THE PERCEPTUAL BASIS OF ALIASING AND ANTI-ALIASING

MARC GREEN
COMPUTER STUDIES PROGRAMME
TRENT UNIVERSITY
PETERBOROUGH, ONTARIO K9J 7B8

1. ABSTRACT

Temporal sampling artifacts may cause jitter in moving video images. Most often, these artifacts are attributed to aliasing in the spatiotemporal spectrum of the image. However, the spatiotemporal spectrum is only a mathematical representation, so the practical value of this approach depends on a number of assumptions about human vision. Most importantly, there must be a linear summation of the individual components. This assumption was tested with both quantitative and informal experiments. Initial results showed that the spatiotemporal model was accurate for simple sine-wave gratings and slightly overestimated aliasing for more complex gratings. However, it was possible to create compound gratings where the model grossly overestimated aliasing. Results showed that, in general, the assumption of linear summation is unwarranted and that jitter in complex images cannot always be predicted from their constituent components. One practical implication of this result is that image quality testing should include natural images.

2. INTRODUCTION

Rosen (1988) characterizes 20th century science as the "Age of Syntax." He means that phenomena are understood by decomposition into context-free primitive elements plus a set of re-write rules. Image processing, of course, operates in the same fashion: most coding schemes and compression algorithms describe images in terms of primitive entities - usually the weights of a collection of basis functions. The rewrite rules are generally simple linear summations. The major advantage of linearity is that it permits the maximum degree of context-independence.

If image processing were merely a matter of engineering, this approach would be adequate. Formal, objective indices of coding fidelity would suffice to predict image quality. However, mathematics, like all formal systems, defines correctness only in terms of internal consistency. Any error measure must, by definition, be "true." The problem with relying solely on formal systems is that they have no inherent validity in the real world. The term "aliasing", for example, comes from mathematics but is used as a synonym for certain types of perceptual distortions which arise from image sampling. Strictly speaking, this confuses physical variables in the image with perceptual variables in the head.

The tendency to confuse natural phenomena with the mathematics used to describe them is very seductive. It often leads to attempts to make the real world fit the formal system rather than the other way around. However, if the relationship between the physical (aliasing) and perceptual (image distortion) were simple and direct, this error would not have practical importance.

0-8194-0820-4/92/$4.00

For the syntactic approach to work efficiently, therefore, mapping from image primitives to human perception must be very simple. Otherwise, there is no direct relationship between the mathematical image description and image quality. This means that *every image coding scheme is an implicit theory of human perception*. In fact, image processing can be viewed as the inverse of psychophysics, which is the science of finding these image primitives with re-write rules through empirical study. In psychophysics, advances are said to occur when someone discovers a new image description which maps simply to perception (Green, *et al.*, 1978).

Discussion of these meta-issues may seem like a lot of philosophical mumbo-jumbo, but it provides important perspective on a fundamental question in electronic imaging: what kind of images should be used to judge quality?

There are two polar extremes in philosophies of image quality measurement. The first approach, usually taken by more applied researchers, suggests that natural images should be used. The logic is that image quality should be tested with the kind of images which will actually be viewed. The tacit assumption here is that the syntactic approach will fail. Among the possible explanations are that 1): we don't have a good image grammar - a good set of primitives with a simple set of rewrite rules. or 2) the whole is more than the sum of the parts - there are emergent properties not present in simple images. For example, co-linear points may form a line, but the line has properties, such as orientation, which go beyond any combination of individual points. Similarly, perception of faces seems to have properties not explained by its individual features.

In contrast, there is the syntactic, scientific approach (e. g., Klein and Carney, 1991). Researchers use very simple image primitives with the hope that they can be easily combined to predict the appearance of complex images. There are several advantages to this method. Since natural images are highly variable, a coding and compression method which works well on one image may not generalize to others. In the syntactic approach, the quality of any complex image can presumably be predicted from the quality of its primitive components. The syntactic approach is therefore more general. Moreover, specific primitives can be used to test for qualitatively different aliasing artifacts.

Many current syntactic approaches use spatiotemporal frequency components as the primitives and linear summation as the rewrite rules. The rationale for the use of spatiotemporal primitives lies both in their mathematical simplicity and in evidence suggesting that they map simply to human vision. Any reader of the journal *Vision Research* is aware of the research by "gratingologists" who have provided numerous demonstrations that the spatiotemporal spectrum produces highly accurate predictions.

With few exceptions, however, most of the psychophysical evidence favoring the frequency primitives comes from studies using single, low contrast sinusoidal components viewed on a uniform field. To generalize these studies to natural images, there is an important assumption of linear re-write rules. Work in video image quality has generally relied on the same assumption because the application of linear transforms works best when the system is linear, isotropic, homogeneous, and context-free (the individual components must be processed locally by independent filters.) Since none of these assumptions is generally true (see DeValois and DeValois, 1988), there is reason to doubt the usefulness of the frequency approach for predicting image quality.

In the spatial domain, there is already evidence that one form of image distortion, apparent aliasing, cannot be predicted from the components of a scene. Nyman and Laurinen (1982, 1985) examined the sampling rates needed for recognition of both edges and single sinusoidal components. The linear filtering model could not predict the results because edges proved far more resistant to aliasing would be predicted from their individual frequency components. The authors concluded that the visual system may contain nonlinear mechanisms for processing particular image features, such as edges.

Below, I describe some experiments to test the syntactic approach to temporal aliasing. The general question is whether apparent temporal aliasing in complex images can be predicted by aliasing in constituent components. A companion paper (Green, 1992) examines whether apparent aliasing in monocular channels can predict apparent aliasing in binocular vision.

3. EXPERIMENTS

3.1. Sinusoidal sampling of a single moving component

Figure 1 shows how moving objects are typically (e. g., Watson *et al.*, 1986) represented in a spatiotemporal space. The top panel shows the representation for a single, moving sinusoidal component. If the component were stroboscopically sampled, the resulting spectrum would be a combination of the original spectrum plus a number of "replicas," whose spacing was determined by the sampling rate (Figure 1B). In theory, the only difference between smooth and aliased motion is the presence of the replica components.

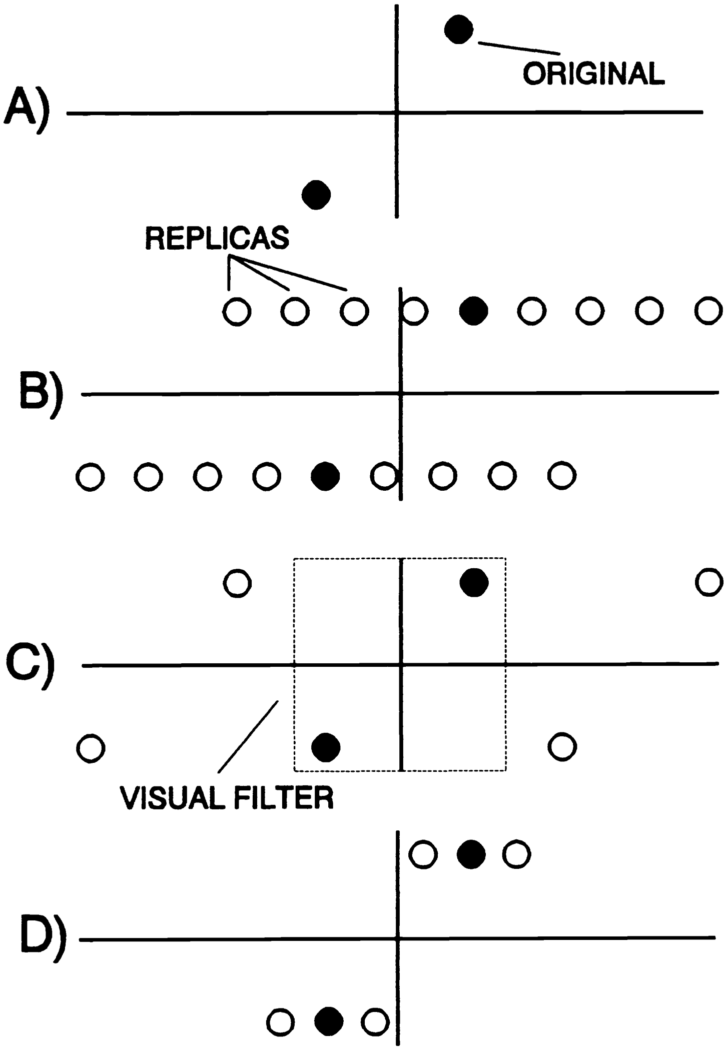

Figure 1

This is merely mathematics. To create a theory of apparent aliasing (as opposed to mathematical aliasing) it is necessary to combine the spatiotemporal representation with a model of human vision. The simplest model is to the view visual system as a single, linear, bandlimited filter. If the replicas fell outside the range of the visual filter, then smooth and sampled motion would be indistinguishable. One way to achieve this effect is to increase sampling rate and push the replicas outside that range of the visual filter (Figure 1C). The staircase motion more typical of video displays can be similarly analyzed (Morgan, 1980; Watson *et al.*, 1986)

These types of sampling are complex in the sense that they produce numerous replicas. The simplest possible sampled motion is the sinusoidal amplitude modulation (AM) of a single, moving sine-wave grating. As shown in Figure 1D, the resulting spectrum consists of only the carrier with only two sidebands. Magnitude of the sidebands is proportional to the depth of modulation. This is simply a generalization of the analysis performed by Sekuler and Levinson (1978) on various types of flicker.

The spatiotemporal model says that motion will appear smooth if the sidebands are undetectable. This leads to the simple prediction that the amount of AM modulation need to perceive the discontinuous motion of the carrier is equal to the amount of modulation needed for detecting the AM component alone. In other words, contrast threshold for detecting the AM component should be the same whether it is presented alone or superimposed on the carrier. This, of course, uses the assumption that the individual components are linear, context-free primitives.

3.1.2 Images

Observers viewed rightward moving sine-wave gratings presented on a Tektronix 603 CRT with frame rate of 200Hz and a mean luminance of 5 cd/m^2. The gratings and all modulations were produced by analog waveform generators. Audio attenuators controlled wave amplitudes.

There were 3 types of images. The first was a sine-wave "carrier" grating with a spatial frequency of 1.1 c/deg, contrast of 10% and temporal frequency of 3 Hz. The second was another sinusoidal component with the same spatial frequency and drift rate which sinusoidally counterphase flickered as it moved. The final image consisted of the counterphase component superimposed on the carrier. The sum was equivalent to sinusoidal amplitude modulation of the carrier at the flicker rate of the counterphase component.

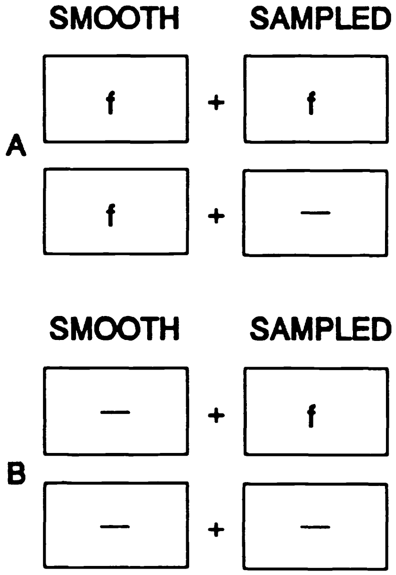

Figure 2

3.2.2 Procedures

The study consisted of two phases, each a series of two-alternative, forced-choice trials with one second intervals. Both experiments (Figure 2) required observers to detect the presence of a moving, counterphase flickering component. In the first phase, one interval contained the carrier while the other contained the carrier + counterphase component. The observers' task was to discriminate the discontinuous motion (carrier + counterphase) from smooth motion (carrier alone). This was accomplished by a staircase procedure which adjusted the amplitude of the counterphase component based on observer responses. I used a rule which produces a threshold equivalent to the 79.6% correct point on a psychometric function (Wetherill and Levitt, 1965). If the observer made three correct responses in a row, depth of modulation was decreased by 0.1 log unit. An error at any time resulted in a similar size increase in modulation.

In the second phase of the experiment, there was no carrier: one interval contained only the moving counterphase flickering signal while the other was blank. The observers' task was to discriminate the counterphase component from a blank field.

3.2.3 Results

Figure 3 shows the results for two observers. At high sampling rates, the model's predictions are upheld: threshold for detecting discontinuous motion and the counterphase component are similar. In other words, the only difference between smooth and jittery motion was the detection of the sidebands.

At low sampling rates, however, threshold for discontinuous motion is somewhat higher than for the counterphase component. This might be explained by masking. Low sampling rates produce replicas with temporal frequencies close to those of the carrier. Temporal mechanisms are broadly tuned (e. g., Green, 1981), so the detectors tuned to the replicas might be activated by the carrier. The carrier is then a background pedestal which raises amplitude threshold because of Weber's Law. When the sampling rate is high, carrier and replicas are presumably detected by mechanisms with nonoverlapping sensitivities. Analogous results have been reported by Bodis-Wollner and Hendley (1979).

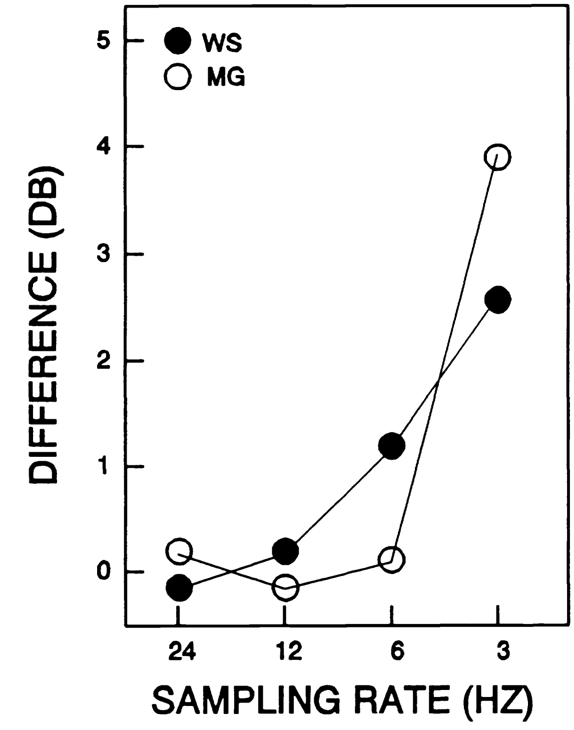

Figure 3

4.2.4 Discussion

The results reinforce the conclusion that the spatiotemporal filtering model is valid, at least at moderate to high sampling rates, when simple stimuli are employed. Previous studies (Watson *et al.*, 1986; Burr and Ross, 1987) reached the same conclusion using less direct tests than the one presented here. The next experiment tested the model's prediction when images were more complex and contained multiple sinusoidal components.

3.2 Sampling in complex gratings

Observers viewed gratings constructed from 2 pieces (Figure 4): 1) the first 3 harmonics of a square-wave grating with a 1.5 c/deg fundamental frequency, a fundamental contrast of 24% and a leftward drift rate of 1 Hz, and 2) the "edge components, the next 5 higher harmonics (7,9,11,13,15) of the same square-wave.

In one set of trials, the observers discriminated an 8 component grating from a second grating constructed from the first 3 harmonics, which drifted smoothly, plus the 5 higher edge harmonics, which counterphase flickered at 15 Hz. That is, the only difference between the two gratings was sampling of the 5 higher harmonics. In the second test, observers discriminated a steady version of the 5 harmonic grating from a counterphase flickering one. This is simply a more complex version of the previous experiment.

If it is possible to generalize from simple case in the first experiment to more complex gratings, then the prediction of the spatiotemporal model is that accuracy in the two tasks should be identical. In both cases, the only difference between the test gratings was the presence of flicker in the 5 highest harmonics.

The observers were again tested in a two-alternative, forced-choice procedure. Each trial block consisted of 200 trials with the two tasks randomly interleaved, and the flicker set to a fixed modulation depth of either 20, 30 or 40 percent.

Results: Figure 5 shows percent correct for the two tasks at modulation depths of 20, 30 and 40%. Aliasing in both conditions increases with depth of amplitude modulation. More importantly, there is a modest suppression of jitter detection when edge components were superimposed on steady low spatial frequency component.

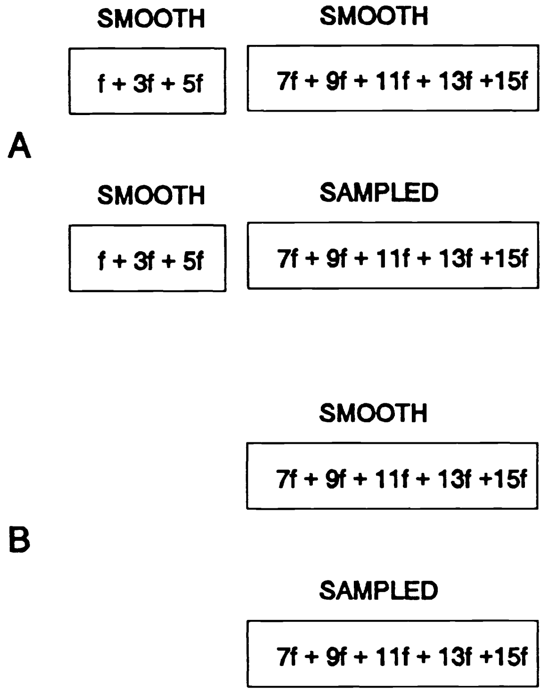

Figure 4

The small loss of perceived aliasing due to the steady low frequency components might again be explained by masking effects due to Weber's Law. The 5f component in the steady grating and the 7f, 9f, and possibly 11f component of the flickering wave fall within the bandwidth of a single detector. The steady 5f component may have provided a pedestal which elevated threshold for the flickering components.

The first two experiments suggest that the spatiotemporal filtering model is a reasonably good predictor of perceived aliasing. There were only modest departures, which might be explained by masking due to Weber's Law. Previous studies (e. g., Watson, *et al.*, 1986) also found only modest errors in the model's estimates.

In summary, these quantitative experiments support the "scientific" approach to image quality evaluation. It appears that apparent aliasing of more complex images can be reasonably predicted from their individual components.

4. OBSERVATIONS WITH COMPLEX GRATINGS

In spite of the seemingly small departure from model predictions, it proved easy to show that aliasing in individual components cannot always predict aliasing in more complex images. Below, I describe some displays which demonstrate that the spatiotemporal model can greatly overestimate aliasing under some conditions. Severe aliasing in high spatial frequency components can be eliminated by the presence of a low frequency component.

In an informal experiment, observers viewed three gratings in staircase motion. The first was an f (2.4 c/deg at 10% contrast) grating which stepped leftward in 60° phase shifts updated every 2 sweeps. When viewed alone, observers perceived only smooth motion in the display. The second grating had frequency 3f (7.2 c/deg) and moved leftward at the same velocity and the same sampling rate. This meant that the 3f component was sampled in 180° steps, i. e., it flickered in counterphase. In the spatiotemporal frequency domain, a flickering grating is simply a highly aliased motion. It is represented as the sum of two sets of components moving in opposite directions at the same velocities (Sekuler and Levinson, 1978). When viewed alone, of course, there was the maximal amount of apparent aliasing - pure flicker.

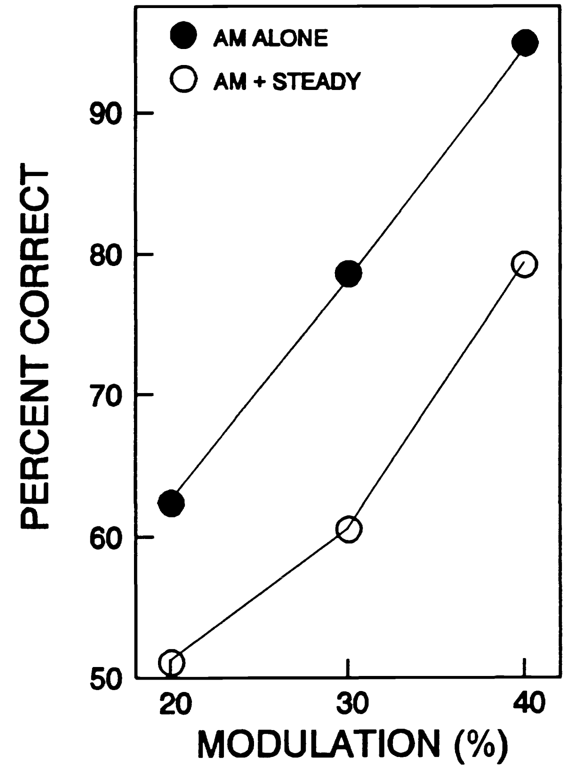

Figure 5

The remarkable observation was, however, that when the two components were superimposed, observers perceived *no apparent aliasing*. Instead of flickering, the 3f component appeared to move smoothly with f component. It was as if all of the rightward moving components, as well as the leftward moving components of different velocities, had become invisible.

To ensure that the effect was not due to involuntary tracking of the f component, observers subsequently viewed a display containing three horizontal bands of gratings with the middle band stepping leftward and the top and bottom moving rightward. The same suppression of aliasing could be seen simultaneously in bands moving in opposite directions. The phenomenon could not be explained by involuntary tracking.

Next, I replicated the basic observation by varying the velocity of the gratings while keeping the sampling rate constant. The f and 3f still moved at the same velocity, but step size increased so that the 3f component was sampled in larger jumps. This still created jitter but because the steps were greater than 180°, there was a clear sense of *rightward* motion. When superimposed on the leftward moving f component, the 3f grating again appeared to move smoothly *leftward*. The 3f component appeared glued to the f component even though neither it nor any of its replicas moved in both the same direction and velocity of the f components. The effect occurred even when the 3f component moved in 252° steps, producing a smooth rightward motion. The suppression proved robust over a wide range of variations in spatial frequency of the f component, phase and velocity. In all cases, any aliasing of the 3f component disappeared when it was superimposed on the f component.

This phenomenon is important for two reasons. First, it shows that the aliasing of complex images cannot in general be predicted by the individual components. Just as Nyman and Laurinen (1982, 1985) found in the spatial domain, linear summation the individual components grossly overestimates the amount of perceived aliasing in sampled images.

Second, the suppression occurs even when the components differ widely in spatial frequency. The 3:1 ratio of components suggests that the components would be detected by different spatial frequency channels. I also found similar results with gratings separated by wider frequency intervals. The results cannot be explained by Weber's Law masking or any similar phenomenon.

The observation further suggests that perceived aliasing does not arise directly from the activity of individual spatiotemporal channels. It seems more likely that perceived aliasing is the result of a mechanism which performs a nonlinear integration of the outputs of individual channels. Previous authors have suggested a number of candidate mechanisms.

One possibility is the opponent-process motion detectors proposed by Stromeyer, *et al.* (1984). They found that a grating moving in one direction masked a second grating moving in either the same or opposite direction. Based on this and other evidence, they concluded that motion detection is mediated by a mechanism which responds to *differences* between motions in opposite directions. However, all of their results were obtained with moving components of the same spatial frequency. Moreover, their effects only occurred at spatial frequencies below 6 c/deg and at very high velocities. The suppression of jitter can be obtained at both high spatial frequencies and low velocities.

Similarly the results cannot be explained by the same mechanism which underlies coherence of tilted gratings moving in different directions (Adelson and Movshon, 1982). Unlike the suppression effect, coherence disappears with differences in spatial frequency of the gratings.

A promising alternative is the MIRAGE mechanism (Watt, 1988) in which the channels, while contributing to perception, are never directly coded for frequency or bandwidth but combine nonlinearly. The significant aspect of such a model is that we have no direct access to the output of individual channels, although we may be fooled into believing this because of single sine-wave experiments. However, no currently proposed mechanism readily explains the jitter suppression.

5. CONCLUDING REMARKS

I started this article with meta-issues concerning the kind of images which should be used to evaluate image quality. The essence of the scientific approach (e. g., Klein and Carney, 1991) is that natural phenomena should be studied by a divide-and-conquer approach. As Simon (1981) notes, division into "nearly decomposable", i. e., context-free, subunits is the best way to understand complex systems. When applied to video image quality, this doctrine means that distortions in a complex image should be predicted by summing distortions in context-free primitives.

The linearity assumption common to most image representations is important because it permits the context-independence necessary for the syntactic approach to work efficiently. In the domain of temporal aliasing, the assumption means that the jitter produced by sampling individual components predicts the jitter seen in complex images. Studies, however, show that the assumption of context-independence is only partly valid. In some cases, it makes fairly accurate predictions, but in other is highly inaccurate.

It appears that, just as Nyman and Laurinen (1982, 1985) found in the spatial domain, the linear frequency models overestimate the amount of aliasing seen in complex images. The model is therefore useful in setting the upper bound on the conditions which produce jitter. However, the overestimates of perceived jitter may cause underestimates in the amount of image compression possible.

The syntactic approach could still work, although less efficiently, if it were possible to characterize the mechanisms performing the nonlinear integration. At this point, however, there is little understanding of the visual mechanism underlying the jitter suppression in high spatial frequency components. Until this occurs, complex, natural images will be important for quality assessment.

6. ACKNOWLEDGEMENTS

This research was funded by grant #46297 from the National Science and Engineering Research Council of Canada.

7. REFERENCES

Adelson, E. and Movshon, J. (1982). Phenomenal coherence of moving visual patterns, *Nature*, **300**, 523-525.

Bodis-Wollner, I. and Hendley, C. On the separability of two mechanisms involved in the detection og grating patterns in humans. *Journal of Physiology*, **291**, 251-263.

DeValois, R. and DeValois, K. (1988). *Spatial Vision.* New York: Oxford Press.

Green, M. (1981). Psychophysical relationships among mechanisms sensitive to pattern, flicker and motion. *Vision Research*, **21**, 971-1084. *Vision Research*, 21, 971-984.

Green, M., Corwin, T., and Zemon, V. (1978). Checkerboards and color aftereffects. **196**, 208.

Green, M. (1981). Psychophysical relationships among mechanisms sensitive to pattern, flicker and motion.

Green, M. (1992). Temporal sampling requirements in stereoscopic displays. *Stereoscopic Displays and Applications III.* In press.

Klein, S. and Carney, T. (1991). "Perfect" displays and "perfect" image compression in space and time. *Human Vision, Visual Processing and Digital Display II*, 190-205.

Morgan , M. (1980). Spatiotemporal filtering and the interpolation effect in apparent motion. *Perception*, 161-174.

Nyman, G. and Laurinen, P. (1982). Reconstruction of spatial information in the human visual system. *Nature*, **297**, 324-325.

Nyman, G. and Laurinen, P. (1985). Visual undersampling in raster sampled images. *IEEE Transactions on Man, Systems and Cybernetics*, **SMC-15**, 655-659.

Rosen, R. (1987). On the scope of syntacs in mathematics and science: the machine metaphor. In (J. Casti and A. Karlqvist, eds.) *Real Brains and Artifical Minds*. New York: North Holland.

Sekuler, R. and Levinson E. (1978). Physiological basis of motion perception. In (H. L. Teuber, H. Leibowitz and R. Held, Eds.) *Handbook of Sensory Physiology, Vol. VII, Perception*, 67-96. Berlin: Springer Verlag.

Shannon, C. (1949). Communication in the presence of noise. *Proc. IEEE IRE*, **37**, 10-21.

Simon, H. (1981). *Sciences of the Artificial, 2nd Edition*. New York: Wiley.

Stromeyer, C., Kronauer, R., Madsen, J. and Klein, S. (1984). Opponent-movement mechanisms in human vision. *Journal of the Optical Society of America*, **1A**, 876-884.

Watson, A., Ahumada A. and Farrell, J. (1985). Window of visibility: a psychophysical theory of fidelity in time-sampled visual motion displays. *Journal of the Optical Society of America A*, **3**, 300-307.

Watt, R. (1988). *Visual Processing: Computational, Psychophysical and Cognitive Research*. New York: Erlbaum.

Wetherill, G and Levitt, H. (1965). Sequential estimation of points on a psychometric function. *British Journal of Mathematical and Statistical Psychology*, **18**, 1-9.