Applied Psychophysics: Studies in Support of Perception-Directed Graphics Rendering

Kevin Larson, Eric Horvitz, Jed Lengyel, and Mary Czerwinski
Microsoft Research
One Microsoft Way
Redmond, WA 98052, USA
a-kevlar, horvitz, jedl, marycz @microsoft.com

ABSTRACT
We present the motivation, methods, and results for a set of perceptual studies that were performed to assess the ability of people to discriminate changes in the quality of the graphical output of an experimental flexible graphics rendering system. The studies were designed to characterize the perceptual sensitivities to reductions in the quality of rendered images along the specific dimensions of degradation that are exploited in the rendering architecture. Two dimensions of the degradation of image quality were explored independently and in concert. The results confirm intuitions about opportunities for guiding the allocation of resources in graphics rendering systems by harnessing information about changes in the perception of image quality associated with alternate degradations.

Keywords
Visual perception, graphics rendering, image quality

INTRODUCTION
Psychophysical research pursues the relationship between properties of physical stimuli and the resulting human experience. Human perceptions of physical qualities such as color, contrast, depth, sizes, and movement have been studied in detail. From this research, we know that what we perceive of each of these physical dimensions does not correspond exactly to the physical properties of the sensory input. Consider visual contrast by way of example. A series of one-inch thick black bars on white paper appears as a pattern of crisp intensity differences between the black and white bars. However, a series of very narrow black bars in close juxtaposition to one another appears as a homogenous gray field. Most color television systems take advantage of similar blending of independent colors to generate the perception of a full range of color.

The Qualia project [1,2] centers on the enhancement of 3D graphics rendering under limited or varying computational resource constraints by taking into consideration aspects of human visual processing. A key focus of the Qualia project is the development of methods for guiding decisions made in graphics rendering systems based on models of *perceived* image quality. Such rendering control decisions attempt to trade losses in the fidelity of graphics for gains in computational tractability of rendering.

In this paper, we describe several empirical studies undertaken to refine models used to control graphics rendering. In a companion paper [2], we describe additional details on the background, motivation, models and array of experiments undertaken within the Qualia project.

Complexity of Rendering
The rendering of high-quality 3D graphics can be computationally intensive. In dynamic computational environments, such as interactive media, an entirely new screen must be drawn thirty times a second in response to users' real-time actions. This is a difficult task if there are 1024 x 768 different pixels that require constant updating.

A traditional approach to addressing limitations in computational resources in graphics systems is to simply drop the frame rate. Such a basic approach to the problem of limited resources often leads to intolerable flickering and fragmentation. To avoid problems with the lowering of the frame rate, researchers have been building and designing rendering systems that provide one or more rendering approximation methods. Such methods include *flexible procedures* that can be controlled to trade off the amount of degradation in quality along different dimensions of approximation for gains in computational tractability. Such dimensions include the fidelity of the shape of objects, spatial resolution, and the shading complexity of rendered objects. The availability of multiple controllable methods poses sets of decision problems about the ideal allocation of resources in graphics systems. Such decisions are enhanced by information about the sensitivity of viewers to the degradations introduced by approximations.

We have pursued characterizations of viewers' sensitivities to support perception-based allocation of resources. Our studies were framed to answer specific questions about the control of flexible procedures made available in an experimental flexible rendering architecture named Talisman [3].

Studies in Support of a Flexible Rendering Architecture
In a distinction from traditional psychophysics research, our experiments are not motivated centrally by general questions about human perception. Rather, the experiments have been designed explicitly to study the sensitivities of the visual system to specific stimulus dimensions defined by the capabilities and tradeoffs of a particular graphics architecture.

We examined key dimensions of quality manipulated by the Talisman graphics architecture. The Talisman architecture provides several flexible approximation strategies. Details about Talisman are described in [3]. Previous work within the Qualia project focused on the development of models of perceptual cost and strategies for optimizing the control of rendering based on these models [2]. These models and control strategies take as inputs specific characterizations of the perceptual sensitivity of viewers to alternative degradations. Questions about the details of the perceptual relationships motivated and prioritized our experimental studies.

We have studied two key dimensions of degradation independently and in concert. These dimensions are spatial and temporal resolution. Spatial resolution refers to the detail with which an object or an entire frame is drawn. Temporal resolution is the rate that individual components of scenes are redrawn accurately, as opposed to approximated by methods that tend to introduce geometric distortions and jumping artifacts. Such jumping artifacts are introduced when a sprite is re-rendered accurately after a series of attempts to re-use sprites rendered earlier via a set of approximations.

Series of Experiments

In the series of experiments, subjects looked at pairs of video of graphics sequences generated by the Talisman architecture, and decided which member of the pair had the better image quality. The content of each video was identical; the videos differed only in terms of spatial and/or temporal resolution. In Experiment 1, only spatial resolution was explored. In Experiment 2, only temporal resolution was examined. In Experiment 3, spatial and temporal resolution were manipulated simultaneously. We examined each dimension in isolation before exploring any interactions that might be observed.

For each experiment, we explored several levels of degradation along a dimension and compared data on all levels of degradation against all other levels of degradation. This approach was adopted to explore the sensitivities of the human perceptual system along a wide range of degradations.

We held several hypotheses about the sensitivity to the degradations prior to running the experiments. We expected that, subjects would perform the best at making discriminations when the difference between the magnitudes of degradations were great (i.e. 100% vs. 25% of a gold standard image). When the difference in levels of degradation were small, we anticipated that performance would be poor. As subjects were challenged with comparisons that involved assessing the quality of image sequences produced with degradations at differences ranging between the greatest and smallest degradations, we predicted that performance would become increasingly poor. We suspected that such diminishment would not be linear [2]. We expected that somewhere between the smallest and largest differences there would be a perceptual leap where the differences would suddenly become distinguishable. Such nonlinearities would provide opportunities for leveraging resource allocation in flexible graphics architectures. We felt confident in these predictions but were more interested in the specific functions we would obtain for each dimension so as to instantiate mathematical models of expected perceptual cost used for guiding rendering decisions.

We did not have preconceived intuitions about the answers to several important questions. For example, we did not know what we would discover about the efficiency with which users would make quality assessments for simultaneous degradations along multiple dimensions. If we manipulated two dimensions simultaneously, would one of the two dimensions be more perceptually salient than the other? Would subjects ignore the non-salient dimension in favor of the salient dimension, or would there be interesting interactions between the two?

EXPERIMENT #1: SPATIAL RESOLUTION

Subjects

14 subjects participated in the first experiment. All subjects were 18 to 26 years old and had normal or corrected-to-normal vision. There were roughly equivalent numbers of males and females.

Stimuli

The video sequence used in this experiment showed two spaceships flying through a canyon. 10 permutations of this video were created using a Talisman simulator that was built previously to explore questions about the flexible graphics architecture. We generated multiple sequences at multiple levels of spatial resolution. The degradations were stepped at increasingly lower resolutions on a geometric scale with the largest changes in spatial resolution occurring at the highest spatial frequencies. Each of the 10 permutations was created by using an 86% step degradation in spatial resolution. Considering the spatial resolution as a percent of the highest possible quality, the 10 levels used were: 100%, 86%, 73%, 63%, 54%, 46%, 40%, 34%, 29%, and 25%.

The spatial degradation used 4x4 subsampling. Each textured triangle in the scene was sampled 16 times per pixel in a regular grid and then these samples were averaged to output a single pixel value. To generate the degraded examples, sprites with lower output pixel resolutions were used. These degraded sprites used fewer pixel samples, and so were more aliased.

The number of pixels used for each video was the total number of pixels available on the screen multiplied by the square of the degradation level. The highest quality video used 349,920 pixels (720 x 486), the second highest quality video used 258,800 pixels (720 x 486 x 0.86^2), and the lowest quality video used 21,870 pixels (720 x 486 x 0.25^2).

The video sequences were stored on an Abekas A65 digital disk recorder and displayed on a calibrated Sony NTSC display at a rate of 60 interleaved frames per second. 80 separate frames were computed for each video sequence and played back over a 1.33-second time period so that each interleaved frame showed a new image.

Procedure

The subjects in this experiment were initially briefed about the purpose of this experiment. They were told that we were interested in studying what differences in video quality people could and could not detect. Subjects were told to be as accurate as possible. They were not told that their reaction time would be recorded.

Over the course of a single trial, subjects saw a video played for 1.33 seconds followed by 500ms of a black screen, and then a second 1.33 second video. Once the second video began playing, subjects were prompted to make their decision as to which sequence had better image quality. Subjects were told that they could answer as soon as the second video began playing, but were not told to answer as quickly as they could; they were informed that the computer would wait indefinitely for their response, displaying a black screen once the second video finished. The subjects received no feedback about the correctness of their answers. After each user response, a note was posted to the screen stating that a new trial was about to begin and reminding the subject that the task was to choose the video with the better image quality. This trial break lasted 1 second.

A single block of trials consisted of each of the ten levels of image quality paired with each of the nine levels of a different degradation for a total of 90 trials or video pairs per block.

Order effects within a trial were not considered to be factor in the study design because the highest quality video is followed by the lowest quality video once, and the opposite ordering of the lowest quality video followed by the highest quality video is also played once. The order of the trials within a block is randomized. Between each block, subjects were presented with an opportunity to stop and rest for a self-selected period of time. Subjects responded to 11 blocks of trials, for a total of 990 trials per session. The first block of trials was considered a practice block, and the data from this block was not analyzed.

Two kinds of data collected from each trial: (1) The accuracy of subjects' responses, and (2) the amount of time it took subjects to respond from the onset of the second video. Although accuracy was always intended to be the primary data source, it was expected that reaction time would provide additional evidence about trends identified in the data. For each subject, 20 instances of each video sequence was paired with every other sequence. The total percent correct and median reaction time for each subject was recorded and then averaged across subjects. Median reaction time was used to minimize the effects on our analysis of reaction time of the few intermittent delays when viewers became distracted and failed to answer for an extended period of time.

RESULTS

Percent Correct

We noticed that subjects showed some tendency toward an inability to distinguish the quality of pairs of videos in two situations. These included trials with pairs of videos with proximal levels of spatial resolution degradation (*e.g.*, 46% vs. 40% of the best spatial resolution), and for greater spans quality for the higher-quality videos pairs (i.e., greater than 54% of the standard).

Table 1 shows the average percent correct assessments of degradation quality for each pair of videos. We assumed that answering at a level of 50% indicated that subjects were guessing about which video sequence had the higher image quality, while answering at a level of 100% would mean that subjects were confident about which video of the pair had higher image quality.

	100 %	86%	73%	63%	54%	46%	40%	34%	29%	25%
100 %	---									
86%	0.58	---								
73%	0.69	0.64	---							
63%	0.79	0.73	0.65	---						
54%	0.89	0.88	0.82	0.75	---					
46%	0.94	0.96	0.92	0.90	0.81	---				
40%	0.97	0.96	0.97	0.96	0.92	0.79	---			
34%	0.95	0.98	0.97	0.98	0.95	0.93	0.83	---		
29%	0.96	0.98	0.98	0.98	0.98	0.97	0.95	0.88	---	
25%	0.99	0.98	0.99	0.98	0.97	0.98	0.98	0.95	0.83	---

Table 1: Accuracy matrix for each pair of video comparisons in Experiment 1. Percent of best spatial resolution on both axis.

Figure 1 displays plots of the performance of the highest quality (no degradation or 100% and lowest quality or 25% of the best quality image) video sequences against each of the other sequences. We found that when the lowest quality video is compared against the second lowest quality video (25% versus 29% of the standard), there is a reasonably high (average of 83%) number of correct guesses. When compared against the next higher video quality (25% v. 34% of the standard), subjects were nearing perfect discriminability. This trend is not seen for the comparisons with the highest quality videos. When the highest quality video is compared with the second highest (100% versus 86% of the standard), subjects were responding with an average of only 58% percent correct, just slightly better than chance. Performance does not reach an average of 94% accuracy until the highest quality

video is compared with the 6th highest quality video (100% versus 46% of the standard).

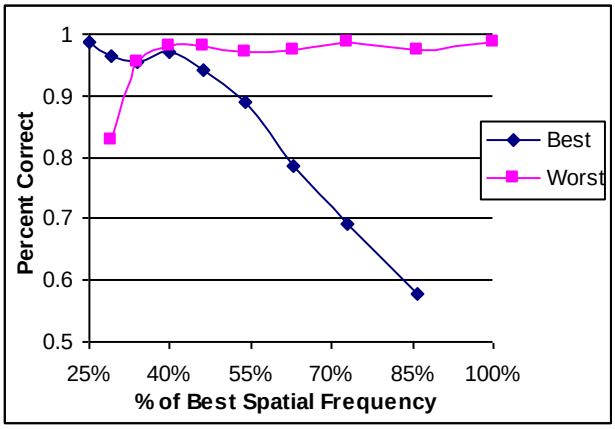

Figure 1: Performance accuracy for the best and worst spatial resolution videos as a function of all other degradation levels in Experiment 1.

Reaction Time

The reaction time data follows a similar trend as the accuracy data. Subjects took longer to decide between adjacent levels of spatial resolution and took comparatively longer at the highest quality videos.

Table 2 shows the average of the median reaction times (in milliseconds) to choose the higher quality video for each pair of videos. We interpret the faster reaction time as indicating that the subjects had less trouble making their decision and answered quickly despite not being told that the speed of their decisions was desired or would be measured. We interpret the longer reaction times as indicating a more effortful decision making effort. The reaction time for an individual subject was calculated by including data from all trials without regard for response accuracy. Because this is a forced choice study, we believe that incorrect choices were just as informative as correct choices and did not remove them from the analysis.

	100%	86%	73%	63%	54%	46%	40%	34%	29%	25%
100%	---									
86%	1364	---								
73%	1309	1338	---							
63%	1252	1246	1353	---						
54%	1159	1228	1265	1292	---					
46%	1039	1007	1091	1210	1212	---				
40%	963	1007	1017	1050	1106	1324	---			
34%	893	882	925	929	995	1125	1418	---		
29%	907	848	838	875	904	986	1068	1243	---	
25%	847	862	874	876	871	913	974	1112	1276	---

Table 2: Reaction time in milliseconds to make a quality decision for each pair of videos in Experiment 1. Percent of best spatial resolution on both axes.

Figure 2 shows the reaction time for the best and worst spatial resolution videos as a function of each of the other levels of degradation. We found that there is a rapid decrease in reaction time with progression from the worst video being compared to the second worst to being compared with the third worst video. The reaction time function for the worst quality video comparisons then plateaus over the rest of the levels of spatial resolution degradation values in spatial resolution.. The best spatial resolution video shows a consistently increasing trend from comparisons with the worst video up to the comparison with the second best video.

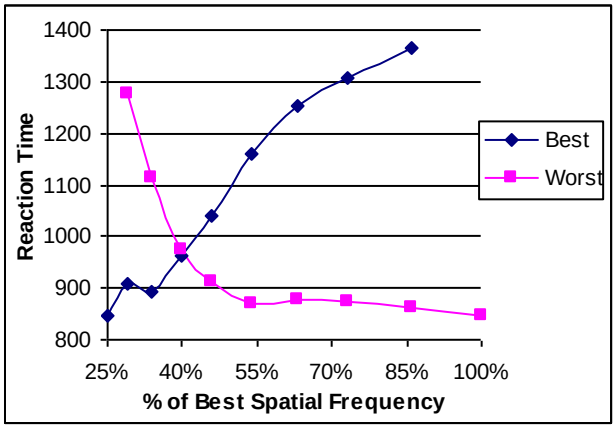

Figure 2: Reaction time for the best and worst spatial resolution videos as a function of all other degradation levels in Experiment 1.

DISCUSSION

discriminate between higher quality levels of spatially degraded videos. In essence, videos degraded down to 54% of a 100% standard in spatial resolution are viewed as being of equally high quality. Such a finding indicates that there is an opportunity for graphics rendering systems to allocate image rendering resources within such a range of quality.

EXPERIMENT #2: TEMPORAL RESOLUTION

Introduction

Experiment 2 follows Experiment 1 systematically, in that we continued our quest to determine how subjects' accuracy and reaction time relate to degradations in image quality, and, more specifically, to identify changes in these parameters that are most efficiently discriminated by subjects. Rather than explore degradations in spatial resolution, we examined the influence of degradations in temporal resolution as a function of an error tolerance in rendered sprites.

Subjects

15 subjects participated in this experiment. All subjects were 18 to 26 years old and had normal or corrected-to-normal vision. There were roughly equal numbers of male and female subjects. None of the subjects participating in Experiment 2 participated in Experiment 1.

Stimuli

The video sequence used in this experiment was the same as the previous experiment. Two spaceships flying through a canyon were displayed to subjects, and their task was to indicate the higher quality member of the pair. 10 permutations of a standard quality video (no temporal resolution degradation) were created using the Talisman simulator.

Degradation in temporal resolution was created by separating the screen into a series of contiguous objects or *sprites*. Whenever a measure of error in the fidelity of the shape and location of sprites in the image exceeds a threshold defined by the level of degradation, the sprite is redrawn with appropriate fidelity and placed in the correct location. Until the error metric is exceeded sprites are re-rendered with approximation algorithms. This introduces errors in the shape and placement of the sprite. Additional details of this graphics approximation method are described in [1,2]. Lower levels of temporal degradation are associated with lower tolerances to infidelity in the location and shape of sprites.

As with the spatial resolution experiments, degradations were performed on a geometric scale with the larger changes in temporal resolution occurring at the higher temporal resolution levels. Each of the 10 permutations was created by using 86% step degradations in temporal resolution from the 100% standard video. Looking at temporal resolution as a percentage of the highest quality, the 10 levels used were: 100%, 86%, 73%, 63%, 54%, 46%, 40%, 34%, 29%, and 25%.

As in Experiment 1, we used the Abekas and a NTSC screen to play the video sequences for this study.

Procedure

The procedure for this experiment was identical to Experiment 1.

RESULTS

Percent Correct

As was observed in the previous experiment, if the levels of temporal resolution degradation presented in a trial were proximal (i.e., 40% vs. 46% of the 100% standard video), the subjects showed some tendency to guess. As in Experiment 1, there was a greater tendency to guess between higher quality temporal resolution video sequences than between the lower quality temporal resolution video sequences. We found that this effect is significantly more pronounced with the temporal resolution degradations than with that of the spatial resolution degradations. Subjects answered at the rate of chance for comparisons between the four highest levels of temporal resolution, and did not reach an accuracy of 0.9 or better until these high quality videos were compared to videos with an image quality of 34% or lower of the standard.

Table 3 shows the average percent correct for each pair of videos in Experiment 2. Answering at a level of 50% suggests that the subjects cannot discriminate the quality of the images, while answering at a level of 100% suggests that the subjects were confident about which the ordering over image quality.

	100 %	86%	73%	63%	54%	46%	40%	34%	29%	25%
100 %	---									
86%	0.49	---								
73%	0.53	0.54	---							
63%	0.52	0.56	0.53	---						
54%	0.56	0.55	0.52	0.52	---					
46%	0.66	0.67	0.65	0.62	0.61	---				
40%	0.80	0.79	0.82	0.78	0.79	0.68	---			
34%	0.92	0.91	0.91	0.93	0.94	0.91	0.84	---		
29%	0.92	0.94	0.93	0.95	0.95	0.93	0.91	0.85	---	
25%	0.94	0.95	0.95	0.96	0.95	0.95	0.95	0.93	0.82	---

Table 3: Accuracy matrix for each pair of video comparisons in Experiment 2. Percent of best temporal resolution on both axes.

Figure 3 plots the performance of the highest quality (standard, no degradation) and lowest quality (25% of the standard image quality) videos against each of the other levels of quality. When the lowest quality video is compared to the second lowest quality video (25% versus 29% of the standard), there is an average percent correct of 82%. When compared to the next higher video quality (34%), subjects seem to approach perfect discriminability. This trend is not seen for the comparisons against the highest quality video. When the highest quality video is compared with the second highest (100% v. 86% of the

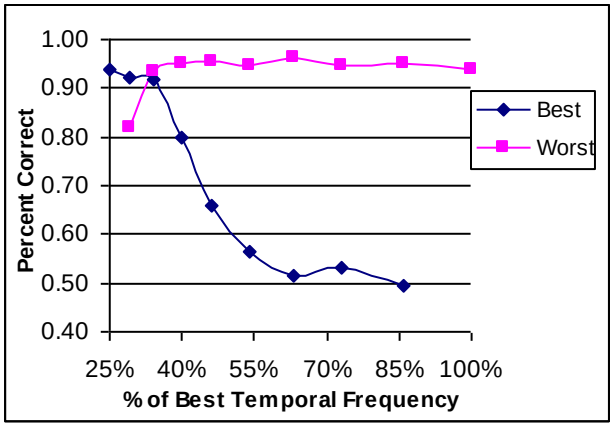

Figure 3: Performance accuracy for the best and worst temporal resolution videos as a function of all other degradation levels in Experiment 2.

standard), subjects respond at chance with an average of 49% percent correct. Performance stays very close to chance through degradation levels represented by images at 54% the quality of the standard. video (100% versus 34%

of the standard). Performance does not reach 92% percent correct, on average, until the highest quality video is compared with the eighth highest quality

Reaction Time

The reaction time data shows very similar trends to the percent correct data. Subjects took longer on average to decide between adjacent levels of temporal resolution and took longer on average to decide between the higher quality temporal resolution videos.

Table 4 shows the average of the median reaction times in milliseconds for each pair of videos. We interpret the faster reaction times to mean that the subjects had less trouble making a choice decision and answered quickly, despite not being instructed to decide as quickly as possible. The average reaction time datapoint for an individual subject was calculated with data from all trials without regard for response accuracy.

	100 %	86%	73%	63%	54%	46%	40%	34%	29%	25%
100 %	---									
86%	1491	---								
73%	1463	1456	---							
63%	1452	1507	1460	---						
54%	1554	1474	1523	1475	---					
46%	1452	1452	1421	1456	1440	---				
40%	1388	1302	1441	1421	1374	1407	---			
34%	1192	1282	1174	1188	1211	1255	1345	---		
29%	1102	1065	1117	1107	1032	1125	1182	1341	---	
25%	1057	1078	1042	1040	1020	1085	1113	1170	1334	---

Table 4: Reaction time in milliseconds to make a quality decision for each pair of videos in Experiment 2. Percent of best temporal resolution on both axes.

Figure 4 shows the average reaction times for the best and worst temporal resolution videos compared against each of the other temporal resolution degradation levels. There is a rapid decrease in reaction time when the worst video is compared to the second and third worst videos, and then reaction time plateaus over the rest of the videos. The function for comparisons with the image with the best temporal resolution function demonstrates an increasing trend from comparisons to the worst video (847ms) up to the comparison with the sixth (46% of the best video), then plateaus over the rest of the quality levels.

DISCUSSION

As was the case in Experiment 1, Experiment 2 has provided us with some interesting insights into how users discriminate degradations in image quality. Specifically, subjects in this study again demonstrated a difficulty in discriminating images of quality in less severe ranges of degradation in temporal resolution. Such information about the inability of users to discriminate the qualities in

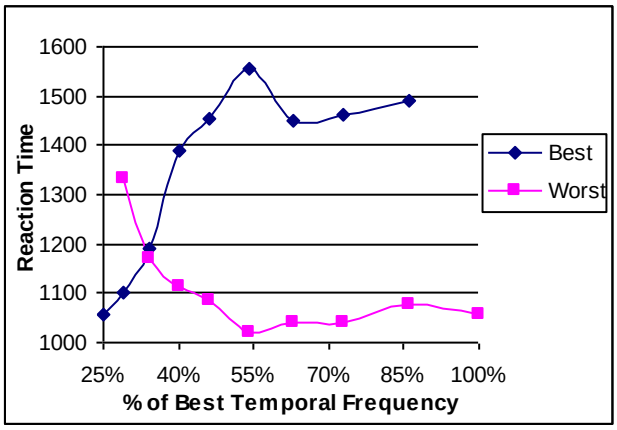

Figure 4: Average reaction time for the best and worst temporal resolution videos as a function of all of the other degradation levels in Experiment 2.

this region provides an opportunity for making flexible decisions about the allocation of resources in graphics systems.

In Experiments 1 and 2, we characterized the ability of users to discriminate among images of a range of quality for both spatial resolution and temporal resolution degradations. In Experiment 3, we performed a third experiment to examine subjects' abilities to discriminate videos in a wide range of combinations of degradation using both temporal and spatial frequencies.

EXPERIMENT #3: COMBINING SPATIAL AND TEMPORAL DEGRADATIONS

Subjects

14 subjects participated in this experiment. All subjects were 18 to 26 years old and had normal or corrected-to-normal vision. There were roughly equal numbers of male and female subjects. None of the subjects had participated in Experiments 1 or 2.

Stimuli

The video sequence used in this experiment was the same as the previous experiments. Nine permutations of this video were created using a Talisman simulator to degrade both the spatial and temporal resolution of this video. Perceptually similar steps of spatial and temporal resolution were chosen from the previous experiments. 100%, 73%, and 54% of the best spatial resolution video and 63%, 46%, and 34% of the best temporal resolution video were chosen for their perceptual distinctiveness (as discovered in Experiments 1 and 2). Each level of spatial resolution was combined with each level of temporal resolution to create the 9 videos.

As in Experiments 1 and 2, we used the Abekas and a NTSC screen to play the video sequences.

Procedure

The procedural sequence for each trial was identical to Experiments 1 and 2. There were several overall differences in this experiment because of the different nature of the stimuli.

A single block of trials consisted of the nine different videos compared with a standard video twice in each order for a total of 36 trials per block. We selected to use as a reference video the middle value of both spatial resolution (73% of best) and temporal resolution (46% of best). The order of the trials within a block was randomized. After every other block, subjects were presented with an opportunity to rest for a self-selected period of time. Subjects responded to 11 blocks of trials for an overall session total of 396 trials. The first block of trials was designated a practice block, and the data from this block was not analyzed.

Two kinds of data were collected for each trial: (1) the resolution that the non-standard video was selected over the standard video and (2) the amount of time it took subjects to respond from the onset of the second video. Each subject was presented with 40 instances of each video paired against the standard video. In contrast with the other studies, we cannot analyze the results in terms of a correct assessment of quality because quality on one dimension was manipulated independently from quality in the other dimension. Since for any pair of videos higher video quality cannot be defined, we instead use the percent of non-standard video preference and the median reaction times were taken for each subject and then averaged across subjects.

RESULTS

Preference Data

Results from Experiment 3 show that, when either the level of spatial or temporal resolution is increased or decreased from the standard video, the likelihood of choosing the non-standard video increases or decreases respectively. When both the level of spatial and temporal resolution increases or decreases from the standard video, the likelihood of choosing the non-standard video increases or decreases respectively at a rate greater than when just one of the independent variables is adjusted. In other words, there was a combinatorial effect of the two dimensions of temporal and spatial resolution degradation on choice performance.

Two interesting phenomena were observed when one dimension, but not the other, was of higher quality compared to the standard. When spatial resolution is increased and temporal resolution is decreased from the standard, and when spatial resolution is decreased and temporal resolution is increased from the reference video sequence. In both of these cases, subjects selected the standard video sequence as the one with the better quality over 80% of the time.

Table 5 shows the average probability that subjects selected the non-standard video over the reference video. Although the subjects were told to choose the video with the higher image quality, we can no longer call the dependent variable "percent correct" because there is no longer a perceptually superior member of a pair. In two cases, the image quality of one dimension increases while the image quality of the other decreases. One of the nine comparisons is identical to the reference video, which subjects should respond to at chance. For this comparison, subjects chose identical video 47% of the time, which is essentially chance.

		Spatial Resolution		
		100%	73%	54%
Temporal Resolution	63%	0.75	0.61	0.16
	46%	0.64	0.47	0.12
	34%	0.18	0.12	0.06

Table 5: Likelihood of selecting each video when compared against the reference video in Experiment 3.

A two-way analysis of variance (ANOVA) showed reliable main effects for spatial [$F(2,13) = 137.8$, $p < 0.01$] and temporal frequency [$F(2,13) = 56.0$, $p < 0.01$] as well as a reliable interaction effect [$F(4,13) = 24.5$, $p < 0.01$]. Given that we chose these levels of spatial and temporal frequency for their discriminability, the main effects are unsurprising. The interaction effect is more interesting, saying that these two dimensions do not act in isolation, but rather are at least slightly additive.

We discovered in Experiments 1 and 2 that there is increasing discriminability along each dimension as viewers move to comparisons at the higher quality video sequences to the lower quality sequences. Thus, it was not surprising to find in Experiment 3 that, in the four conditions that considered only one dimension of degradation, that subjects could more accurately discriminate the videos when the quality decreased for either dimension than when the quality increased for the respective dimension. Subjects accurately identified a decrease in spatial resolution an average of 88% of the time, but only correctly identified an increase in spatial resolution an average of 64% of the time. Likewise, subjects accurately identified a decrease in temporal resolution an average of 88% of the time, but only identified an increase 61% of the time, on the average.

We believe that subjects may change their dimensional focus during each trial. If people preferentially watch for changes in either spatial or temporal resolution, then in the two cases where one dimension had higher quality while the other had lower quality, the selection rate should have been higher than 50% for one case and lower than 50% for the other. For example, if people were selectively attending to temporal resolution degradations, then the selection rate should have been lower than 50% when temporal resolution decreased and spatial resolution

increased (relative to the reference video sequence). Likewise in this scenario, we would expect that the selection rate would have been higher than 50% when the temporal resolution increased and spatial resolution decreased (relative to the reference video). Instead, people are selecting the non-standard video more when either spatial or temporal resolution is decreased.

The most plausible explanation (though not the only one) of these findings is that subjects are noticing the most salient dimension. When spatial resolution increases and temporal resolution decreases compared to the reference video, subjects choose the standard video as higher quality at a rate almost identical to the rate when spatial resolution is the same as the standard and only temporal resolution has decreased. There is a difference (6%) between the case where spatial resolution increases and temporal resolution does not change. However, this does not appear to be a significant difference. When temporal resolution increases while spatial resolution decreases compared to the reference video, the rate that the standard video is selected as the higher quality video sequence is remarkably similar to our findings in Experiment 1, when spatial resolution was decreased in isolation. In both cases, the dimension that was observed to be more salient in isolation was a better predictor of performance in the mixed conditions.

Reaction Time

Table 6 shows the average of the median reaction times for Experiment 3. The pattern of data maps nicely to the preference data described above. It took subjects longest to choose between the videos when they were identical. For each of the other comparisons, the further the probability of selecting the non-standard is from 50% (see previous section), the faster the reaction time. It took subjects longer to respond when either spatial or temporal resolution was increased rather than decreased, but the rate of responding when both spatial and temporal resolution were increased or decreased was faster than when only one dimension was changed. Again, the reaction time data provides converging evidence for a combinatorial effect with the two dimensions studied.

		Spatial Resolution		
		100%	73%	54%
Temporal Resolution	63%	1851	1912	1695
	46%	1892	1936	1657
	34%	1587	1571	1416

Table 6: Average reaction times in milliseconds for Experiment 3.

We performed three studies of the perception of image quality as a function of degradations employed in a flexible graphics architecture. The studies were framed to answer specific questions about the control of flexible procedures made available in an experimental rendering architecture. Our experiments were designed explicitly to study the sensitivities of the human visual system to the specific degradations defined by the capabilities and tradeoffs of the system. However, we suspect that the results have applicability to other systems employing similar rendering approximation methods.

For the range of degradations studied, we demonstrated that the ability of subjects to discriminate proximal changes in the degradation of the quality of images varies significantly as a function of the general level of quality of video sequences being compared. More specifically, we found that subjects are more sensitive to small changes in the quality of images when these changes occur in regions of lower quality than they are to changes in degradation in regions of higher quality.

We worked to characterize the functional form of discriminability of degradations for specific dimensions of degradation employed by Talisman. The functions reveal plateaus and sharp swings in discriminability. Flexible graphics rendering systems can exploit such functional forms and qualitative information to make intelligent resource allocation decisions.

In studies of the influence of multiple degradations on perceived quality, we found that multiple degradations acting simultaneously on images tend to cooperate, further decreasing the perceived quality of the images more than either degradation does alone. Furthermore, we noted, for the simultaneous degradations in spatial and temporal resolution over the ranges of quality studied in our experiments, that the dimension of degradation with the greatest salience tends to dominate the influence of other dimensions on perceived quality.

We are currently working to leverage these results to update the models of perceptual cost that are used for dynamically controlling the allocation of computational resources in Talisman. As part of continuing work on the Qualia project, we are pursuing questions about the mapping between the studies of discriminability of image quality and measures of the overall perceived quality of a viewer's experience as a function of multiple dimensions of degradation and graphics content.

References

1. E. Horvitz, M. Czerwinski, K. Larson, and J. Lengyel. *The Qualia Project: Perception-Directed Rendering of Graphics under Scarce Resources*. Microsoft Research Technical Report, Microsoft Research, Fall 1997.

2. E. Horvitz and J. Lengyel. Perception, attention, and resources: A decision-theoretic approach to graphics rendering. In *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence* (UAI-97)}, pages 238-249, Providence, RI, 1997. Morgan Kaufmann Publishers, San Francisco, CA.

3. J.Lengyel and J.Snyder. Rendering in a layered graphics architecture. In *Proceedings of SIGGRAPH 97*. SIGGRAPH, August 1997.