed. Note, that corresponding yellow and green lines intersect near the point of zero initial disparity (screen plane). We hypothesize that it is related to accommodation-vergence coupling, which attracts vergence towards the screen plane, where the A/V conflict disappears.

Additionally, given the step magnitude and direction (Fig. 3, either one yellow or one green line), with decreasing initial disparity, convergent steps get slower whereas divergent steps get faster. This effect could be convincingly explained by the amount of A/V conflict which increases with disparity magnitude. At *negative* initial disparities, divergent steps work towards resolving the conflict, whereas convergent steps work towards increasing it. With *positive* initial disparities the roles are reversed. The larger the magnitude of the initial disparity, the more stress is put on the visual system, and the demand to resolve (or not to increase) the conflict is higher. Thus, the larger discrepancy between convergent and divergent steps. These effects should be taken into account while optimizing stereoscopic content, as simple minimization of *disparity difference* will not necessary lead to shorter adaptation times.

Another interesting finding is that with fixed *target disparity*, adaptation times for convergent steps are hardly dependent on the step magnitude. This phenomenon, at first unintuitive, could be explained by the A/V coupling as well: Larger step magnitudes, which should intuitively contribute to longer adaptation times, may be offset by varying initial stress exerted by the A/V conflict on the visual system.

In our experiment we considered only a computer display observed at a relatively short distance. On the one hand, at larger viewing distances the depth of field increases, thereby reducing the importance of the A/V coupling, the hypothesized cause of the observed variation in vergence adaptation time. On the other hand, discomfort induced by step-like motion in depth has been observed even for disparities within the DOF [Yano et al. 2004]. Answering the question, if similar effect of initial disparity on the adaptation time can be observed in other viewing conditions, e. g., in cinema, requires further investigation.

3.3 Evaluation

The obtained model was derived using simple stimuli (flat white-noise patterns). On the one hand, this approach has several advantages: the exact disparity is known, regardless of fixation points; the measurements can be repeated easily; and the learning effect is reduced, since the subject has no memory related to spatial arrangement of objects in case of repeated images. On the other hand, it is unclear how well the model predicts response to cuts between natural images: the presence of complex luminance patterns or high-level processes related to scene understanding may very well influence the transition times. Therefore, we conducted a validation experiment, to test if the model can be generalized.
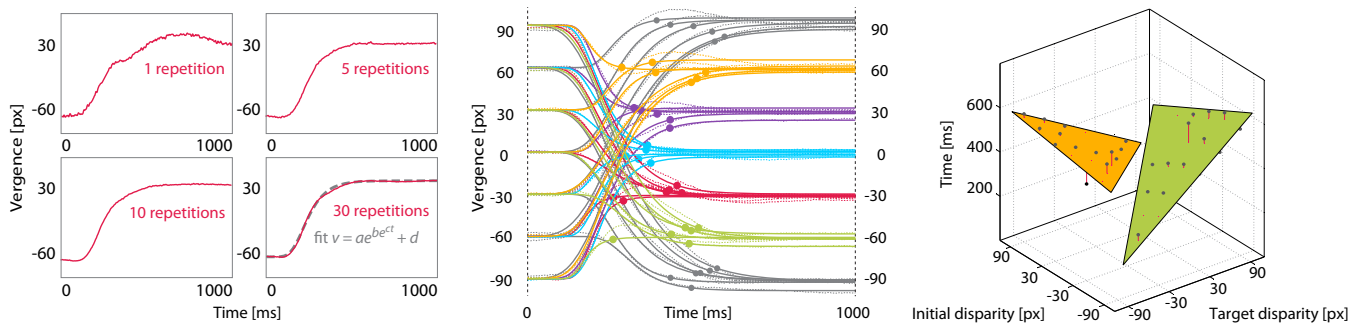
Figure 2: *The results of the pilot experiment. In the left panel we presented averaged vergence responses of subject S7 to a $-60\,\text{px} \to 30\,\text{px}$ step, after 1, 5, 10, and 30 repetitions. The curve after 30 repetitions is shown together with a fit of a Gompertz function. In the middle panel average vergence responses to all 42 possible disparity steps for subject S7 are shown (dashed lines), together with fitted curves (solid lines), and points where the curves reach 95% of vergence change (solid circles). In the right panel, we plotted the transition time against the initial and target disparity. These points are almost perfectly modeled by two planes – the standard deviation of the error is approximately 27 ms. The two planes represent divergence (green) and convergence (yellow). We leave a gap between the planes, where times begin to increase due to Panum's fusional area and tolerance of the visual system to vergence errors. The diagonal is a singularity, where no transition is present, because the initial and target disparities are equal.*
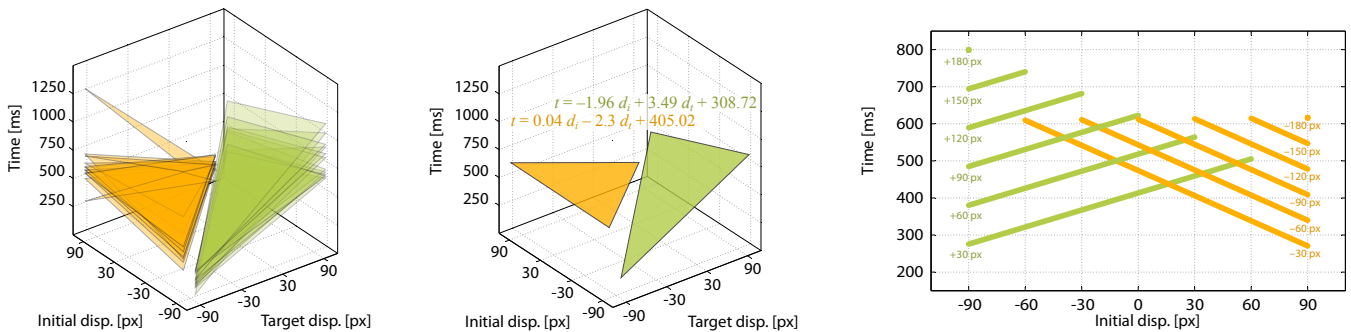

Figure 3: *The results of the main experiment. In the left panel we presented fits for all subjects, after exclusion of 4 outliers (subjects S1, S6, S8, and S14). These subjects were excluded due to serious difficulties with correct fusion of the stimuli. For completeness we provide their data in the supplemental materials. The middle panel shows the average of all fits from the left panel, along with equations of the planes. These planes describe transition times for the average observer. In the right panel we showed a number of diagonal sections of the average planes. Each line represents disparity steps of the same magnitude and direction, but with different initial disparities. See the text for a detailed discussion.*

Participants and Stimuli Four participants (S3, S7, S11, and S16) from original 16 were invited to take part in the evaluation of the model. Six 3D photographs taken with an LG Optimus 3D P725 smartphone were used (see Fig. 4). They were divided into two groups of three, one with smaller and the other with larger disparity changes across pictures. The disparities in the picture were estimated using the SIFT flow algorithm [Liu et al. 2011]. In a single trial a 6.5-minute random sequence composed of the three photographs from one of the groups was shown. As previously, a single appearance of a picture lasted between 1.25 s and 2.5 s (chosen randomly), and there were no breaks between appearances. The task was to simply observe the pictures, and the participants were asked to perform one trial for each group.

Data Analysis and Results After cleaning and segmenting of the tracking data, a semi-automatic procedure was employed to group segments of the same type, enabling averaging of measurements. In the first, automatic step segments where a saccade occurred at the time of the cut, or within the first 100 ms after the cut, were discarded. Then, initial disparity was estimated using the disparity map and the fixation coordinates just before the cut (initial fixation). The target disparity was chosen using the following heuristic: whenever the duration of the first fixation was shorter than 300 ms, the second fixation was used; otherwise, the initial fixation was assumed to be

also the target fixation.

In the second, manual step, all segments were briefly reviewed to correct filtering and target fixation errors. The false negatives were the cases when the saccade near the cut was small enough not to change significantly the vergence response. The false positives were the non-typical cases, including, but not limited to, eye-tracker errors, clearly incorrect vergence response indicating lack of fusion, segments with unusually large saccade-to-fixation ratio, erratic saccades indicating partial fixations, etc. In the end, 718 out of 3028 segments were discarded. We provide all annotated segments along with a custom viewer/editor as additional materials, and encourage the readers to inspect the data we used in this evaluation.

In the end, segments with the same initial/target disparities were grouped; groups with 5 or more members were averaged and compared against the model prediction for the respective subject. The results of the experiment are presented in Fig. 5.

Discussion Although our prediction slightly overestimated the time of transition for photographs, our model correlated well with the actual time, as indicated by relatively low standard deviation of the error. The study proves that our model is a good predictor (up to an additive constant) of transition time for natural images. We hypothesize that improved performance was due to the presence of
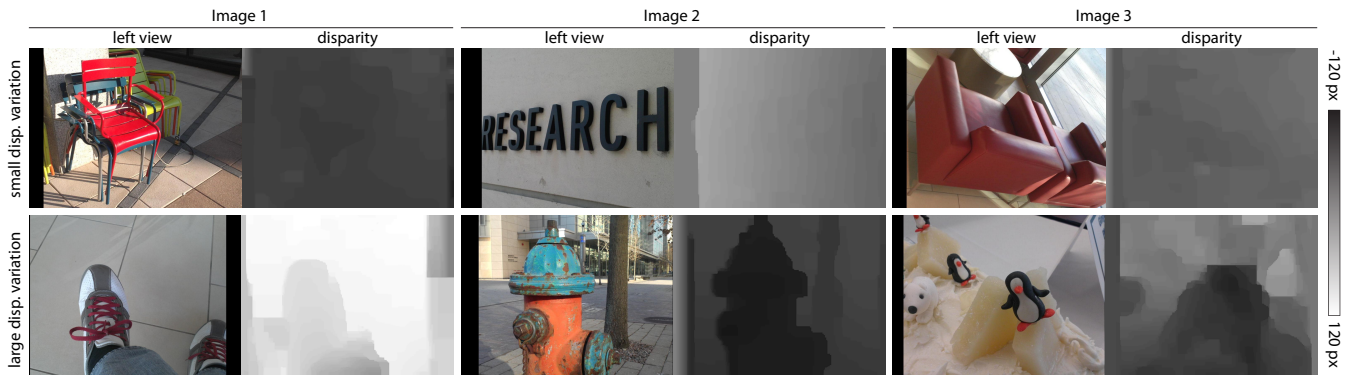
Figure 4: *The two groups of stimuli used in the evaluation, one with larger, and one with smaller disparity variation across pictures. The black bars on the sides are floating stereoscopic windows added to avoid frame violation or large disparity steps at the edge (the shoe example).*
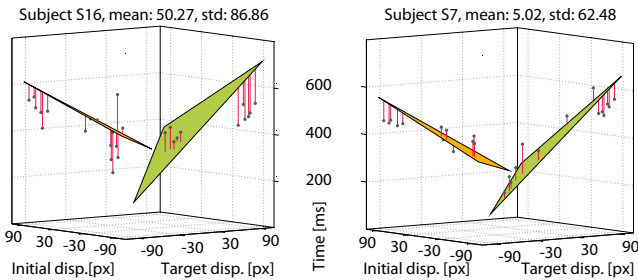

Figure 5: *The results for subjects S16 (left) and S7 (right); plots for subjects S3 and S11 are provided in the supplemental materials. The planes show model predictions, whereas the solid circles represent the observed data. The mean and standard deviation of the error for subjects S16, S7, S3, and S11 are respectively 50 ± 87 ms, 5 ± 62 ms, 107 ± 90 ms, and 61 ± 84 ms.*

higher-order cues, absent in white-noise stimuli, where the sole depth cue was binocular disparity. It is also possible that the transition was facilitated to some extent by the learning effect.

4 Applications

In this section, we propose a set of tools for aiding in the production of stereoscopic content, that utilizes our model to minimize vergence adaptation times. We also analyze the impact of the minimization on visual quality in one of the proposed tools using an object-recognition experiment.

4.1 Production Tools

Transition Time Visualization A straightforward application of the model is a visualization tool providing stereographers and VFX artists with an interactive analysis of transition times. In order to evaluate stereoscopic transition and estimate transition time, we first need to determine the pairs of disparity values between which the transitions occur. A naïve approach would be to measure the transition time between corresponding pixels in both sequences; however, it is not very useful, as in most cases people change the fixation point immediately after the transition, and no change in vergence happens (see the data browser provided in supplemental materials). Therefore, the fixation points in both sequences need to be precisely determined.

Such data can be obtained from various sources, e. g., it is possible to use eye-tracker data. This does not require many subjects, as it

has been shown that eye scan-paths form highly repetitive patterns between different spectators for the same video sequences [Wang et al. 2012]. Moreover, skilled directors are capable of precisely guiding and predicting viewers' attention. Such prediction is further facilitated by the tendency of increasing object motion in modern movies [Cutting et al. 2011] and by the fact that typical 2D-movie cuts trigger saccades towards the screen center [Mital et al. 2011; Wang et al. 2012]. Thus, the information about fixation points for our methods can be very reliably provided by the directors. Besides, Carmi and Itti [2006] observed that the saccades immediately after the cut are driven mostly by the bottom-up factors and can be predicted relatively well by existing saliency models. Once the fixation points before and after the cut are known, the corresponding disparity values need to be determined. This can be obtained directly from the rendering pipeline for animated movies, using user input in the case of 2D-to-3D conversion, or using disparity estimation techniques for natural scenes when the depth map is not available. Once the fixation points along with disparity values are known, transition times can be directly calculated from the model. Since computing model predictions is inexpensive, it can be used to provide real-time preview of transition times.

Camera Parameters Optimization Apart from predicting transition times and visualizing them for editing purposes, one can automate the process of stereoscopic content preparation. An optimization problem for cuts can be defined, and our model can serve as the core of the cost function.

As discussed in Sec. 2, stereoscopic content can be optimized by manipulating various parameters. These can be changed for the entire sequence (e. g., from cut to cut), or selectively around the cuts, with smooth blending back to original parameters [Lang et al. 2010; Koppal et al. 2011]. There is a wide range of manipulations that can be used to adjust stereoscopic content. They range from very simple ones, like changing camera separation and convergence (i. e., the plane of zero parallax), to more complicated ones, such as depth remapping. All such manipulations can be easily integrated and used with our model.

Cut Positioning If the two sequences between which the cut occurs overlap in time, it is also possible to find the best moment for the cut. To this end, we optimize not only stereoscopic parameters, but also the position of the cut. This can be performed efficiently by simply iterating over all possible cut positions, in addition to all horizontal shifts of the left/right views. The optimal cut can be chosen automatically or can be shown to the editor as a suggestion. The design of a tool performing these tasks is shown in Fig. 1, right, and in the supplemental video.

4.2 Impact on Visual Quality

Visual quality can be defined in many ways, using various objective and subjective criteria. In the following experiment, we focus on the time necessary to recognize the 3D arrangement of objects after a cut. We assume shorter recognition times to be an indicator of higher quality. We measured the time needed to recognize object arrangement, and showed that this time closely matches our model. In practice, this means that when cuts are optimized using the proposed production tools, the time necessary to recognize objects in the scene is minimized.

Methods The equipment and viewing conditions were the same as in other experiments, but no eye-tracker was used. As stimuli, we used two shots corresponding to a cut in the 3D version of the *Big Buck Bunny* animation. We modified them by placing two small dark-gray circles between the eyes of the character, with approximately the same disparity as the character (see Fig. 6, left, inset). Two 3D arrangements of circles for each shot were considered: one with the upper, and one with the lower circle closer to the observer. The disparity difference between the circles was 2 px. The convergence in the shots was modified so that the average disparity of the circles was equal to d_i before and d_t after the cut. Seven pairs of disparity steps were used: $-75 \rightarrow -105/90$, $-60 \rightarrow -90/-30$, $-30 \rightarrow -90/60$, $0 \rightarrow -30/30$, $30 \rightarrow -60/90$, $60 \rightarrow 30/90$, and $75 \rightarrow -90/105$ px. For each initial disparity, both a convergent and a divergent step was possible, which prevented anticipatory eye movements in subjects. In order to determine the arrangement recognition time for all 14 steps, we performed 14 independent QUEST threshold estimation procedures [Watson and Pelli 1983], each estimating time of 75% correctness. A single trial of each procedure had the following structure: First, the first shot was shown for 2 s. Next, the second shot was shown for a period between 0.1 and 1.5 s (controlled by QUEST). The arrangement of the circles was chosen randomly in every trial. After the screen was blanked, the subject was asked to indicate if the arrangement was the same in both shots: If the same circle (i. e., upper or lower) was closer to the observer both before and after the cut, the subject had to press the Y key, and the N key otherwise. Such a task definition ensured that the subject actually performed the vergence transition $d_i \rightarrow d_t$. All 14 procedures were performed in parallel, randomly interleaved. A session of the experiment lasted 20 min (average standard deviation in a QUEST instance 73 ms). Subjects S3, S11, S12, S15, and S16 took part in the experiment. S11 participated in three sessions, S16 in two, and the remaining three in one session.

Results The data obtained using the above procedure was fitted with two planes minimizing the RMSE. The planes obtained from all subjects were averaged (first within subjects, then between subjects), and compared to their average model. The results are presented in Fig. 6. A corrective constant shift of 83 ms minimizes the RMSE, and yields a low prediction error of 42 ms. This correlation implies that optimizing camera convergence using our model instead of disparity distance as the cost function will produce cuts with shorter recognition times. Similar improvement can be expected when optimizing other camera parameters or cut positions. This illustrates the practical importance of our model for S3D games and films.

5 Conclusions

We proposed a new model which predicts the time a human observer needs to adapt vergence to rapid disparity changes. We first presented measurements of transition times for simple stimuli, and demonstrated that these times are valid also for complex scenes. The experiment revealed interesting facts about viewer behavior during
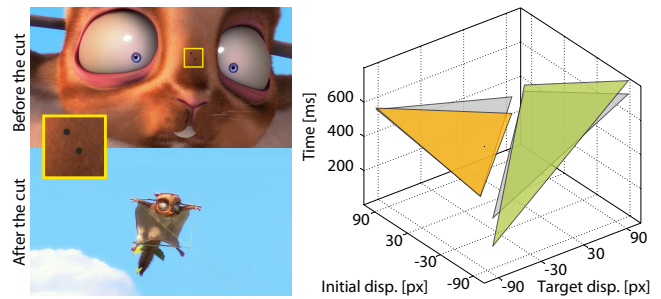

Figure 6: Left: *Stimuli used in the object recognition experiment.* Right: *The results of the experiment; the gray planes represent the obtained data after corrective shift of 83 ms, and the yellow/green planes represent the average model of the subjects, predicting the data with the* RMSE *equal 42 ms.* Pictures from Big Buck Bunny *CC-BY Blender Foundation and Janus B. Kristensen*

scene cuts, which provides valuable knowledge for stereoscopic content creators. Additionally, we proposed a set of tools for the editing of stereoscopic content to minimize the vergence adaptation time after cuts. An important property of the proposed optimization techniques is that the manipulations are applied only locally around cuts, which has limited effect on the depth impression created by the artist. To our knowledge, this is the first work that proposes to automatically edit stereoscopic cuts taking into account varying performance of the human visual system in adapting to rapid disparity changes. Finally, we demonstrated the impact of minimizing adaptation times on the visual quality of S3D content as measured by a subject's performance in the 3D object recognition task. An interesting avenue for future work would be an extensive user study quantifying how shorter transition times influence visual fatigue.

Acknowledgments

References

ALVAREZ, T. L., SEMMLOW, J. L., AND PEDRONO, C. 2005. Divergence eye movements are dependent on initial stimulus position. *Vis. Res. 45*, 14, 1847–55.

BERNHARD, M., DELLMOUR, C., HECHER, M., STAVRAKIS, E., AND WIMMER, M. 2014. The effects of fast disparity adjustments in gaze-controlled stereoscopic applications. In *Proc. ETRA*. To appear.

CAMPBELL, F. W., AND WESTHEIMER, G. 1959. Factors influencing accommodation responses of the human eye. *J. Opt. Soc. Am. 49*, 6, 568–71.

CARMI, R., AND ITTI, L. 2006. Visual causes versus correlates of attentional selection in dynamic scenes. *Vis. Res. 46*, 26, 4333–45.

CUTTING, J., BRUNICK, K., DELONG, J., IRICINSCHI, C., AND CANDAN, A. 2011. Quicker, faster, darker: Changes in hollywood film over 75 years. *i-PERCEPTION 2*, 6, 569–76.

DU, S.-P., MASIA, B., HU, S.-M., AND GUTIERREZ, D. 2013. A metric of visual comfort for stereoscopic motion. *ACM Trans. Graph. 32*, 6, 222.

EADIE, A. S., GRAY, L. S., CARLIN, P., AND MON-WILLIAMS, M. 2000. Modelling adaptation effects in vergence and accommodation after exposure to a simulated virtual reality stimulus. *Ophthalmic Physiol. Opt. 20*, 3, 242–51.

ERKELENS, C. J., VAN DER STEEN, J., STEINMAN, R. M., AND COLLEWIJN, H. 1989. Ocular vergence under natural conditions. II. Gaze-shifts between real targets differing in distance and direction. In *Proc. of the Royal. Soc.*, 441–6.

FINKE, R. 1989. *Principles of Mental Imagery*. MIT Press.

HEINZLE, S., GREISEN, P., GALLUP, D., CHEN, C., SANER, D., SMOLIC, A., BURG, A., MATUSIK, W., AND GROSS, M. H. 2011. Computational stereo camera system with programmable control loop. *ACM Trans. Graph. 30*, 4, 94.

HOFFMAN, D., GIRSHICK, A., AKELEY, K., AND BANKS, M. 2008. Vergence-accommodation conflicts hinder visual performance and cause visual fatigue. *J. Vision 8*, 3, 1–30.

HUNG, G. K., CIUFFREDA, K. J., SEMMLOW, J. L., AND HOR, J.-L. 1994. Vergence eye movements under natural viewing conditions. *Invest. Ophthalmol. Vis. Sci. 35*, 3486–92.

HUNG, G. K. 1992. Adaptation model of accommodation and vergence. *Ophthalmic Physiol. Opt. 12*, 3, 319–26.

HUNG, G. K. 1998. Dynamic model of the vergence eye movement system: Simulations using Matlab/Simulink. *Computer Methods and Programs in Biomedicine 55*, 1, 59–68.

HUNG, G. K. 2001. *Models of oculomotor control*. World Scientific Publishing, Singapore.

KOPPAL, S. J., ZITNICK, C. L., COHEN, M., KANG, S. B., RESSLER, B., AND COLBURN, A. 2011. A viewer-centric editor for 3d movies. *IEEE Comput. Graph. Appl. Mag. 31*, 1, 20.

KRISHNAN, V., FARAZIAN, F., AND STARK, L. 1973. An analysis of latencies and prediction in the fusional vergence system. *Am. J. Optometry and Arch. Am. Academy of Optometry 50*, 933–9.

KRISHNAN, V., FARAZIAN, F., AND STARK, L. 1977. Dynamic measures of vergence accommodation. *American Journal of Optometrics and Physiological Optics 54*, 470–3.

LAMBOOIJ, M., IJSSELSTEIJN, W., FORTUIN, M., AND HEYNDERICKX, I. 2009. Visual discomfort and visual fatigue of stereoscopic displays: A review. *J. Imaging Sci. Technol. 53*, 3, 1.

LAMBOOIJ, M., IJSSELSTEIJN, W., AND HEYNDERICKX, I. 2011. Visual discomfort of 3D TV: Assessment methods and modeling. *Displays 32*, 4, 209–18. Visual Image Safety.

LANG, M., HORNUNG, A., WANG, O., POULAKOS, S., SMOLIC, A., AND GROSS, M. 2010. Nonlinear disparity mapping for stereoscopic 3D. *ACM Trans. Graph. 29*, 4, 75.

LIU, C., YUEN, J., AND TORRALBA, A. 2011. Sift flow: Dense correspondence across scenes and its applications. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 33*, 5, 978–94.

MEESTERS, L., IJSSELSTEIJN, W., AND SEUNTIENS, P. 2004. A survey of perceptual evaluations and requirements of three-dimensional tv. *Circuits and Systems for Video Technology, IEEE Transactions on 14*, 3, 381–91.

MENDIBURU, B. 2009. *3D Movie Making: Stereoscopic Digital Cinema from Script to Screen*. Focal Press.

MITAL, P., SMITH, T., HILL, R., AND HENDERSON, J. 2011. Clustering of gaze during dynamic scene viewing is predicted by motion. *Cognitive Computation 3*, 1, 5–24.

OKUYAMA, F. 1998. Human visual accommodation and vergence eye movement while viewing stereoscopic display and actual target. In *Proc. IEEE Eng. Med. Biol. Society*, vol. 2, 552–5.

OSKAM, T., HORNUNG, A., BOWLES, H., MITCHELL, K., AND GROSS, M. H. 2011. Oscam-optimized stereoscopic camera control for interactive 3d. *ACM Trans. Graph. 30*, 6, 189.

OWENS, C., 2013. Invited talk. 2nd Toronto International Stereoscopic 3D Conference.

RUSHTON, S. K., AND RIDDELL, P. M. 1999. Developing visual systems and exposure to virtual reality and stereo displays: some concerns and speculations about the demands on accommodation and vergence. *Applied Ergonomics 30*, 1, 69–78.

SCHOR, C. M. 1979. The relationship between fusional vergence eye movements and fixation disparity. *Vis. Res. 19*, 12, 1359–67.

SCHOR, C. M. 1992. The relationship between fusional vergence eye movements and fixation disparity. *Optometry and Vision Science 69*, 4, 258–69.

SCHOR, C. 1999. The influence of interactions between accommodation and convergence on the lag of accommodation. *Ophthalmic Physiol. Opt. 19*, 2, 134–50.

SEMMLOW, J., AND WETZEL, P. 1979. Dynamic contributions of the components of binocular vergence. *JOSA 69*, 639–45.

SEMMLOW, J., HUNG, G., AND CIUFFREDA, K. 1986. Quantitative assessment of disparity vergence components. *Invest. Ophthalmol. Vis. Sci. 27*, 558–64.

SHIBATA, T., KIM, J., HOFFMAN, D. M., AND BANKS, M. S. 2011. The zone of comfort: Predicting visual discomfort with stereo displays. *J. Vision 11*, 8, 11.

TAM, W. J., SPERANZA, F., VÁZQUEZ, C., RENAUD, R., AND HUR, N. 2012. Visual comfort: stereoscopic objects moving in the horizontal and mid-sagittal planes. In *Proc. SPIE*, A. J. Woods, N. S. Holliman, and G. E. Favalora, Eds., 8288:13.

UKAI, K., AND KATO, Y. 2002. The use of video refraction to measure the dynamic properties of the near triad in observers of a 3-d display. *Ophthalmic Physiol. Opt. 22*, 5, 385–8.

WANG, H. X., FREEMAN, J., MERRIAM, E. P., HASSON, U., AND HEEGER, D. J. 2012. Temporal eye movement strategies during naturalistic viewing. *J. Vision 12*, 1, 16.

WATSON, A. B., AND PELLI, D. G. 1983. QUEST: a Bayesian adaptive psychometric method. *Perception and Psychophysics 33*, 2, 113–20.

YANO, S., EMOTO, M., AND MITSUHASHI, T. 2004. Two factors in visual fatigue caused by stereoscopic HDTV images. *Displays 25*, 4 (Nov.), 141–50.

ZILLY, F., KLUGER, J., AND KAUFF, P. 2011. Production rules for stereo acquisition. *Proc. IEEE 99*, 4, 590–606.

ZWICKER, M., MATUSIK, W., DURAND, F., PFISTER, H., AND FORLINES, C. 2006. Antialiasing for automultiscopic 3D displays. In *Proc. EGSR*, 73–82.