On the Parameters of Human Visual Performance
An Investigation of the Benefits of Antialiasing

K. S. Booth, M. P. Bryden, W. B. Cowan, M. F. Morgan and B. L. Plante

Computer Graphics Laboratory
University of Waterloo
Waterloo, Ontario, Canada N2L 3G1
519/888-4534

Abstract. A two-part experiment was conducted to investigate the effects of aliasing artifacts and screen resolution on a simple visual recognition task. The results indicate that in many cases far less realism may be necessary in synthetic computer-generated imagery than is often assumed in the literature. The first part of the experiment comprised a subjective rating of image quality, the second part measured task effectiveness of image quality. In the second part subjects were asked to discriminate between images of two types of objects built from cubes, similar to objects used in experiments involving mental rotation.

At higher resolutions the elimination of aliasing artifacts did not significantly improve subjects' performance. At intermediate and low resolutions, comparable to what might be used for iconic menus, the reduction in aliasing artifacts resulted in improved performance. The subjective ratings indicate that for both high and low resolution the elimination of aliasing artifacts does not improve "quality," whereas images rendered at intermediate resolutions are significantly degraded by aliasing artifacts to the extent that antialiasing improves the subjective rating.

An interpretation of these results is given in the context of an ongoing research program aimed at identifying the parameters of real-time human performance for graphics workstations.

Résumé. Nous avons réalisé une expérience en deux parties afin d'analyser les interactions entre la résolution d'un écran et les problèmes d'aliasing dans le contexte d'une simple tâche de reconnaissance visuelle. Les résultats indiquent que dans beaucoup de cas, les images réalisées par ordinateur n'ont pas besoin d'être aussi réalistes qu'indiquées par la littérature sur le sujet. La première partie de l'expérience consistait en une évaluation subjective de la qualité des images. Tandis que la deuxième partie mesurait l'impact de cette qualité sur la tâche de reconnaissance visuelle. Dans cette deuxième partie, les sujets devaient différencier les images de deux types d'objets formés par un amalgame de cubes. Ces objets étant semblables à ceux déjà utilisés pour des expériences de rotations mentales.

Dans le cas d'un écran à haute définition, l'utilisation de techniques d'antialiasing n'affectait pas de façon significative la performance des sujets testés. Pour des définitions basse ou intermédiaire, définitions comparables à celles qu'on retrouve dans des menus à base d'icones, la réduction de l'aliasing conduit à une meilleure performance. L'évaluation subjective indique que pour une haute ou basse définition, l'élimination de l'aliasing n'améliore pas la "qualité" des images. Alors que pour une définition intermédiaire, les images sont suffisament dégradées par l'aliasing que des techniques d'antialiasing améliorent le jugement qualitatif.

Une interprétation de ces résultats est donnée dans le contexte d'un programme continu de recherche cherchant à identifier les paramètres intervenant dans l'interaction, en temps réel, entre un humain et une station graphique.

Keywords. antialiasing, image quality, vision.

Introduction

Most graphics displays are used by people performing particular tasks and needing information from the display in order to perform those tasks. Significant resources are allocated to the production and presentation of realistic imagery containing that information, yet we have little theory and few rules of thumb to evaluate the effectiveness of that imagery on graphics displays.

There have been studies that examine the effectiveness of various computational methods for enchancing the realism of computer-generated images. Atherton and Caporael derived measures of effectiveness for rendering curved surfaces using polygonal approximations [1]. Their measures were based on subjective judgments made by subjects who compared images rendered using different shading rules and degrees of approximation. The results indicated specific cutoff values beyond which additional computation provided little gain in realism. In drawing their conclusions, Atherton and Caporael noted that at or beyond these cutoffs the polygon-based techniques produced "image quality reasonably close to an 'ideal' image."

Cohen and Greenberg reported an experimental study by Meyer and Rushmeier to validate the realism of the radiosity method for rendering environments [3] [4]. Subjects were shown photographs of a physical model constructed of wood and cardboard (painted to provide diffuse surfaces) and photographs of synthetic images of the same scene computed using the radiosity technique. The results indicated that subjects were unable to discriminate between the two photographs. Cohen and Greenberg claimed "conclusive validation of the radiosity method," in part because of the experiment.

While these results provide useful information for those whose primary concern is attaining the maximum possible realism in synthetic images, the studies do little to address the underlying problem of providing effective information at reasonable cost in graphics displays.

For example, in a flight simulator it is less important to deceive a pilot into believing that he is in a real airplane than it is to provide him with all of the information in the right form to allow him to function as if he were in a airplane. In practice, many compromises must be made to balance the high cost of real-time synthetic image generation against the actual needs of an application. These compromises may result in much less than full realism in the images.

To continue the example of a flight simulator, it seems to us that studies such as those of Atherton and Cohen seem primarily aimed at asking how much computational work is necessary to deceive the pilot; they examine the work necessary to make the image indistinguishable from reality rather than determining the characteristics of the image that are necessary to enable the pilot to perform the task at hand. Indeed, both studies explicitly make reference to the imitation of reality as a criterion for evaluating rendering techniques. We have taken a more pragmatic approach in our work and have embarked on a series of investigations whose goal is the determination of the parameters of real-time human performance, with a particular emphasis on performance while interacting with imagery presented on graphics workstations.

In areas other than graphics imagery there have been numerous studies examining the performance of users on many specific tasks. Nickerson amply surveys this work and provides extensive references to the literature [6]. Typical of this area of research are studies of command sets for text editors, readability of individual characters in dot matrix printers, and the use of blinking to direct attention in visual displays.

More directly related to computer graphics are investigations of some aspects of color usage on graphic displays, an area not characterized by intensive computational requirements. Schwarz, Beatty and Cowan's study suggests appropriate means for specifying color in a task where the goal is to provide a mechanism that maximizes the user's effectiveness in performing the task [7]. Ware and Beatty examined various strategies for using color to enhance the display of multi-dimensional data, again with the goal of maximizing the user's effectiveness in performing a selected task: visual cluster analysis [9].

Experiments of this kind provide useful engineering guidelines for the design of computer software, hardware devices and user interfaces, but there has been very little of this work performed to provide such guidelines for image rendering quality in computer graphics, especially in the area of interactive graphics displays.

The spatial structure of synthetic images is typically characterized by intensive computation, often in real-time for applications such as flight simulators. The conclusions drawn from studies such as those by Atherton and Caporael or Cohen and Greenberg tend to ignore the realities of computational load and may lead to inappropriate and overly stringent goals for practical image rendering systems. This paper discusses three aspects of an ongoing research program and presents the results of one experiment that we have conducted that was designed to determine the level of image quality that is required for a specific recognition task.

We first describe the general directions of our research into the parameters of real-time human performance, the role that image quality plays in that performance and the importance we feel that this research has for the design of future graphic workstations.

We then present the results of the experiment we conducted involving the effect of image quality, as measured by spatial resolution and the presence or absence of spatial aliasing artifacts, on the performance of subjects in a simple recognition task. We interpret those results as they pertain to graphic workstation rendering systems and the recognition of icons in menu-based applications.

Finally, we discuss the methodology that we are developing for use in future studies and highlight those aspects of our approach that we feel are unique or important.

Parameters of Real-Time Human Performance

In a talk presented in 1985 to this audience, Allen Newell criticized current research on computer-human interfaces for concentrating too heavily on the exacting details of the psychological mechanisms taking place [5]. Instead, he urged that we seek to understand fundamental limits of computer-human interaction in an "engineering" sense so that designers of systems have parameters to work with that can be tuned to attain appropriate levels of performance. Newell concluded that methodologies for building interactive systems must include tools that allow performance constraints to be specified as part of the design at its inception, rather than being applied after the design, during acceptance testing.

Our principal interest is in establishing such rules of thumb for using graphic displays in workstation environments where real-time performance is important. This falls naturally into two parts: the performance characteristics of systems that users expect to see for real-time applications and the properties a designer must be able to specify and control in order to produce such systems. As a first step in this research we are conducting a multidisciplinary project whose goal is the design and control of cognitive psychology experiments, most of them associated with visual stimuli.

A subject in such an experiment sees stimuli presented in real-time. Spatial, temporal and chromatic artifacts must be significantly below a subject's perceptual threshold to maintain the integrity of the experiment. Interaction must be instantaneous with respect to human time scales (a few milliseconds) and it must be natural enough that the experimenter measures the effect in which he or she is interested, not properties of the interaction technique.

It is here that we may differ with the standards of image realism used in the studies cited previously. We are interested in determining the level at which the artifacts being investigated by the psychologists are the only artifacts influencing the outcome. For this to be true, we must understand the thresholds at which such artifacts interfere with real-time human performance. In some cases this may lead to criteria more stringent than merely the subjective illusion of "realism," although in other cases we expect less stringent criteria.

To facilitate an effective implementation of systems that achieve these criteria, the designer should have an easy and straightforward task in translating his or her conception of the experimental procedure into a functioning computer program (we note that this conception almost always involves multi-tasking and real-time). Development and monitoring tools must be available to allow the designer to be confident that the experiment is as desired, even though the discrepancies involved may be beyond the limits of the designer's own perception.

Similar requirements apply to the design of any system which interacts with humans. Other systems, such as bank teller machines or flight simulators, have no reason to go beyond the artifact-free performance of the psychology experiment. Thus, in a fairly strong sense we can claim that the real-time requirements of cognitive psychology experiments provide a good measure of the "envelope" within which real-time systems must operate if they are to be successfully used by humans.

Task-Oriented Image Quality

A two-part experiment was conducted to determine the relationship between aliasing artifacts and image quality. The first part of the experiment involved a subjective rating of a variety of images depicting simple geometric objects constructed of cubes. The second part required subjects to discriminate between two classes of those objects based on the number of cubes in the objects.

The objects were chosen to be similar to those used in our on-going investigation of mental rotation phenomena, in part because we wanted to validate our belief that the use of computer-generated 35mm slides had not introduced aliasing artifacts into the images that would interfere with the underlying cognitive processes that were being examined. Each image depicts a fully shaded 3-D rendering of an object composed of either 9 or 10 cubes (Shephard and Metzler used computer-generated line drawings of 10-cube objects in their original mental rotation experiments [8]).

The various images differ in the degree of antialiasing (we used only two levels: none at all and 8×8 subpixel sampling) and rendering resolution (we varied the number of "tiles" in a full image from a high of 512×512 to a low of 64×64). We refer to the images that were computed using subpixel sampling as "antialiased" images and those for which a single sample determined the shade for the tile as "aliased" images. More sophisticated antialiasing algorithms are often used in practice, but 8×8 subpixel sampling was deemed sufficient for the images in our study.

We expected to observe a marked preference for "realistic" renderings of the objects in the first part of the experiment, both in terms of resolution and antialiasing quality. The subjective rating was made primarily to allow us to correlate perception of image quality with its impact on performance in the recognition task that was conducted in the second part of the experiment.

As predicted, subjects preferred antialiased images to those with aliasing artifacts and they preferred images rendered at higher resolution to those rendered at lower resolution. Perhaps not surprisingly, the preference for antialiased images was reduced at the highest resolution, presumably because the limits of human visual acuity were being approached. Although the subjective rating dropped consistently for both aliased and antialiased images as the resolution decreased, it appears that at very low resolution the images are so degraded from the original objects that subjects did not discriminate between the aliased and antialiased images.

The second part of the experiment required subjects to discriminate between objects composed of 9 cubes and objects composed of 10 cubes; each object was displayed in various 3-D spatial orientations. This particular task was chosen because of its similarity to the recognition task used in mental rotation studies (subjects there are asked to decide whether two objects are the same or mirror images of each other; each object may have undergone a 3-D spatial rotation).

We expected subjects to perform the task better (more quickly) when presented with high resolution antialiased images. The purpose of the experiment was to obtain a quantitative measure of performance as a function of resolution and level of aliasing.

The expected improvement did in fact happen, but only for lower resolution images. Not until image resolution dropped below 3×3 tiles (an effective resolution of 171×171 pixels) was there a significant difference in subjects' response times between aliased and

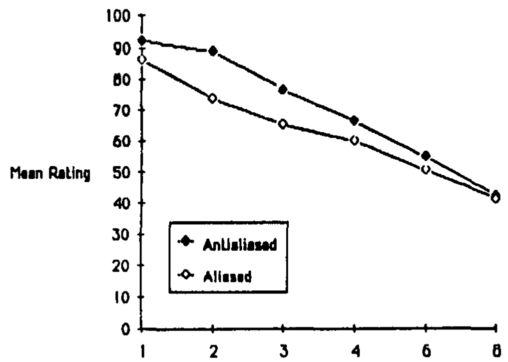

Figure 1. Subjective mean ratings of picture quality as a function of pixel (tile) size for images with and without aliasing artifacts.

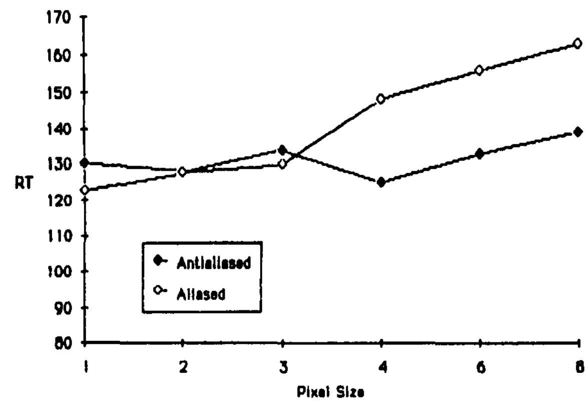

Figure 2. Discrimination response time (RT) as a function of pixel (tile) size for images with and without aliasing artifacts.

antialiased images. After that, performance for both levels of antialiasing degraded consistently as a function of decreasing resolution.

The lack of effect in the recognition task between aliased and antialiased images at high resolution is in marked contrast to the subjective ratings, where a preference for antialiased images was expressed as soon as image resolution dropped below 1X1 (an effective resolution of 512X512) so even at 2X2 tiling the preference was strong. In fact it was strongest in this case; the differential preference dropped steadily after that as a function of decreasing image resolution.

We draw a number of conclusions from this experiment.

- The first is that our on-going investigation of mental rotation phenomena has not been compromised by the introduction of aliasing artifacts (the images used in that work are 512X512 with the same level of antialiasing tested here).
- Second, there appears to be a difference between assessing image quality in terms of how closely a rendering algorithm approximates physical realism versus how effectively rendered images assist in performing a specific task.
- Finally, for many practical applications 512X512 or even lower resolution may be sufficient for some tasks, especially if antialiasing is used to moderate the artifacts introduced by lower resolution.

This last conclusion must be interpreted carefully. Our experiment examined only a single recognition task. We expect that similar results would apply to other tasks such as discriminating among icons used in a menu-driven application on a graphics workstation (with perhaps a shift along the curve to compensate for the overall scale change between our objects, which occupied much of the visual field, and the icons, which typically cover a smaller portion of the visual field) although we would be hesitant to project our results to the flight simulator example cited earlier.

One reason for our caution in generalizing these findings to other tasks is a study that was conducted for a prototype of an air traffic control system. The prototype featured a real-time graphic display in which line-drawing images representing the actual planes were generated for a 512X512 raster display. The system employed no antialiasing. The spatial artifacts introduced by the rendering algorithms were exacerbated by the movement of the airplanes, so that a temporal pattern emerged for each type of plane as it moved in the traffic pattern. Particularly when gliding in on final approach, each type of plane had a characteristic "click" that allowed its rate of descent to be easily determined.

We can only imagine the results of a hypothetical design based on a hypothetical study using this prototype. We envision great enthusiasm when it is discovered, to the delight of the designers, that air traffic controllers are indeed able to work more efficiently using the prototype display system! Because of this, a multi-million dollar contract might be let to build a production air traffic control system, of course employing ultra-high resolution displays with state-of-the-art antialiasing algorithms, because "everyone knows" this will improve the air traffic controllers' performance in the recognition task. There will be great wringing of hands (and possibly necks) when it turns out that the performance of the expensive production system lags far below that of the simpler prototype because it lacks the characteristic "clicks" that provided clues to the aircrafts' rates of descent.

These temporal aspects of spatial aliasing (and the more complicated interaction between spatial and temporal aliasing) remain an area for further study. That research should remain closely linked to specific tasks, trying to discover classes of tasks that produce similar performance constraints. Only in this way can the misinterpretation that could have occurred in the hypothetical air traffic control study be avoided.

Future Directions

The work reported here is a first attempt to quantify the effects of artifacts within computer-generated imagery. Subsequent experiments to be performed will test the validity of displaying visual stimuli, usually displayed using card tachtistoscopes, directly on crt displays. This rather specialized line of research actually addresses fundamental questions about the presentation of analog images in a quantized medium. Thus, it involves questions concerning spatial, temporal and chromatic artifacts introduced by real-time display technology. Once known, we believe that these artifacts can be controlled by combining our experimental findings with known properties of the human visual system to produce displays that are artifact-free in all three domains over a much wider range of parameters than is currently possible.

The results reported here should be extended to determine the interaction between block size and image size. In our study the block size was approximately four times the largest tile size, large enough to discount this effect in the analysis. The "clicks" noted in the air traffic control prototype are examples of the interaction between block size and tile size where there is a significant effect on the task being performed. Block size may be a factor in icon recognition for menus because icons are usually smaller than the figures we used. We intend to continue this experiment using blocks of different sizes to determine if the antialiasing effects we found are related to an absolute limit in the human perceptual system or if they are related to an interaction between size of detail in the image and tile size.

The results should also be extended to examine more complicated tasks. The discrimination task we tested is quite simple. The mental rotation tasks that we looked at prior to this study were more complicated. Those studies used only high resolution, antialiased images. A follow-up experiment will examine the effect of aliasing artifacts in the mental rotation task to determine if the same cutoff levels exist for the more complicated task or if a higher degree of image quality is required.

The final step in achieving our goal of understanding the parameters of real-time human performance will be to apply our research on display quality to a complete system for the design and control of cognitive psychology experiments. Because the design tools useful to the experiment designer are very similar to the design tools needed by the designer of any real-time interactive system (be it a humble bank teller machine or an exotic flight simulator), we expect to derive a family of tools for a broad range of applications that differ only in their level of performance.

The tools necessary to permit the designer to build and monitor the experiments will be based on a real-time, multi-tasking operating system kernel customized for use in cognitive psychology experiments. Preliminary work on aspects of that project has been reported elsewhere [2].

Acknowledgement

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada through the strategic grant program and a number of operating grants. Additional equipment support was provided by Digital Equipment Corporation of Canada. The Résumé was prepared by Alain Brossard.

References

[1] Peter R. Atherton and Linnda R. Caporael, A subjective judgment study of polygon based curved surface imagery, *CHI '85 Proceedings* (April, 1985) pp. 27-34.

[2] Kellogg S. Booth, William B. Cowan and David R. Forsey, Multitasking support in a graphics workstation, *Proceedings 1st International Conference on Computer Workstations* (November, 1985) pp. 82-89.

[3] Michael F. Cohen and Donald P. Greenberg, The hemi-cube — A radiosity solution for complex environments, *Computer Graphics 19:3* (July, 1985) pp. 31-40.

[4] Gary W. Meyer, Holly E. Rushmeier, Michael F. Cohen, Donald P. Greenberg and Kenneth E. Torrance, Assessing the realism of computer graphics images, to appear in *ACM Transactions on Computer Graphics*.

[5] Allen Newell, unpublished address to CHI '85 Conference (April, 1985).

[6] Raymond S. Nickerson, *Using Computers: The Human Factors of Information Systems*, MIT Press (1986) 432 pages.

[7] Michael W. Schwarz, William B. Cowan and John C. Beatty, An experimental comparison of RGB, YIQ, LAB, HSV and opponent color models, to appear in *ACM Transactions on Computer Graphics*.

[8] Roger N. Shepard and Jacqueline Metzler, Mental rotation of three-dimensional objects, *Science 171* (1971) pp. 701-703.

[9] Colin Ware and John C. Beatty, Using colour to display structures in multidimensional discrete data, *Color Research and Application 11* (June, 1986) pp. S11-S14.

[10] Marceli Wein, personal communication.

Appendix A — Experimental Design

A collection of 35mm color slides depicting random block configurations, similar to those used in mental rotation studies by Shephard and Metzler [8], were produced using an Adage RDS 3000 frame buffer and a Dunn 631 film recorder. Images were generated on the frame buffer as 512×512×24 full color images, gamma-

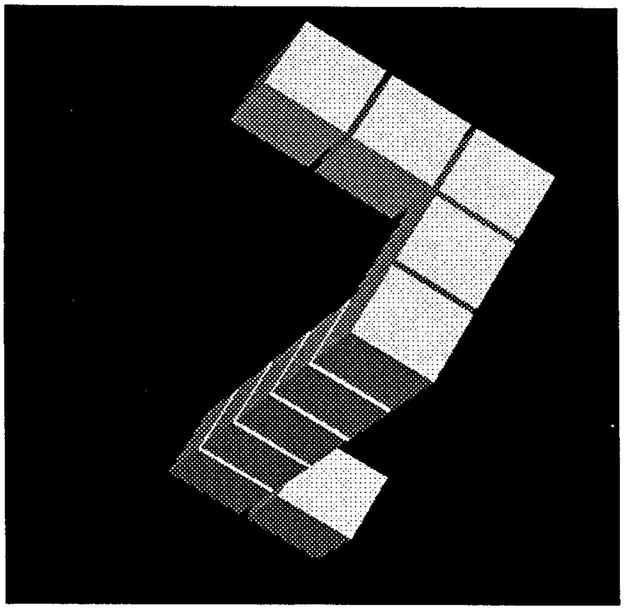

Figure 3(a). A 512X512 image of an object with 10 cubes. Aliasing artifacts have been reduced using 8X8 subpixel sampling.

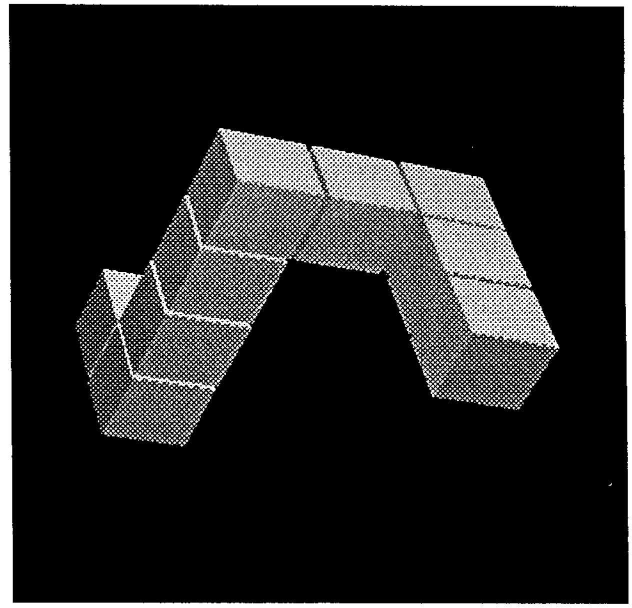

Figure 3(b). A 512X512 image of an object with 9 cubes. Aliasing artifacts have not been reduced.

corrected using 10-bit LUT's for the Dunn film recorder, and recorded on 35mm Ektachrome slide film.

The images were rendered at various resolutions defined by tiling size on the 512X512 pixel matrix: tile sizes were 1, 2, 3, 4, 6 and 8. Thus the size 1 figures were rendered at the full 512X512 resolution of the display but the size 8 figures were rendered using tiles composed of 8X8 hardware pixels, yielding an effective resolution of only 64X64.

A Gourard shading model was used to assign a color to each tile. In one half of the slides the color of each tile was determined by the shading model at the pixel's center; the images exhibited aliasing artifacts. In the other half of the slides the color of each tile was determined using a simple antialiasing algorithm that performed 8X8 subpixel sampling with equal weighting of subpixels.

One half of the configurations had 9 blocks, the other half had 10 blocks. Six separate spatial rotations of the base figures were used. In all there were 6 different slides for each combination of aliasing, number of blocks and tile size.

Examples of the slides are shown in Figure 3(a–d). The digital half-toning process employed in printing may introduce artifacts into the images that are not present in the 35mm slides used for the experiment.

At the viewing distance used (182 cm), the base pixel size is approximately 0.032 degrees, so that one tile subtends about 0.032 degrees in the 1-pixel condition and about 0.25 degrees in the 8-pixel condition. The 1-pixel condition is near the limit of acuity under normal illumination (about 30 cycles/degree). A circle circumscribed around each object would have a diameter of approximately 450 pixels (about 90% of the display width).

In the first part of the experiment subjects were shown two slides simultaneously: a 1-pixel antialiased slide and the corresponding slide from one of the other cells. Subjects were told that the 1-pixel antialiased slide was of quality "100" and that they should assign a rating to the other slide indicating how good a representation it was in comparison to the standard – if it was 90% as good, they were to assign the number 90, if it was half as good, the number 50, etc.

In the second part of the experiment, slides were shown one at a time. Subjects were asked to indicate as rapidly as possible whether there were 9 blocks or 10 in the object.

All viewing conditions were the same in both parts of the experiment except for the use of two slides side-by-side in the first part and only a single slide in the second.

Appendix B – Mathematical Analysis

An analysis of variance for the subjective ratings revealed that both tile size ($F(5,45) = 62.27$, $p < .001$) and antialiasing ($F(1,9) = 41.82$, $p < .001$) had significant effects on the ratings. Fairly obviously, the ratings of quality were higher for images with small tile size and the antialiased figures were consistently judged to be better than those that had not been antialiased.

In addition, there was an interaction between the

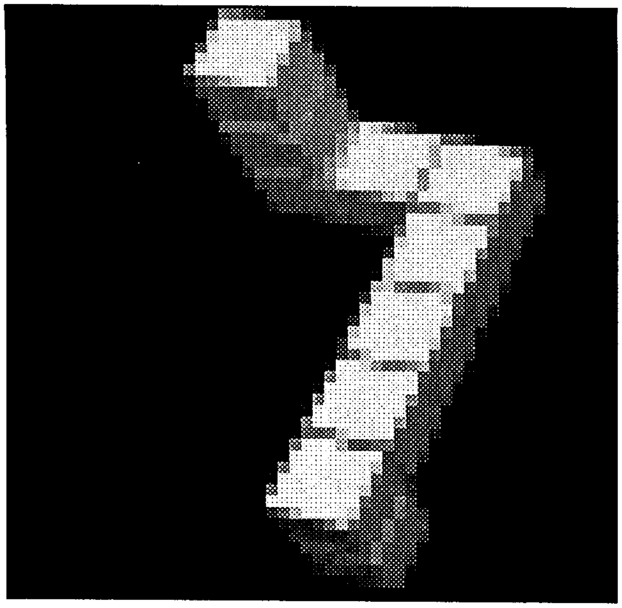

Figure 3(c). A 64X64 antialiased image of an object with 10 cubes. Aliasing artifacts have been reduced using 8X8 subpixel sampling.

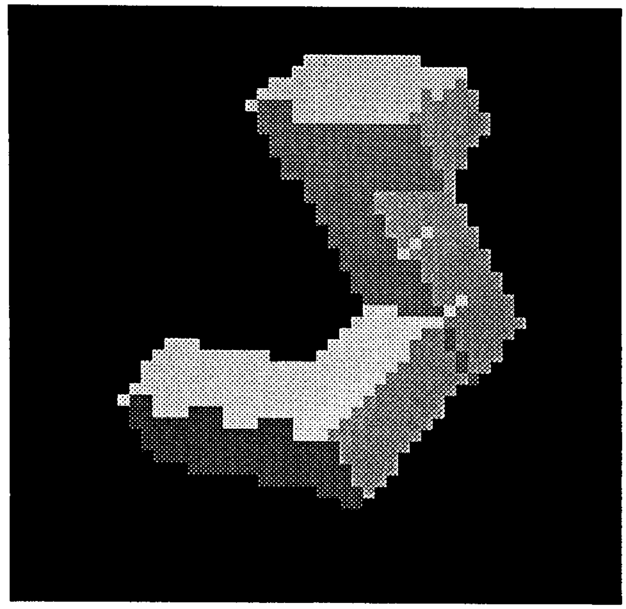

Figure 3(d). A 64X64 image having aliasing artifacts of an object with 9 cubes. Aliasing artifacts have not been reduced.

two variables ($F(5,45) = 7.36$, $p < .001$). The superiority of the antialiased figures was largest for tile sizes 2 and 3, and smallest for tile size 8. At the greatest degradation (tile size 8) both types of image were so poor that antialiasing did not result in any improvement in judged quality. The benefit of antialiasing increases as tile size decreases, down to tile size 2. At tile size 1, the benefit of antialiasing again reduces, as the limits of visual acuity are approached.

In the second part of the study, antialiasing was superior to aliasing only for tile sizes 4, 6 and 8. At higher resolution (tiles sizes 1, 2 and 3) there was no difference between the two conditions. The linear interaction of tile size with antialiasing was statistically significant ($F(1,8) = 5.38$, $p < .05$). Thus, although subjects thought the antialiased versions were perceptually better than the aliased ones, the addition of antialiasing did not make it any easier for them to carry out a simple discrimination task.

An alternative way of looking at these data is to combine the two parts of the study by plotting the median discrimination times against the quality ratings. Overall, faster response times are found with higher quality ratings ($r = -.62$, $p < .05$). This, however, is due entirely to the figures that were not antialiased. For those figures alone there is a high correlation ($r = -.96$, $p < .05$), whereas judged quality has little to do with response time for antialiased figures ($r = .04$, n.s.). For figures up to about 75% of the quality of a 512X512 antialiased figure, antialiasing makes the discrimination easier, given constant perceptual quality. Beyond this limit antialiasing does not improve performance on this particular task.

The specific discrimination task we chose was a very simple one. In essence subjects were required only to determine if the displayed object was "large" or "small" with respect to its number of cubes. Our data indicate that antialiasing does not improve performance for figures of even moderate quality (tile size 2, or 256X256 resolution). Of course, if we had elected to use a more complex task, we might well have found that the improvement with antialiasing extended to even higher resolutions.

In general, then, antialiasing is a good thing if it is important for the figures to look good, at least to the limits of visual acuity. However, if the primary concern is only with whether or not a person can operate on the representations, one need not be so fussy about antialiasing.