INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

Visual Temporal Masking at Video Scene Cuts

**M.A. Thesis
Carol English**

Carleton University

May 15, 1997.

National Library
of Canada

Acquisitions and
Bibliographic Services

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale
du Canada

Acquisitions et
services bibliographiques

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-26916-7

Short Abstract

Two experiments were performed to study visual temporal masking using naturalistic images. Masking was evaluated in three frames on either side of a scene cut. In Experiment 1, perceptual thresholds (75% accuracy) were measured for all six frames using a PEST algorithm. The strongest masking effects were observed in the first frames on either side of the cut. Masking strength was found to vary with image content. Frame one, Forward masking was found to hide more noise than Backward masking, supporting previous research on masking at scene cuts but contradicting classical masking literature. In Experiment 2, suprathreshold masking effects were evaluated using a subjective quality scale. The first frame, on either side of the cut could be degraded severely without affecting perceived image quality. Modest levels of image degradation were tolerated in the second frame. Results varied by Image content.

Quality recommendations for coder design were specified from experimental results.

Abstract

Two experiments were performed to study visual temporal masking with naturalistic images. The goal of the research was to measure the level of image compression/quantization that could be hidden by visual masking at a scene cut. Masking was studied in three frames following a scene cut (Forward masking), and in three frames prior to a scene cut (Backward masking).

In Experiment 1, three Forward, and three Backward thresholds were estimated for each participant. A threshold was considered to be the level of quantization at which the participant could reliably distinguish a degraded image from a non-degraded image 75% of the time. The strongest masking effects were observed in the first frame after a scene cut, and in the last frame before a scene cut. Some masking effects were still noticeable in the second frame before and after the scene cut. Strength of masking varied as a function of image sequence content, and practice. In the first frame, Forward masking was found to hide more noise than Backward masking. This result supported previous research on visual temporal masking at scene cuts but contradicted classical masking literature.

In Experiment 2, suprathreshold masking effects were evaluated. It was found that the image in the first frame, before or after a scene cut, could be degraded severely without affecting perceived image quality. Modest levels of image degradation were tolerated in the second frame. As in Experiment 1, there were reliable differences between the target images. Overall, subjective ratings were higher in the Forward masking condition than in the Backward condition, indicating that masking was stronger in the Forward than in the Backward direction.

Visual Temporal Masking at Video Scene Cuts

Introduction

The purpose of the present research was to study the time course of visual temporal masking using naturalistic images, and to consider ways to exploit this knowledge in the area of image coding and compression, specifically as it impacts on buffer demand at scene cuts.

The main thesis of the research can be summarized as follows: Because successive video frames are generally highly redundant (i.e., because successive frames undergo minimal change), video coding and compression schemes are able to exploit this temporal redundancy by coding only frame to frame differences, a very economical strategy for storage or transmission. Unfortunately, this strategy fails at scene cuts within video sequences, where the correlation between successive frames drops close to zero. This "unexpectedness" presents a data overload problem for the coding and decoding buffers necessitated by compression algorithms. In order to accommodate this sudden temporal decorrelation at scene changes, buffer capacity must be increased (an expensive option) or image quality must be drastically reduced.

The properties of the human visual system, however, suggest another solution. A convenient artifact of human visual processing is its momentary insensitivity to any temporal image decorrelation, that is, any sudden image change. This visual insensitivity, typically referred to as visual masking, is coincident in time with the decorrelation in image content (which causes the buffering "overload"). Because of this, visual temporal masking can be exploited by data coders for managing buffer overloading. What this means is that images near scene cuts may, potentially, be degraded (through severe image compression) with little or no effect on the perceived

image quality. Degradation may occur either before or after scene cuts, presenting an economical alternative to increasing buffer size.

This study establishes threshold and suprathreshold visibility parameters for image degradation at scene cuts. Using JPEG/MPEG standards for image quantization and degradation allows experimental results to be applied directly to the design of coders as they perform around scene cuts, and can be generalized to any image coding system.

Two experiments were performed: The first experiment established threshold parameters for the detection of image degradation in the first three frames following a scene cut (Forward masking), and in the last three frames prior to a scene cut (Backward masking). The second experiment established suprathreshold parameters for acceptable image degradation in the first three frames following a scene cut (Forward masking), and the last three frames prior to a scene cut (Backward masking). All experiments used naturalistic images as masks and targets.

The first section of this thesis will cover some basic theoretical issues in visual masking, provide an overview of image coding and compression, and review the psychophysical masking literature related to video coding and compression. The next section will describe the two experiments that were carried out, including experimental results. The final discussion will discuss these results and their potential applications, as well as suggesting directions for future research.

Visual Masking

Visual masking is commonly understood as a condition in which exposure to one image, the **mask**, blocks or interferes with the visual processing of a second, **target** image. Howard and Rogers (1995, p. 100), define masking as "A briefly presented suprathreshold stimulus [that] tends to elevate the threshold of a briefly presented test stimulus presented in the same location or in a neighbouring location at the same time

or in close temporal contiguity." Kohlers (1983, p. 136) defines masking more simply as an "alteration of perception due to sequential presentation of stimuli."

Masking can be temporal or spatial: Spatial masking refers to masking within an individual image, where a part of that image affects the perception of another part of the same image. Temporal masking refers to masking in which images, presented sequentially in time, interfere with one another.

There are two main categories within temporal masking, **Forward masking** and **Backward masking** (Turvey, 1973; Breitmeyer & Ganz, 1976). **Forward masking** occurs when the mask precedes the target in time, affecting perception of the target. **Backward masking** occurs when the mask follows the target in time. Kohlers (1983, p. 136) states: "The principal observation is that stimulus event M occurring after the event T nevertheless affects the perception of T."

In surveying the literature on temporal masking there are two common terms, stimulus onset asynchrony (SOA) and interstimulus interval (ISI), that are used to define the time course of masking. Breitmeyer and Ganz (1976) define stimulus onset asynchrony (SOA) as "the temporal interval separating the onsets of the target and mask", and state that it is conventional to use positive SOA values to indicate Backward masking and negative SOA values to indicate Forward masking. In other words, when the mask follows the target, the SOA takes a positive value; when the mask precedes the target, SOA is negative. The interstimulus interval (ISI) is defined as the delay between the offset of the first stimulus and the onset of the second¹ (See Figure 1).

One of the results to emerge from the traditional masking studies is that two distinct processes appear to be at work, **integration** and **interruption** masking (Turvey, 1973; Breitmeyer & Ganz, 1976; Breitmeyer, 1980; Kohlers, 1984). **Integration** masking

¹ Note that providing the ISI and the SOA in forward masking will give you the duration of the mask ($SOA - ISI = \text{duration M}$), whereas in backward masking it will give you the duration of the target ($SOA - ISI = \text{duration T}$). See Figure 1 for clarification.

occurs when the mask and target image combine and a single undifferentiated image is seen (Dodwell, 1971; Turvey, 1973; Breitmeyer & Ganz, 1976; Loftus & Hanna, 1989).

Interruption masking occurs when the "mask" interrupts or interferes with the processing of the "target", such that the target is not seen, or is seen indistinctly (Spencer & Shuntich, 1970, Turvey, 1973). **Integration** masking is believed to occur under conditions of both Forward and Backward masking. **Interruption** masking is believed to occur primarily under Backward masking conditions, that is, when the processing of the target is interrupted by the mask (Humphreys & Bruce, 1989), specifically for non-overlapping masks and targets (Breitmeyer & Ganz, 1976).

For short target:mask intervals ($SOA < \sim 48$ ms), **mask** and **target** are interchangeable and integration masking takes place. In other words, for very short SOA's, there is no functional distinction between Forward and Backward masking. For longer target:mask intervals (~ 48 ms $< SOA < 100$ ms) with Backward masks, interruption masking may occur (Turvey, 1973), (See Figure 2). It has generally been assumed that interruption masking takes place only under conditions of Backward masking, since the, by definition, it is not possible to interrupt a stimulus prior to its occurrence. For intervals of over 100 ms, interruption masking becomes negligible as well (ibid.). Some research has shown that this period may extended to as much as 200 ms, with latencies that depend both on semantic, and structural properties of the target (Breitmeyer & Ganz, 1976). Both these observations support the view that Backward masking is stronger than Forward masking. Humphreys and Bruce (1989, p. 116) interpret Turvey's results to mean that "Forward masking...[is] weaker than Backward masking at longer target durations" (see Figure 3).

According to Breitmeyer and Ganz (1976, p. 7), in integration masking "sensory representations of target and mask stimuli can combine...to form a representation in which the mask camouflages or obscures the target"; however, this is not always the case. The detailed manner in which the target or mask combine depends on the relative

energy of the stimuli: The more proportionately intense the target, the more resistant it will be to masking (Turvey, 1973). Bloch's law states that for stimuli presented for less than 100 ms, the stimulus energy equals the stimulus intensity multiplied by the stimulus duration (Humphreys & Bruce, 1989).²

Integration and interruption masking effects are thought to be artifacts of the visual system's normal functioning: It is likely that integration masking is an artifact of the limited temporal resolution of the visual system. This necessarily limited resolution can also be understood as a feature which allows the summation of visual activity over time, and may facilitate other visual processes such as the computation of movement (See Bridgeman, 1978, for a model of how this may take place; also see, Braddick, 1980). Interruption masking may be an artifact of the visual system's requirement for distinguishing separate images.

Breitmeyer and Ganz (1976) argue for a dual channel explanation of interruption masking; a "transient" channel in the visual system that actually suppresses a "sustained channel". It is this "transient-on-sustained" suppression that "interrupts" an ongoing process of visual information integration, "masking" any information that is still being processed. In normal visual processing, periods of visual information integration are naturally interspersed with saccadic eye movements. Timing of the saccades and the periods of integration varies considerably depending on intrusions into the visual field, but the process itself is normally involuntary, as well as being necessary to normal visual processing. In order to perceive the world, we need both continuity and interruption of flow. Thus, we require the ability to perceive correlated change (integration), as well as individual, separate events (interruption) (Humphreys & Bruce, 1989).

² For example, a stimulus with a duration of 2 ms and an intensity of 10 units is perceptually identical to a stimulus of 10 ms with an intensity of 2 units.

The distinctions between Backward and Forward masking, and between masks and targets, can be confusing simply because there is no such thing as a generic mask or target. A **mask** may mask a **target**, and a **target** mask a **mask**. What we are really talking about are two images (and sometimes more, see Breitmeyer, 1978) that may or may not integrate/interfere with each other. Traditionally, by definition, we have looked only at the processing of the **target** and remained unconcerned with the processing of the **mask**. But, in practice, an abrupt change from one stimulus image to another stimulus image may create a condition in which each image integrates or interferes, simultaneously, with the other. With naturalistic video sequences, the video frames defined as the mask and target are interchangeable. For the purposes of this study, the target is defined as the video segment containing impairments. The mask is defined as the video segment without impairment. Though for practical purposes this definition is suitable, from a theoretical perspective it is an oversimplification: The sequences before and after the scene cut are composed of a sequence of individual images, each of which is displaced from, and may have a masking effect upon its neighbours.

Video Coding and Compression

"Efficient image representations are important for systems with finite resources... No matter how much computer memory...we have, we can always perform better computations, transmit more information, or store higher quality images if we use efficient storage algorithms. If we fail to consider efficiency, then we waste resources that could improve performance." (Wandell, 1995, p249)

Digital transmission of video images requires enormous data storage capacity, and places huge demands on data transmission systems. The "cost" of sending a video frame compared with that of a page of ASCII text is 1000 to 1 or more. A single video

pixel can use as much as twenty-four data bits to code position, hue, and value information. On a VGA graphics display, the least complex of high-quality colour graphics monitors, this translates into 200 by 320 pixels by 8 bits, a total of 512,000 bits. On a higher resolution monitor this can easily increase by a factor of three -- to one and a half million bits (Wandell, 1995; Nelson, 1992).

Lossless Coding

Algorithms exist which allow for "*lossless*" graphics compression, sometimes known as redundancy coding or entropy coding. This is a type of compression in which no information is lost through the coding and decoding process. Though lossless algorithms can compress images to as little as twenty percent of their original size, through redundancy reduction (Nelson, 1992), lossless compression will, generally, achieve a data savings factor of 2 to 3 (Wandell, 1995). Though this is an obvious improvement, storage requirements are still prohibitive, making lossless coding appropriate only for cases, such as archiving of images, where it is necessary to maintain the original data without loss, or where transmission speed is not an issue.

Lossy Coding

There are alternative means of compression which, though not lossless, will still produce unnoticeable, or at least acceptable, image degradation when decoded. Lee and Dickinson (1994, p. 513) state "To achieve high compression, one must resort not only to redundancy reduction [lossless coding] but also to irrelevancy reduction, lossy coding that exploits characteristics of human visual perception." Algorithms for this alternative "*lossy*" coding have allowed digital compression of images to as little as five percent of their original size.

Lossy coding includes both threshold and suprathreshold image degradation. Threshold degradation is unnoticeable even to "expert" viewers, subjects who have

spent many hours learning to distinguish degraded images from non-degraded images. This compression is sometimes referred to as being "perceptually lossless" (Wandell, 1995). Suprathreshold degradation is degradation that is noticeable, but still acceptable to viewers. Establishing these psychophysical threshold parameters is important in that it allows us to determine data coding parameters. In other words, the trade-off between psychophysical needs and system limitations can then be optimized. The key issue here, then, is that of determining an appropriate ratio of image quality to image cost. To reiterate, measuring psychophysical masking effects around scene cuts may allow us to find the optimal compromise between buffer size limitations at scene-cuts and image quality.

Still Image Compression/ Spatial Compression (JPEG)

Joint Photographic Experts Group (JPEG) is a standards group that has produced specifications for both lossless and lossy encoding of still images. Lossless encoding uses a predictive/adaptive model (with an entropy coding method, such as Huffman coding or arithmetic coding), and may or may not follow part of the sequence for lossy coding, described below. However, lossless coding algorithms are beyond the scope of this paper and will not be dealt with here. (See Wallace, 1991; Nelson, 1992, for more information on lossless encoding and decoding.)

In lossy encoding, each digitized image is divided into 8 X 8 blocks (64 pixels), which are then encoded independently.³ This encoding takes place in three basic stages (see Figure 4): In the first stage, the 8 X 8 matrix, which consists of spatial values (i.e., values representing individual pixels in individual locations in image space), is transformed into an 8 X 8 matrix in frequency space. This is done through a

³ This is not strictly true. Nelson (1992) tells us that "since adjacent blocks in an image exhibit a high degree of correlation, coding the DC [discrete cosine] element as the *difference* from the previous DC element typically produces a very small number (p. 369)."

mathematical function called the Discrete Cosine Transformation (DCT) (Figure 5), a subclass of mathematical operations known as Fourier Transforms. This frequency matrix contains the relative energy of each discrete frequency value, listed sequentially from the lowest frequency, in the top left hand corner of the matrix, to the highest frequency, in the bottom right hand corner of the matrix. (see Figure 6). The coefficient at position (0,0) is known as the DC coefficient and represents the average energy for the entire block.

This first encoding step is still lossless, except for a small roundoff error. At this stage, the matrix can be transformed from frequency space back into image space with no essential loss of information (through the inverse discrete cosine transform). Each coefficient in frequency space represents the relative contribution of that frequency to the overall image. All representative frequency waves combine, in their respective weights, as a wave interference pattern that reproduces the original image.

The second stage of the algorithm is where the "lossiness" takes place, that is, where data is "lost". During this stage, the matrix coefficients are quantized, in preparation for compression. Nelson (1992) defines quantization as "the process of reducing an integer value by reducing the precision of the integer" (p364). This quantization or reduction takes place mathematically; each matrix co-efficient is divided by a quantum value. Quantum values can range from 1 to 255. These values establish step size. Step size, in turn, represents a measure of the degree of compression of the final image, as well as its quality.

Two things are important to note here: The first is that the matrix is arranged sequentially from low to high frequency. The second is that, in general, naturalistic scenes contain most of their energy in the low frequencies. Thus, a uniform quantum value will tend to code higher frequency information as zeros, unless there are high frequency matrix values with unusually high energies/amplitudes. At this point, we have arranged our data in the format which will statistically allow the most efficient

compression.⁴ Although we have discarded information, we have done so in a way that will have the least impact on image quality following decompression.⁵

The third stage of the coding process is the compression stage. There are three sub-sections within this third stage. The DC coefficient in any given block (except the first) is coded as a difference value with respect to the previous DC coefficient. Reading the data in a "zig-zag sequence" produces a data string ordered from the lowest to the highest frequency information. The sequences of zeros are then compressed out, using a "Run-Length Encoding" (RLE) algorithm. Since it is common for more than half the coefficients to be reduced to zero after quantization, and for most of them to be in the higher frequencies, this facilitates good compression (Nelson, *ibid.*). Finally, "Entropy" Coding, generally either Huffman coding or arithmetic coding, is used to further compress the frequency data (Nelson, *ibid.*).

Excellent overviews of still and moving image coding and compression are given by Bhaskaran & Konstantinides (1995); Wandell (1995); and Nelson (1992).

Temporal image compression / MPEG standards

Still image compression exploits **spatial** redundancy. Video compression exploits both spatial and **temporal** redundancy. Normally an image changes gradually between

⁴ Note that even in lossless coding it may be worthwhile to transform our data into frequency space (without subjecting it to a lossy quantization process), since even losslessly we will still tend to get longer strings of zero values, because many high frequency contributions will be reduced to zero. These can then be losslessly compressed and decompressed.

⁵ Nelson (1992) points out that there are two ways to determine the best range of quantizing coefficients: The first is to use a global error term measurement, established by comparing the decompressed image with the original image. The second is to use a psychophysical measure based on what the eye can distinguish. It cannot be assumed that the mathematical measure will produce that same result as the psychophysical one. Girod (1989, 1992) discusses this in detail, noting that the most effective coding will be that which is matched, by frequency, to the variable sensitivity of human eye. Still the use of a uniform quantizing coefficient, weighting in favour of the lower frequencies, is a simple approximation of the ideal.

frames, so the difference between frames is not very great. Thus, in addition to exploiting the redundancy within each individual image frame, it is possible to take advantage of the high degree of redundancy across successive frames.

According to Bhaskaran & Konstantinides (1995), using strictly a DCT-based coding scheme, it is possible to achieve a data rate of 14 Mbits/s for normal video transmission, which, they claim, is still too high for most practical uses. Digital television broadcasting can manage a bandwidth of 4 to 6 Mbits/s; CD-ROM's, 1.5 Mbits/s (a compression ratio of 110:1). Thus, temporal image compression is critical. LeGall (1991) claims that temporal compression can reduce storage size by a factor of three.

Moving Picture Experts Group (MPEG) have established generic standards for temporal compression (LeGall, 1991) in the same way that JPEG has established still picture compression standards. These standards are particularly suited to lossy compression, including compression down to 1 to 1.5 Mbits/s, making them ideal for CD-ROM storage (LeGall, 1991.).

In MPEG protocol, individual video image frames are processed and transmitted in groups (see Figure 7). A group commonly consists of 15 frames, but MPEG standards allow this length to vary, depending on needs and applications. Each group of pictures (GoP) consists of an Intra-Frame (I), one or more Predicted-Frames (P), and several Inter-Frames, or Bi-directionally predicted frames (B). I-Frame pictures are used as references for all the other pictures in the GoP. They are compressed losslessly, using only spatial compression algorithms. P-frames are predicted from the previous I-frames, or the previous P-frames. B-interpictures are predicted bidirectionally through interpolation, using the closest P or I-frame on either side, or an average of both, depending on which strategy produces the smallest prediction error (Aravind et al., 1993). To supplement prediction, difference information, that is, the difference

between the predicted and the true image, is also transmitted (LeGall, *ibid.*; Aravind et al., *ibid.*).

Buffering

The purpose of a buffer is to convert a variable bit rate at the encoder to a fixed bit rate in the transmission channel. It is necessary to protect the output buffer from overflowing at times of high demand (Wang, 1995). Most algorithms accomplish this by having feedback control from the buffer to the quantization stage of the coder. As the buffer becomes full, the severity of quantization increases. When this occurs, image quality fluctuates, sometimes severely.

The critical issue here is that the added data load at scene cuts can lead to buffer overflow. Thus, it is important to implement algorithms to manage this added load. Forward and Backward masking studies imply that image quality can be safely degraded around scene cuts without affecting perceived image quality. This suggests a natural way of coping with buffer overflow problems, based on human visual performance.

Wang (1995) suggested that temporal masking effects could be exploited to enhance and stabilize the performance of MPEG-2 coders, delaying or smoothing out sudden data demands on buffers. No psychophysical or subjective tests were performed in Wang's study.

Review of Related Literature

Early masking studies include Alpern (1953), on metacontrast masking; Barlow (1958), on temporal and spatial summation in human vision; Boynton (1961), on temporal factors in vision; and Averbach and Coriell (1961) on short-term memory in vision. Other masking studies include Townsend (1973); Bowen, Pola, and Hanna

(1974); von Graunau (1978); Coltheart (1980). Spencer and Shuntich (1970) looked at interruption effects in Backward masking; Scheerer (1973) studied integration and interruption in Backward masking; DiLollo (1977), DiLollo and Dixon (1988), and Dixon and DiLollo (1992) looked at integration effects in masking. General review articles on masking include Turvey (1973); Breitmeyer and Ganz (1976); Kohlers (1983). Suggested models of visual masking have been proposed by Allport (1968); Dodwell (1971); Bridgeman (1978); Loftus and Hanna (1989); Loftus and Ruthruff (1994); DiLollo and Dixon (1988). Humphreys and Bruce (1989) include a chapter on masking in *Visual Cognition*. Breitmeyer published *Visual Masking: An Integrative Approach* in 1984.

Though there is a large body of literature on masking in general, very little research has been done either on temporal masking using naturalistic video images, or on the use of temporal masking effects for video coding. Lee and Dickinson (1994) note that spatial visual masking has been recognized and utilized for video compression, but to date, almost no consideration has been given to the possibilities offered by temporal visual masking.

An early study by Seyler and Budrikis (1964) used a transient low-pass filter to reduce bandwidth in order to look at subjective response to reduced spatial detail in images after scene cuts. They found that, provided the image had returned to full bandwidth by the end of 780 ms, observers judged the images as sufficiently good. This is considerably greater than the measured 50 ms - 100 ms masking effect (Turvey, 1973) shown in more traditional studies, however it must be remembered that Seyler and Budrikis were using subjective judgment criteria, rather than "objective" psychophysical threshold measurements. They suggest that this finding could be used to "reduce extensive (and costly) buffer storage in the technical coding system for the redistribution of these transients in the flow of frame difference samples."

Seyler and Budrikis state that "the sensory perception process is not noticeably affected if the external filtering operation is matched to the sensory one" (p. 42). Their results imply is that there is considerable leeway between the perceptual noise threshold at a scene cut ("can't see") and the level of subjectively acceptable image degradation ("can see, but don't care") that may be exploited when coding. This is an important point, because most investigations of temporal masking have investigated threshold visibility, not suprathreshold effects. The present thesis investigates both threshold and suprathreshold temporal masking effects in naturalistic images.

Girod (1989) investigated the bit-rate savings allowed by spatial and temporal masking in video signals. He found that Forward temporal masking was significant only in the first 100 ms after a scene cut, but did not investigate Backward masking. He also investigated temporal masking due to image movement, finding that, due to the visual system's tracking ability, this was insignificant.

Girod (1992) derived a model of the system composed of a video screen, fovea, retina, and optic nerve. This model is highly non-linear, its parameters were fitted to a number of psychophysical results. Using his model, Girod predicted strong Forward masking effects but, again, did not address Backward masking effects.

Lee and Dickinson's (1994) study considered Backward and Forward temporal masking for the purposes of optimizing I-frame intervals. They were able to improve considerably on the method used by Seyler and Budrikis, using a discrete cosine transformation (DCT) on full colour images. As well, they were able use actual video scene cuts, rather than switching between a live video image and a still image.

Examining reduced sensitivities to a single frame, immediately preceding and following a scene cut, they found that this frame could be coded with much reduced information. As little as twenty percent of the usually required information was needed in the frame immediately preceding a scene cut (Backward masking), and only five

percent in the frame immediately following a scene cut (Forward masking). Lee and Dickinson based their conclusions on informal viewing of processed sequences.

Tam, Stelmach, Wang, Lauzon and Gray (1995) studied Forward masking after a scene cut, and established psychophysical visibility thresholds for impairments in the first, second, and third frames following a scene cut. As well, they varied impairments in the first and second frames together in order to establish a temporal error buildup threshold. Tam et al., found that masking effects occurred primarily in the first frame following a scene cut, and were almost negligible by the third frame. Further, they found that impairments in the first frame, although below threshold, lowered the threshold level in the second frame. The present study replicated Tam et al's findings for Forward masking effects in the first three frames after a scene cut. As well Backward masking effects were measured for the last three frames prior to a scene cut, in order to compare Backward and Forward masking thresholds.

Turvey (1973) showed that integration masking takes place during both Forward and Backward masking, whereas interruption masking only takes place in Backward masking. Humphreys and Bruce (1989) noted that Forward masking effects have been measured as weaker than Backward masking effects, at least for dichoptic pattern masking. This is presumably because Backward masking combines integration and interruption masking. However, this appears to contradict Lee and Dickinson's findings that Forward masking is stronger than Backward masking. Turvey found that noise masks, in both Forward and Backward masking, seemed to exhibit only integration masking. It is possible that the dominant effect in naturalistic images is one that most closely resembles noise masking (integration masking) rather than pattern masking (interruption masking).

Brietmeyer and Ganz (1976) argue that metacontrast masking (their term for interruption masking) only takes place when masks and targets do not overlap -- which they clearly do in video sequences. However they still believe that physiological

interchannel interference will still result in stronger Backward than Forward integration masking effects under dichoptic viewing conditions.

Experiments

Introduction

Temporal decorrelation occurring at scene cuts produces a sudden increase in the volume of data that needs to be transmitted, presenting a problem for the image coding algorithm. The temporal decorrelation coincidentally also produces a momentary lag in visual responsiveness, both Forward and Backward in time. Thus, it is possible to exploit visual temporal masking to avoid this buffer overload.

Previous research on masking using naturalistic images has shown that masking can hide some image degradation both before and after scene cuts (Lee & Dickinson, 1992). Tam et al. (1995) systematically measured psychophysical thresholds in their study of Forward masking using naturalistic images. However, prior to this study no one had measured thresholds for Backward masking. This is a curious omission, given that classical masking studies have indicated stronger Backward masking effects than Forward masking effects.

Two studies have established that there are **suprathreshold** masking effects (Seyler & Budrikis, 1964; Lee and Dickinson, 1992). Seyler and Budrikis, looked only at suprathreshold Forward masking; Lee and Dickinson looked only at suprathreshold effects in the first frame on either side of a scene cut. No one has carried out suprathreshold testing for Backward masking in order to determine the degree of subjectively acceptable image degradation over the time course of masking. The difference between threshold limits for image degradation ("can't see") and suprathreshold limits ("don't mind") are of considerable practical use for determining minimum buffer size, since, ultimately, visible image degradation is not a problem if it

is not actually bothersome to viewers. If the masking time course extends over more than just the first frame before, and the first frame after the scene, this allows for the possibility of progressive image buildup, further easing buffer strain.

The present study addressed these two questions. In Experiment 1 thresholds for Forward and Backward masking using naturalistic images were measured. Maximum threshold quantizations, corresponding to the severity of image degradation, were established for the three frames following the scene cut (Forward masking) and the three frames preceding the scene cut (Backward masking). In Experiment 2, observers responses to suprathreshold levels of impairment (where compression artifacts were clearly visible) were recorded. This was done for the three frames before and after the scene cut.

Experiment 1

Threshold measurements (Forward and Backward masking)

Experiment 1 was conducted to estimate the threshold sensitivity of the human visual system to artifacts in the three frames just following a scene cut (Forward masking), and just preceding a scene cut (Backward masking). We used the same methodology as Tam et al. (1995), (who studied Forward masking using naturalistic video images), extending it to Backward masking. This was accomplished by taking the same video sequences that we used in the Forward condition and playing them backwards. This meant that each Forward masking presentation had a Backward masking counterpart that was identical except that it was shown in reverse. For example, an impairment seen in the first frame following the scene cut (for the Forward condition) would appear in the first frame prior to the scene cut (for the Backward condition). By matching Forward and Backward conditions in this way we were able to

analyze and compare the strength of Forward and Backward threshold masking effects, using identical image content .

Threshold estimates were obtained for the first, second, and third frames following a scene cut, and the first, second and third frames prior to a scene cut. Forward and Backward visual masking thresholds were defined as the level of objective image impairment at which subjects were able to detect coding artifacts accurately 75% of the time.

The results of the present study have direct applicability to coder design.

Method

Video Sequences

Individual test sequences were created in the CCIR-601 (4:2:2, 720 X 480) format. Each sequence consisted of two, 1-second (30 frame) naturalistic image video segments which were shown consecutively. The sequence "Flower" functioned as the masking segment throughout the study. It depicted a panned view of a Flower garden. The target was varied randomly between three alternative target images: A sports action scene (Football), an indoor studio scene (Mobile), and an outdoor crowd scene (Women), were each used as target images, and each was combined with the "Flower" masking image. For the Forward threshold masking condition, the three two-second test sequences were: "Flower-Women", "Flower-Mobile", and "Flower-Football". For the Backward threshold masking condition the three two-second test sequences were reversed, thus appearing in the opposite order:: "Women-Flower", "Mobile-Flower", and Football-Flower".

Image Processing

Impaired video frames were produced by processing target images at a range of quantization levels ranging from $Q = 5$ to $Q = 255$, in steps of 5. A quantization coefficient of 5 created a high quality image, a quantization coefficient of $Q = 255$ created an image with the greatest possible impairment; one in which most of the discrete-cosine transformed (DCT), 8 X 8 pixel, blocks in the frame were displayed at the mean level of luminance and mean colour for the block. Initially, all unimpaired video frames were processed at $Q = 5$, for uniformly high image quality. For the Forward condition, Q was increased in either the first, second, or third frame following the scene cut. For the Backward condition, impairments were introduced in either the first, second, or third frame preceding the scene cut. (See Figure 8). Video sequences were processed using an MPEG-2 codec. In order to enable individual frame quantization, the Group-of-Pictures length was set to one ($GoP = 1$).

Participant /Viewers

The eight participants had either normal, or corrected to normal, visual acuity, and normal colour vision. Two participants were paid for their participation. Six participants, including the author, were student volunteers. Viewers ranged in age from twenty to fifty-eight.

Displays

A broadcast quality, 19-inch (Sony BVM-1910) colour monitor was used to display the test sequences. The testing room was adjusted to an ambient illumination of 10 lux, using a dimmable fluorescent system refreshed at 1000 Hz. The wall behind the video monitor was draped with medium gray velvet-textured curtains, and had a reflected luminance measured at 5 cd/m^2 . Test sequences were stored and displayed using a

DVSR-100, RAM-based video storage and display system. Viewing distance was equal to three picture heights.

Design and Procedure

Participants were tested one at a time, and initiated each trial with a button press. Each trial was made up of two sequences and each sequence was made up of two segments, a "target" segment and a "mask" segment. For the Forward condition the "mask" segment appeared first in the sequence, followed by the "target" segment. For the Backward condition the "target" appeared first, in the sequence, followed by the "mask". Backward condition trials were produced by actually running the Forward condition trial sequences backwards, thus reversing "mask" and "target". As well, this reversal caused the degraded frame, which had appeared in the first, second or third frame following the scene cut, to now appear in the first, second or third frame preceding the scene cut.

In each trial, the sequence was shown twice, once with a degraded frame in the "target" segment (the impaired sequence) and once with no impairment (the unimpaired sequence). The order of presentation, that is, whether the unimpaired sequence was shown preceding the unimpaired sequence or following it, was varied randomly.

Subjects were required to view both sequences successively before responding with a button press, a two alternative, forced choice procedure. The "button box" consisted of three buttons; a left hand button which was used to initiate trials, a central, and a right hand button. If they believed the degraded frame was in the first sequence they were instructed to press the button in the central position, if they believed the degraded frame was in the second sequence they were instructed to press the right hand button.

For example, to estimate the quantization threshold for the second frame, prior to the scene cut, the participant would be shown two sequences, one with the second frame degraded to a mid-range quantization value, and one with all frames at $Q = 5$ (i.e.

unimpaired). The participant would then indicate, by a button press, which of the two sequences contained the degraded frame. If, over several trials, the participant was able to reliably determine which of the two sequences contained the degraded frame, the next sequences would present that frame degraded to a lower quantization value, making identification more difficult. If over several trials the participant was unable to reliably determine which sequence contained the degraded frame, in subsequent trials that frame would be degraded to a higher quantization value, making identification easier. This process would continue, adjusting the quantization either upwards and downwards, in gradually decreasing increments, until the presentation algorithm was able to settle on the lowest quantization level for which the participant was able to choose the sequence containing the degraded frame correctly at least four times out of six attempts. In cases where the participant was unable to distinguish reliably between the two sequences at any quantization level, the maximum quantization level of 255 was assigned.

The actual quantization level of the degraded frame, for each trial, was selected according to an adaptive psychophysical procedure known as Parameter Estimation of Sequential Tracking (PEST) (Taylor and Creelman, 1967). Threshold estimates were obtained, using PEST, for each of the three target sequences (Women, Mobile, and Football), at each of the Frame positions: the first, second, and third frames following the scene cut, and the first, second, and third frame preceding the scene cut. Eighteen threshold estimates were made by each subject. Subjects were tested on three different days, to assess practice effects.

Overall, the experiment consisted of 18 conditions in a 3 X 3 X 2 factorial, within-subjects, repeated measures design (Image sequence(3) X Frame position(3) X Direction (Forward/ Backward(2)). Eighteen psychophysical threshold estimates were made each day. The order of presentation was randomized for each subject, and rerandomized for each day.

Results & Discussion

In Experiment 1 I estimated threshold sensitivities of the human visual system to artifacts in the three frames just following a scene cut (Forward masking), or just preceding a scene cut (Backward masking). Threshold visual sensitivity was measured for eight subjects. Each participant made a total of 54 individual threshold estimates using a PEST type algorithm (18 per day, for three days). The threshold estimates measured the minimum quantization level (Q) at which visual impairments could be detected with 75 % accuracy. Results were analyzed using a four-factor, within subjects, analysis of variance (ANOVA): Direction (Forward, or Backward), Frame (first, second, or third frame from the scene cut), Image (Woman, Mobile, or Football), and Day (first, second, or third day). Estimated quantization thresholds are shown by Frame in Table 1:

Main Effect of Frame

Figure 10 shows a significant Main Effect of Frame [$F(2,14) = 44.3$, $MS_{Error} = 7879.0$, $p < .01$]. Looking at the graph see we that threshold values were greatest in Frame 1, dropping considerably between Frame 1 and Frame 2, and then slightly to Frame 3. Though this suggested that there may have been some masking still present in Frame 2, Newman-Keuls Post Hoc Tests revealed that the significant main effect of Frame must be attributed to the large difference between Frame 1 and Frame 2 [$p < .01$], as the difference between Frame 2 and Frame 3 was not significant. These results support the view that masking effects, though strong in the Frame 1, were virtually gone by Frame 2.

Forward condition: Quantization thresholds

Frame 1	Frame 2	Frame 3
155	55	35

Backward condition: Quantization thresholds

Frame 1	Frame 2	Frame 3
105	40	35

Table 1: Estimated Quantization thresholds

Main Effect of Direction (Forward vs. Backward)

There was no Main Effect of Direction, in other words there was no overall difference between thresholds for Forward or Backward masking. However there was an interaction between Direction and Frame.

Interaction between Direction and Frame

If we look at Figure 11, a graph of Forward and Backward conditions by Frame, we see that the Forward threshold was higher than the Backward threshold for Frames 1 and 2. There was a relatively large difference between Forward and Backward thresholds for Frame 1 [154.0 vs. 114.5] a smaller difference for Frame 2, [65.8 vs. 48.6], and no significant difference for Frame 3. [43.1 vs. 42.2]. Overall, the interaction between Direction and Frame was significant [$F(2, 14) = 4.5$, MS Error = 3014.3, $p < .05$].

Newman-Keuls Post Hoc Tests showed that Forward and Backward conditions were significantly different for Frame 1 [$p < .01$], but not for Frame 2 or Frame 3. Thus, there was a reliable difference between Forward and Backward masking thresholds for Frame 1, but not for Frame 2 or Frame 3. It is possible that with a larger sample size the Frame 2 comparison [$p = .08$] would have reached significance.

Newman-Keuls Post Hoc Tests also revealed that the significant main effect of Frame resulted from the large difference in masking effect between Frame 1 and Frame 2, for both Forward [$p < .01$] and Backward [$p < .01$] masking. The difference between Frame 2 and Frame 3 was not significant for either Forward or Backward masking conditions. However, it may be noted that the difference between Frame 2 and Frame 3 approached significance in the Forward condition [$p = .063$], and might realistically have been expected to reach significance with a larger sample size. On the other hand, there was no difference between Frame 2 and Frame 3 in the Backward condition. Thus, one can argue that masking may extend to Frame 2 for the Forward condition

only. Generally, though, masking was fairly minimal, even the second frame following the scene cut.

The assumption that masking was absent by the third Frame in both Forward and Backward conditions was further supported by the convergence of threshold values for both Forward and Backward conditions: Looking at Figure 11 we see that by the third Frame the two points lie on top of one another.

The Forward masking results of Experiment 1, replicate and support the results of a previous study by Tam et al (1995). They found Forward masking effects to be strongest in the first frame following a scene cut (SOA = 33 ms), minimal in the second (SOA = 66 ms), and negligible in the third. However, Tam et al., did not study the effects of Backward masking.

These results also support the research of Seyler and Budrikis (1964), who found that masking effects were greater near the scene cut for Forward masking than for Backward masking when using naturalistic video images. As well, Lee and Dickinson (1994) found that Forward masking effects were greater than Backward masking effects for the first frame on either side of a scene cut. The magnitude of the difference that I found (for Frame 1) is not as large as that measured in Lee and Dickinson's study: They found the Backward masking effect to be about 1/4 as strong as the Forward masking effect. Our results showed the Backward masking effect to be about 3/4 of the strength of the Forward masking effect.

Main Effect of Image

Figure 12 shows that there was an overall difference in threshold value depending on the target image, [$F(2, 14) = 18.7$, $MSE_{Error} = 4358.4$, $p < .01$]. Newman-Keuls Post Hoc Tests show that there was a significant difference between threshold values for "Women" [77.5] and "Football" 102.1], [$p < .01$], and between "Mobile" [54.5] and "Football", [$p < .01$], as well as between "Women" and "Mobile"

[$p < .05$]. Thus "Football" hid more impairment than either "Women" or "Mobile", and "Women" hid more impairment than "Mobile". In order to look more closely at this effect we next considered the interaction between Image and Frame.

Interaction between Image and Frame

Figure 13 shows a clear interaction between Image and Frame [$F(4, 28) = 15.4$, $MSE_{Error} = 1428.8$, $p < .01$]. We see that, as in the Main Effect, for each frame the image "Football" produced the highest threshold for all Frames, [181.3, 74.6, 50.4], followed by the image "Women" [135.8, 55.9, 40.7], and the image "Mobile" in the lowest, overall, position [85.6, 41.1, 36.7]. Thus, at each Frame position, impairments in the image "Football" were the most difficult to see, followed by impairments in the image "Women", and the image "Mobile".

Newman-Keuls Post Hoc Tests comparing Images by Frame, revealed that the for Frame 1 there was a significant difference between all three Images: "Women" vs. "Mobile", [$p < .01$]; "Women" vs. "Football", [$p < .01$]; "Mobile" vs. "Football", [$p < .01$]. For Frame 2 there was a significant difference between "Women" and "Football" [$p < .05$]; between "Mobile" and "Football" [$p < .01$]; but not between "Women" and "Mobile". For Frame 3, there was no significant difference between any of the images.

Thus, the effect of Image depended on Frame, and masking strength depended on Image content for Frames 1 and 2. Finally, the lack of a reliable difference between Images for Frame 3, supported the contention that masking effects were virtually absent by the third frame.

Main Effect of Day

Figure 14 shows that there was a small Main Effect of Day [$F(2, 14) = 31.8$, $MSE_{Error} = 1421.0$, $p < .01$]. Threshold values were $Q = 98.5$ for Day 1, dropping to $Q = 68.5$ on Day 2, and, again, very slightly to $Q = 67.1$ on Day 3. Newman-Keuls Post

Hoc Tests confirm what appears obvious here: There is a significant difference between Day 1 and Day 2 [$p < .01$], but no significant difference between Day 2 and Day 3.

The overall difference in threshold between Day 1 and Day 2, indicates that subjects improved in their ability to detect quantization impairments through practice. The lack of difference between threshold between Day 2 and Day 3 suggests that subjects, now familiar and comfortable with the task, had reached a stable level of performance. Thus, we chose to use only Day 2 and Day 3 results in our estimate of quantization thresholds. Table 9 includes a list of quantization thresholds determined by averaging over subjects, for Days 2 and 3.

It may be noted that thresholds reached an asymptote at a Q of approximately 35 - 40. At these levels the test images are perceptually indistinguishable from Reference quality at $Q = 5$. It is important to remember that impairments were present for extremely short durations (33.3 ms.) within longer moving sequences. Though the main masking effects were understood to occur after a scene cut, there may have been slight masking effects simply due to the imbedding of the impaired frame within a sequence of unimpaired frames.

Experiment 2: Subjective Quality Assessment

(Forward and Backward Masking)

It would be valuable to verify and extend threshold masking results to suprathreshold levels in order to determine whether larger amounts of impairment can be comfortably tolerated near scene cuts. Girod (1992) points out that we still have much research to do to adapting image coding to the parameters of the human visual system. He says, "... until this search has come to a successful end, we will have to

include a subjective evaluation of image quality in the design of image communications systems" (p. 250). In response to this need I performed a second experiment. In Experiment 2, I assessed image quality at suprathreshold levels of impairment in frames both preceding and following a scene cut. The goal of the experiment was to establish the level of image degradation which, though noticeable, was not subjectively bothersome to viewers. It was anticipated that some degree of detectable image degradation would be acceptable.

Seyler and Budrikis (1964), studying only Forward masking, found considerable latitude in what was considered by viewers to be subjectively acceptable. Their techniques, while ingenious, were less sophisticated than those available today. Using a low-pass filter, they reduced image bandwidth, and thus spatial detail, after a scene change. This provided a variation in the degree of image "blurring", undifferentiated across all frequencies. Furthermore, though Seyler and Budrikis were able to use a moving video image prior to a scene cut, they were technically limited to the use of a still scanned video slide after the scene cut.

This study attempted to improve on the Seyler and Budrikis technique, both through improved technical means to test subjects, and by an improved image degradation algorithm using the discrete cosine transformation, as outlined in the JPEG/MPEG Compression Standards. Forward and Backward masking were both measured. As in Experiment 1, the results of the present study have direct applicability to coder design.

Method

Video Sequences

Video sequences followed the same format as in Experiment 1.

Image Processing

Impaired video frames were produced by processing target images at a range of quantization levels ranging from threshold Q values (for each individual Frame) to $Q = 255$, in five equal steps. Table 2 shows actual quantization values for each of the six Frames: The lowest value shown for each Frame is the measured threshold for that Frame. The highest value shown for each Frame is 255, the maximum possible quantization.

A quantization coefficient of 5 created a high quality Reference image for comparison. A quantization coefficient of $Q = 255$ created an image with the greatest possible impairment; one in which most of the discrete-cosine transformed (DCT), 8 X 8 pixel, blocks in the frame were displayed at the mean level of luminance and mean colour for the block. As in Experiment 1, all unimpaired video frames were processed at $Q = 5$, for uniformly high image quality.

For the Forward condition, impairments were introduced in either the first, second, or third frame following the scene cut. For the Backward condition, impairments were introduced in either the first, second, or third frame preceding the scene cut. (See Figure 8). Video sequences were processed using an MPEG-2 codec. In order to enable individual frame quantization, the Group-of-Pictures length was set to one ($GoP = 1$). The unique aspect of this experiment was that the masked frames were all compressed at quantizations above threshold visibility.

Forward condition: Quantization levels

	Frame 1	Frame 2	Frame 3
Level 1	155	55	35
Level 2	180	105	90
Level 3	205	155	145
Level 4	230	205	200
Level 5	255	255	255

Backward condition: Quantization levels

	Frame 1	Frame 2	Frame 3
Level 1	105	40	35
Level 2	140	95	90
Level 3	180	150	145
Level 4	215	200	200
Level 5	255	255	255

Table 2: Quantization Values shown by Frame

Participant /Viewers

Twenty-four viewers participated in the study. None had any previous exposure to the assessment task or to the video sequences. All had either normal or corrected to normal visual acuity, and normal colour vision. Viewers were paid for their participation. Viewers ranged in age from 14 to 56 years of age.

Display

The display was the same as for Experiment 1.

Design and Procedure

The viewing apparatus allowed two viewers to be tested at a time. For the Forward masking condition the masking segment image "flower" was presented first, and the target segment image (either "Women", "Mobile", or Football") was presented last. For the Backward masking condition, this order was reversed: the target segment was presented first, and the masking segment was presented last. All Frames, except the target Frame, were always shown at a quantization level of 5 ($Q=5$), that is, unimpaired.

There were two sequences presented for each trial, a 'Test' sequence and a 'Reference' sequence: Each of these sequences was composed of a mask segment and a target segment. The Test sequence contained an impaired frame in the target segment. In each trial presentation these two sequences were shown sequentially: The first sequence shown was called the 'A' sequence, the second sequence shown was called the 'B' sequence. The impaired frame appeared in either the 'A' sequence or the 'B' sequence, varying randomly from trial to trial.

In each trial presentation the 'A' and 'B' sequences were shown twice (AB AB). There was a one second pause between the 'A' and the 'B' sequences, and a pause of approximately three second between the first 'AB' set and the second 'AB' set. Viewers

were instructed to watch the first AB set, and to look for differences, but not to actually make a rating until the second AB presentation. Viewers were asked to rate the 'A' sequence and the 'B' sequences separately. This provided independent estimates of degraded (Test) sequences and non-degraded (Reference) sequences.

Each viewer rated 90 'ABAB' sequences, rating the 'A' and 'B' sequences independently using the Double Stimulus, Continuous Quality, subjective rating scale, recommended by the International Telecommunications Union (ITU) in Recommendation 500. A sample page of the booklet given to subjects in the study, showing this scale is provided in Figure 9.

As in Experiment 1, impairments were studied in the three frames following the scene cut (Forward masking) and the three frames preceding the scene cut (Backward masking). The masking image (Flower) was combined with each of the three target images (Women, Mobile, Football). Five impairment levels were used, ranging from threshold level to $Q = 255$ (maximum impairment) in five equal steps. Threshold and step size varied by frame. (See Table 2).

There were 90 conditions in the Experiment defined by Image (3), degree of image degradation (5), and position of image degradation (3). For each condition, subjects rated the Reference image and the Test image separately. Five impairment levels were used for each of the three frames preceding, and the three frames following the scene cut. Three different images were rated at each impairment level, for each frame position. Thus, we used a 5 (Impairment Level) X 3 (Frame position) X 3 (Image) X 2 (Direction), within subjects design. Each participant rated each condition twice. In all, each participant made 360 individual ratings.

Results & Discussion

In Experiment 2, we estimated the image quality of video sequences which contained impairments in an individual frame following (Forward masking), or preceding (Backward masking) a scene cut. Using the Double Stimulus Continuous Quality rating scale method (ITU recommendation 500), each participant rated 180 impaired Test sequences and 180 unimpaired Reference sequences. The results were analyzed using six separate within-subjects ANOVAs, one for each of the six frame positions (3 following the scene cut, and three preceding it).

Image quality ratings for each Frame were analyzed using individual four-factor analyses of variance (ANOVA's): unimpaired (Reference) sequence vs. impaired (Test) sequence, Quantization (5 levels), Image (Woman, Mobile, or Football), and Repetition (first and second presentation). The three frames following the scene cut (Forward masking), and the three frames preceding the scene cut (Backward masking) were each analyzed separately. An ANOVA across frames was not performed as the five quantization (Q) values varied by frame according to individual frame thresholds. Figures 15 a and 15 b show mean image quality ratings for Forward and Backward masking studies, respectively. Filled symbols show results for Reference conditions for each Frame. Open symbols show results for Test conditions. Open triangles show results for the Frame 1, open squares show results for Frame 2, and open circles show results for Frame 3. The X-axis shows actual Quantization values in the Test sequences, the Y-axis shows the image quality rating on a 100 point scale.

Main Effect of Reference vs. Test ratings:

In Figure 15 a, Forward condition, and Figure 15 b, Backward condition, filled symbols represent Reference data, open symbols represent Test data. As expected, the image quality of the Reference sequences were rated uniformly high because these were

Table of Abbreviations:

Forward condition:

Frame 1: **(FF1)**

Frame 2: **(FF2)**

Frame 3: **(FF3)**

Backward condition:

Frame 1: **(BF1)**

Frame 2: **(BF2)**

Frame 3: **(BF3)**

Table 3: Abbreviations used in the Texts

presented at $Q = 5$ for all Frames. The first Frames (FF1, BF1), show little noticeable drop for Test, compared to Reference. This means that subjective image quality was affected negligibly by image degradation in the first Frame before and after the scene cut. However, in the second and third Frames (FF2, FF3, BF2, BF3), Test ratings drop noticeably below Reference, revealing that image quality was reduced by the introduction of degradation in the second and third Frames.

ANOVA's indicated that there was a significant main effect of masking, when comparing non-degraded (Reference) to degraded (Test) sequences for all three Frames. The F ratios for Reference vs. Test were as follows: FF1 [$F(1, 23) = 30.7$, $MSE_{Error} = 89.2$, $p < .01$]; FF2 [$F(1, 23) = 45.7$, $MSE_{Error} = 1135.5$, $p < .01$]; FF3 [$F(1, 23) = 55.9$, $MSE_{Error} = 1250.4$, $p < .01$]; BF1 [$F(1, 23) = 19.7$, $MSE_{Error} = 286.6$, $p < .01$]; BF2 [$F(1, 23) = 73.4$, $MSE_{Error} = 1275.9$, $p < .01$]; BF3 [$F(1, 23) = 96.2$, $MSE_{Error} = 1491.1$, $p < .01$]. Thus, for each Frame, subjects were able to reliably distinguish between Reference and Test video sequences, rating Test lower than Reference. This was expected as quantization levels ranged from threshold (as measured in Experiment 1) upwards.

Interaction of Reference and Test ratings by Quantization level⁶

In Figure 15 a and Figure 15 b we see that as Quantization (Q) level increased, image quality ratings dropped in Frames 2 and 3. In Figure 15 a we see that the largest drop occurred in the third Frame following the scene cut (FF3). There was a slightly smaller drop in the second Frame (FF2), but no noticeable drop in the first Frame (FF1). In Figure 15 b we see a similar pattern. Ratings dropped as Q increased for the

⁶ In considering the interaction of Reference and Test scores it is important to keep in mind that Reference ratings were, by definition, unchanged at any quantization level, because quantization impairments were only introduced in the Test condition. Thus, the interaction represents the effect of Quantization level on image quality ratings, and measures how this compares with unimpaired images.

third Frame preceding the scene cut (BF3), dropped less for the second Frame (BF2), and dropped only slightly for the first Frame (BF1).

Individual ANOVA's for each Frame indicate that there was a significant interaction of Quantization (Q) level by Reference and Test scores, for all Frames, with the exception of FF1 where changes in Q had no effect on Test image ratings. FF2, FF3, BF1, BF2, and BF3 all showed significant interactions. The F ratios were as follows: FF2 [$F(4,92) = 23.1$, $MSE_{Error} = 152.0$, $p < .01$]; FF3 [$F(4,92) = 37.1$, $MSE_{Error} = 190.6$, $p < .01$]; BF1 [$F(4,92) = 7.6$, $MSE_{Error} = 58.8$, $p < .01$]; BF2 [$F(4,92) = 34.6$, $MSE_{Error} = 115.7$, $p < .01$]; BF3 [$F(4,92) = 53.1$, $MSE_{Error} = 145.8$, $p < .01$].

Thus, for FF2, FF3, BF1, BF2, and BF3, increases in Q-level significantly lowered subjective quality ratings in the Test condition. In other words, subjects rated more highly compressed images more poorly. In order to examine whether the difference between Reference and Test ratings were significant at all Q levels, and whether Test scores decreased significantly between adjacent Q levels, Newman-Keuls Posts Hoc Tests were performed on individual subject means. Detailed comparisons of Reference and Test ratings are presented in Table 4; detailed comparisons of adjacent Q levels are presented in Table 5:

Newman-Keuls Tests show that (with the exception of BF1) Test ratings fall significantly below Reference ratings for all Levels of Quantization except Q1. Given that Q1 was set at threshold for each Frame position, it was not expected that there would be a large difference between Reference and Test for that Level.

Although there was a significant interaction between Reference and Test by Quantization Level for the first Frame prior to the scene cut (BF1), individual comparisons of adjacent Q values in BF1 are all non-significant. This means that

Post Hoc Test results: Reference vs. Test by Quantization (Q) Level

	Q1	Q2	Q3	Q4	Q5
FF1 [†]	---	---	---	---	---
FF2	n.s.	**	**	**	**
FF3	n.s.	**	**	**	**
BF1	n.s.	n.s.	**	**	**
BF2	**	**	**	**	**
BF3	n.s.	**	**	**	**

** $p < .01$

[†] Post Hoc Tests were not performed for FF1 because the ANOVA for FF1 showed that there was no significant interaction between Reference and Test.

Table 4: Post Hoc Comparisons of Reference and Test scores by Quantization level

Post Hoc comparisons of ratings for adjacent Q levels within Test:

	Q1-Q2	Q2-Q3	Q3-Q4	Q4-Q5
FF1 [†]	---	---	---	---
FF2	**	**	*	**
FF3	**	**	*	**
BF1	n.s.	n.s.	n.s.	n.s.
BF2	**	**	n.s.	n.s.
BF3	**	**	**	n.s.

* $p < .05$

** $p < .01$

[†] Post Hoc Tests were not performed for FF1 because the ANOVA for FF1 showed that there was no significant interaction between Reference and Test.

Table 5: Post Hoc Comparisons of Test scores by adjacent Quantization levels

quality did not drop enough between adjacent Q values to be significant. However, because actual quantization increments in BF1 were about half those in Frames 2 or 3, I also compared every other Q Level of BF1. Results were as follows: Q1 - Q3, significant [$p < .05$]; Q2 - Q4, significant [$p = .05$]; Q3 - Q5, approached significance [$p = .055$].

Saturation Effects: I considered why the final Post Hoc BF2 and BF3 comparisons (Q4 - Q5) failed to reach significance: This may suggest something about the overall time course of masking. Looking at Figures 15 a and 15 b, we see that the curves for the second and third Frames following the scene cut (BF1 and BF2), dropped more steeply than those of their Forward counterparts (FF1 and FF2), but flattened out at the end. Curves for the second and third Frames following the scene cut FF2 and FF3 dropped more gradually, though the final quality ratings for Frame 3 (FF3 and BF3) at Q = 255 were very close (47.6 and 45.1, respectively). This observation is consistent with the view that saturation had taken place in BF2 and BF3, but not in FF2 or FF3. What saturation means is that viewer rating had reached an asymptote; perceived image quality was as bad as it could get under the present viewing circumstances. In other words, prior to saturation, increases in Quantization level resulted in decreases in perceived quality, thus decreasing ratings. After saturation, further increases in Quantization had no affect on perceived quality, so ratings remained stable.

Forward vs. Backward masking: If we compare the curves in Figures 15 a and 15 b we see that in Frame 1, subjective quality remained high for both Directions (FF1 and BF1), though BF1 had dropped slightly in quality by Q5. In Frame 2, subjective quality dropped more quickly at first in the Backward direction (BF2) than in the Forward direction (FF2), though the slope of BF2 leveled off by Q4 in the saturation effect mentioned above. Quality ratings dropped even more quickly in Frame 3,

overall. They dropped more steeply in the Backward direction at first (BF3) than in the Forward direction (FF3) but also leveled off by Q4. This observation supports the view that, until saturation, Forward masking is stronger than Backward masking

Interaction of Reference and Test scores by Image:

Looking at Figure 16 we see that, for both Forward and Backward masking studies, Test sequences using the image "Football" were, on average, rated the highest for each Frame, the image "Women" was rated intermediately, and the image "Mobile" was rated the lowest. From our graph it appears that the image "Football" could tolerate the most severe level of quantization without a drop in perceived quality. The image "Women" could tolerate an intermediate amount of quantization, and the image "Mobile", the least amount of quantization, overall.

We can see, however, from Figure 16, that these differences were greatest in Frame 1, reduced in Frame 2, and quite small by Frame 3. Our analysis showed that there was a significant interaction of Reference and Test scores, by Image, for all frames; FF1 [$F(2,46) = 16.0, p < .01$]; FF2 [$F(2,46) = 30.3, p < .01$]; FF3 [$F(2,46) = 26.8, p < .01$]; BF1 [$F(2,46) = 3.9, p < .05$]; BF2 [$F(2,46) = 25.8, p < .01$]; BF3 [$F(2,46) = 42.5, p < .01$]. Thus, overall, Image had a significant effect on how much Test ratings differed from Reference ratings.

In order to examine the effects of each picture independently, Newman-Keuls Post Hoc Tests were performed for each Frame: Reference and Test ratings were compared by Image, and then Test ratings for the three Image sequences were compared. Table 6 summarizes Post Hoc Reference and Test comparisons, and Table 7 summarizes individual image comparisons.

Post Hoc Comparisons: Reference vs. Test ratings by Image sequence

	Women	Mobile	Football
FF1	*	**	n.s.
FF2	**	**	**
FF3	**	**	**
BF1	*	**	n.s.
BF2	**	**	**
BF3	**	**	**

* $p < .05$

** $p < .01$

Table 6: Post Hoc comparisons of Reference vs. Test quality ratings by Image

Significant values in Table 6 indicate that perceived image quality for a given Test Image sequence was significantly different from its Reference counterpart. The lack of significance in Frame 1 (FF1 and BF1) for the image "Football" indicated that viewers perceived no difference in quality between impaired and unimpaired Images for the first frame on either side of the scene cut with this Image. Significant results for the other two image sequences showed that viewers did perceive a decrease in quality when the image was impaired. Significant results for the second and third frames (FF2, FF3, BF2, and BF3) indicated that viewers perceived a decrease in quality for all images when impairment was added to the Image sequence.

In Table 7, the first two frames preceding the scene cut (BF1 and BF2), and for the third frame following the scene cut, no difference in quality was perceived between the Image "Women" and the Image "Mobile". All other comparisons showed significant differences in perceived quality by Image. In other words, other than the noted exceptions, any two images in a given Frame were perceived to be of different relative qualities. Whenever there was a perceived quality difference, the image sequence "Football" was always rated the most highly, the Image sequence "Women" was rated intermediately, and the Image sequence "Mobile" was rated as being of the poorest quality.

In a final set of Post Hoc Tests, Reference Images were compared in order to determine whether this variation could be attributed, in part, to viewers' preference for one Image over another. There was no difference in perceived quality for any Reference Images in the Frames preceding the scene cut (BF1, BF2 and BF3), or in Reference Images in the second frame following the scene cut (FF2). There was a difference in perceived quality when comparing Reference Images in "Women" and "Mobile", for the first and third frames following the scene cut (FF1 and FF3). It may

Comparison of Images

	Women vs. Mobile	Women vs. Football	Mobile vs. Football
FF1	**	**	**
FF2	**	**	**
FF3	n.S.	**	**
BF1	n.S.	**	**
BF2	n.S.	**	**
BF3	**	**	**

****** $p < .01$

Table 7: Post Hoc comparisons of Test Images by Frame

be noted that these differences were all relatively small ($.01 < p < .05$) and in no way coincided with the differences found between Test Images. It was therefore concluded that differences between Test Image ratings were not due to Image preference independent of impairment.

However, it may be possible to account for the rating difference by considering how the image coder (which performs the actual quantization on frames) processes images. The image "Mobile" contains the largest amount of high frequency information, and thus places the highest demand on the image coder. The image "Women" is of intermediate complexity. The image "Football" contains the least amount of high frequency information, and so places the least strain on the coder. Thus, at a given quantization actual image quality may vary considerably. A high demand image will produce a lower quality image. What this means is that, for a given Q level, a frame within a relatively low demand image sequence like "Football", will actually be less severely impaired than a frame within a higher demand image like "Mobile". If this is true than, for example, the image "Football" would tend to be perceived as higher quality than the other images simply because, at any given Q Level, it actually was less impaired. Thus, our results may indicate a confound between perceived image quality differences and actual quality differences.

Three-way interaction of Reference and Test scores by Image, and by Quantization:

Looking at Figures 17 a - f by Frame, we see that, for all Frames except the first frame following the scene cut (FF1), perceived image quality varied by Quantization and Image. In the first frame preceding the scene cut (BF1), perceived quality decreased very slightly as quantization increased, and the rate of this decrease was somewhat effected by Image. For frames 2 and 3 (FF2, FF3, BF2, and BF3), perceived

image quality decreased steadily as quantization increased, at a rate that varied by Image.

For all Frames except FF1, the Image "Football" was rated of highest overall quality, the image "Women" of intermediate quality, and the image "Mobile" of lowest quality.

F ratios for these interactions are as follows: FF1 n.s. ; FF2 [F (8,184) = 6.0, MSEError = 69.1, $p < .01$]; FF3 [F (8,184) = 14.9, MSEError = 102.3, $p < .01$]; BF1 [F(8,184) = 5.5, MSEError = 40.0, $p < .01$]; BF2 [F(8,184) = 4.4, MSEError = 82.8, $p < .01$]; BF3 [F(8,184) = 3.4, MSEError = 83.5, $p < .01$]

Newman-Keuls Post Hoc testing confirmed what Figure 17 a-f shows: With the exception of the first frame following the scene cut (FF1), perceived quality of the Image "Football" dropped the most slowly as Quantization (Q) increased; perceived quality of the Image "Mobile" dropped the most quickly as Q increased; perceived quality of the Image "Women" dropped at an intermediate rate.

For the first frame preceding the scene cut, (BF1), as we can see from Figure 17 d, perceived quality for the Image "Football" did not drop at all as Quantization increased. Post Hoc Tests show that there was no significant difference between Reference and Test ratings at any Q level for this Image. It was noted, also, that in the third frame preceding the scene cut (BF3), at the maximum quantization of $Q = 255$, quality ratings for the "Football" dropped less than twenty points (a single grade on our subjective rating scale).

The argument that some images that are less demanding for image coders and so may be of higher actual quality at a given quantization level, is supported here. Ratings of all three images seem consistent with this interpretation. Quality ratings fell consistently from Frame 1 to Frame 2 to Frame 3. When looking at Ratings for independent images (Figure 17) we see that each image followed this pattern fairly consistently, falling by Frame. We see, also, that images generally maintained their

relative perceived quality order as they did so, though the graphs for Frame 3 show some crossover between ratings for "Women" and ratings for "Mobile". These crossovers are most likely due to reduced overall masking effects.

The image "Football" dropped the most slowly, overall, and was rated the highest; "Mobile" dropped the most quickly and was usually rated the lowest; "Woman" dropped at an intermediate rate and was usually rated intermediately. This is the general trend that would be expected if there were actual differences in image quality as a result of differing levels of image artifacts.

Interaction of Test and Reference scores by Repetition:

There was some effect of Repetition on Test ratings for all Frames. Generally, participant ratings averages decreased slightly for the second Repetition. Examination of these effects showed that they were all relatively small, and could be attributed to a slight improvement in the ability of subjects to detect impairments with practice.

General Discussion

Two experiments were performed to measure the level of image quantization /compression that could be hidden by visual masking at a scene cut. Masking was studied in each of the first three frames following a scene cut (Forward masking), and each of the last three frames prior to a scene cut (Backward masking).

In Experiment 1, quantization thresholds were estimated for each subject, using a PEST algorithm (Taylor and Creelman, 1967). A threshold was defined to be the level of quantization at which a participant could reliably distinguish a degraded image from

a non-degraded image 75% of the time. Estimates were made for three images ("Women", "mobile", and "football").

A strong masking effect was observed (Figure 10) in the first frame following, and preceding a scene cut, where a high degree of compression was found to be unnoticeable. A weaker masking effect was found in the second frame following the scene cut. Masking effects were virtually absent by the second frame before the scene cut, and by the third frame after the scene cut. The results of the Forward condition in Experiment 1 closely replicated those of a recent study by Tam et al. (1995) who had previously established threshold parameters for Forward masking. Like the results of the Tam et al. study, this study found that threshold masking effects did not extend reliably beyond the first 33 ms, the length of a single video frame. It was noted that this was considerably less than the 100 ms masking duration predicted by Girod's (1989) w-model. Neither Tam et al. nor Girod address Backward masking.

Lee and Dickinson (1994) looking at masking in the first frame on either side of a video scene cut found that Forward masking effect was stronger than the Backward masking effect when using naturalistic targets and masks. Our results were in agreement with theirs (Figure 11), though the magnitude of the difference in masking strength was found to be only about 25 % of the difference they reported.

Classical literature has shown that under dichoptic viewing conditions, Backward masking is stronger than Forward masking (Breitmeyer and Ganz, 1976). Thus, these results cannot be explained in any straight forward way by the classical masking literature. Given this apparent discrepancy, it must be assumed that our masking "task" differed in some critical way from the tasks presented in classical masking literature. We will address this discrepancy later in the discussion.

An overall effect of type of target image sequence was also found (Figure 12); the sequence "Football" hid the most compression so could be quantized the most severely.

The sequences "Women" and "Mobile" hid less, respectively, permitting less quantization.

An examination of the interaction between Frame and Image revealed that the effect of Image depended on Frame (Figure 13). In Frame all 1, image sequences were found to be different from each other. In Frame 2, the image sequence "Football" was different from "Women" and "Mobile", though they did not differ from each other. No difference was found between image sequences for Frame 3. Different sequence thresholds suggest a possible interaction between spatial and temporal masking. In order to know whether this was actually the case it would be necessary to determine how much of the difference between image thresholds was caused by differential image artifact content (due to variation in image sequence demand on the quantization coder).

A practice effect was found between Day 1 and Day 2., but not between Day 2 and Day 3 (Figure 14). Day 1 thresholds reflected a lack of task familiarity, and a lack of 'strategy'. The overall difference in threshold between Day 1 and Day 2, indicates that subjects improved in their ability to detect quantization impairments through practice. By the end of Day 1 most subjects reported that they had developed strategies for maximizing their ability to see impairments. These usually included maintaining a visual fixation on some part of the screen, commonly an area where movement during the target sequence was less noticeable. Day 2 and Day 3 thresholds did not differ. The lack of difference between Day 2 and Day 3 thresholds suggests that subjects, now familiar and comfortable with the task, had reached a stable level of performance. Thus, we chose to use only Day 2 and Day 3 results in our estimate of quantization thresholds. Table 9 includes a list of quantization thresholds determined by averaging over subjects, for Days 2 and 3.

In Experiment 2 suprathreshold masking effects were studied. Subjects were asked to rate image quality for both Reference and Test images using a subjective rating

scale⁷. This made it possible to assess how artifacts near a scene cut affected perceived image quality. Paralleling the threshold study, the first three frames on either side of the scene cut were examined. The same image sequences were used in both Experiments to facilitate comparison.

Viewer ratings remained high in Frame 1 even at maximum impairment (Figure 15 a, b), indicating the same pronounced masking effect found in Experiment 1. Ratings diminished rapidly over the next two frames. The second and third frames following the scene cut (Forward masking condition) were rated highly enough to indicate some residual masking.

Using criteria described in the Recommendations section below, Frame 1 could be highly compressed with no reduction in perceived Image quality (Forward and Backward masking conditions). The second and third frames following a scene cut could be compressed to slightly more than the threshold levels measured in Experiment 1 (Forward masking condition). Considerable further compression could be achieved for Frame 2 and Frame 3, in applications for which moderate quality video transmission was acceptable (Forward and Backward masking conditions). Overall, these results are of particular interest for practical applications because they lend support the use of progressive image buildup in image coders.

Ratings for Forward masking appeared slightly higher, in general, than ratings for Backward masking, (see Figure 15 a & b). The particularly small difference in Frame 1 (FF1 and BF1) is probably attributable to a ceiling effect. Both Forward and Backward masking effects were strong enough to maintain high viewer ratings for this Frame even at $Q = 255$. Had we used an encoding process that allowed for more

⁷International Telecommunications Union (ITU) recommendation 500 subjective rating scale. (A sample of the scoring sheet used in the study is shown in Figure 9.)

extreme compression, Backward masking might have failed earlier than Forward masking, making the difference more pronounced, as it was in Experiment 1.⁸

For all Frames studied, there were reliable rating differences between image sequences (Figure 17) similar to those found in Experiment 1. Subjects rated the image sequence "Football" more highly at any given quantization value than the other two images. The image "Women" was rated intermediately, and the image "Mobile" was rated lowest. In other words, subjects found image compression to be less objectionable at any given Q value for the image "Football", then for "Mobile" or "Women".

The actual difference between image sequences in Frame 1 (Figure 17 a, & d), though reliable, was very small. This probably results from the same ceiling effect described above. If masking effects were strong enough relative to the amount of impairment present in the Image, perceived quality would remain high for all Images. However, if images had been degraded more severely, differences between them would have been more pronounced.

There were greater differences in perceived quality between Images for second frame (FF2, BF2) and the third frame (FF3, BF3), than in the first frame (Figure 17). This is what would be expected if there were true differences between Images that masking strength was no longer strong enough to overcome. The findings that apparent masking strength varied by Image (in Experiment 1), and that perceived quality varied by Image (Experiment 2), supported the general conclusion that masking strength varied as a function of spatial image properties. This lead to an obvious question about what these properties might be:

⁸Using this argument it is possible that a difference in the magnitude of introduced impairment might account for the discrepancy between the results of the present study, and the results reported by Lee and Dickinson (1994).

It was noted in the Results and Discussion section of Experiment 2 that differences in coder demand for the three images might explain differences in threshold and quality rating. The image sequences used in this study did vary in the amount of high frequency information they contained and, thus, were of differential difficulty for the coder to process. This would be expected to result in differences in the visibility of coding artifacts at any given quantization level.

In Experiment 1, for example, the Image sequence "Football" did not contain much high frequency information and so was less demanding to process. As a result there were fewer coding artifacts in "Football" at a given quantization level. It would be expected, therefore, that a frame in this sequence would have to be processed at a higher quantization level than a frame from a more demanding sequence in order to become impaired enough, i.e., in order to develop enough coding artifacts, to reach threshold.

In general, this means that coding artifacts varied with image processing demand. More demanding images had lower thresholds, while less demanding images, because they produced relatively fewer coding artifacts, had higher thresholds. Our experimental results support this interpretation. The less demanding Image "Football", had the highest threshold, overall, in Experiment 1 and the highest quality ratings in Experiment 2. The most demanding Image "Mobile", had the lowest threshold, overall, in Experiment 1, and the lowest quality ratings in Experiment 2. The Image "Women" was of intermediate processing difficulty, and was found to have both an intermediate threshold, and quality rating. Though this view was supported by the results of both experiments, it did not negate the possibility that other spatial and temporal properties of the image sequences affected Image perception as well. These possibilities will be addressed later in the discussion.

Generally, there were two main theoretical issues raised by this study: (1) Why a difference was found between Images in both experiments and, (2) why Forward masking appeared to be stronger than Backward masking in Experiment 1. We have already begun to address the first issue. In order to address the second issue an attempt was made to determine how masking at scene cuts differed from masking in classical masking studies. It was mentioned earlier that it is an oversimplification to assume that a single mask and target were used in this study. The "mask" and the "targets" used were each composed of a series of independent images. Each of the independent images on either side of the scene cut was correlated with its neighbours, while being displaced from them to a varying extent. Thus, each could be seen as, potentially, acting as a mask for its neighbours.

The basic argument for why a stronger apparent masking effect was found in the Forward masking condition than in the Backward masking condition, depends on recognizing this multiple masking effect. It is argued that what was observed in the Backward masking condition was primarily integration masking. What was observed in the Forward masking condition, however, was actually a combination of Forward integration masking across the scene cut, augmented by a Backward masking metacontrast/ interruption masking effect from the frame just following the impaired frame. It was this combined Forward and Backward effect that created the appearance of a stronger Forward masking effect. Due to the mechanics of the visual processing task, discussed below, a parallel combination effect did not take place in the Backward masking condition. For practical purposes this distinction is unimportant. Nonetheless, it is of theoretical interest since it permits the reconciliation of our results with classical masking theory.

Though this sounds simple, in principle, the mechanics are relatively complex: We know that basic visual system functioning consists either of smooth tracking movements, or of visual fixations of up to 350 ms broken up by saccadic jumps lasting

from 20 - 50 ms (Breitmeyer & Ganz, 1976; Humphries & Bruce, 1989). It has been hypothesized that there are two main visual processing channels responsible for these functions, a sustained processing channel, and a transient processing channel (Bruce & Green, 1990; Hogben & DiLollo, 1985; Breitmeyer & Ganz, 1976). Zeki (1987) has established physiological support for these claims. The sustained visual system responds relatively slowly and is responsible for processing of form and colour information. The transient visual system responds relatively quickly and is responsible for orientation to movement. Head and body movement, or continuous movements within the visual field will trigger smooth tracking movements of the eyes, allowing a continuously displaced image to be tracked and processed by the sustained visual system. Masking experiments have suggested that sudden movements within the visual field activate the transient visual system, which overrides sustained channel processing, "interrupting" any processing taking place at the time. On a macro level the two channels augment one another, allowing high speed responses when necessary, and slower assimilation of detail information when circumstances allow it. However, on a micro level, this cooperative process requires creates competition. It is this "competition" that brings about the masking effects we observed.

Breitmeyer and Ganz (1976), have theorized in detail about how this processes takes place. They distinguish between Type A (integration) masking and Type B (interruption) masking. In Type A masking they claim that target and mask either compete "for common peripheral spatial-frequency analyzing channels", or create a composite integration in a "spatial-frequency synthetic, ...contour-forming process". These are both types of within-channel integration effects that "[depend] on the sharing of spatial-frequency components by the mask and target " (p. 20). It is most likely that

Type A integration took place in the experiments reported here, and that its effect was relatively equal in both Forward and Backward conditions.⁹

Hogben and DiLollo (1985) cite Burr (1980, 1981) as theorizing that what takes place in integration masking is actually an inappropriate activation of sustained-system processing. Just at the scene cut, the visual system will most likely be engaged in a sustained-channel processing (either fixation, or smooth-tracking). Though the discontinuity of the sudden change at the scene cut is expected to induce transient-channel processing, Hogben and DiLollo measured a latency of approximately 30 -40 ms before the transient processing system was activated. They noted that if more than one frame was displayed within that period time a blurring effect occurred. This is what would be expected if the visual system were unable to differentiate between independent images and continued its sustained-system activation as if it were processing a single image, effectively an integration effect. This latency is almost exactly the length of a single video frame, and thus is in agreement with our experimental results for Backward masking.

Brietmeyer and Ganz (1976) have described Type B metacontrast (interruption) masking as the masking that occurs due to transient-on-sustained visual channel interference. In other words, rather than the intrachannel competitive process that occurs in Type A integration masking, Type B masking happens through interchannel interference. We have discussed the integration effect that appears to take place just after a scene cut, and prior to the activation of the transient-processing system. What appears to be happening next is that, at approximately 30 ms after the scene cut, or immediately following the first frame, the now activated transient system will have a

⁹ Though Breitmeyer and Ganz (1976) argue that even in integration masking Backward Masking should still be slightly stonger under dichoptic masking conditions.

Backward masking, transient-on-sustained inhibitory effect, effectively masking the previous frame.^{10, 11}

Thus, when impairments were placed in the first frame following the scene cut, in the Forward masking condition described in our experiments, this frame was masked both by Forward integration masking and by Backward transient-on-sustained inhibition. In the Backward masking condition, this dual effect did not occur, as the visual system was already engaged either in fixation, or in smooth tracking. Thus, at the point in the sequence where the impaired frame occurred, only an integration effect took place. Given the similarity of the impaired frame to the unimpaired frames that followed, this integration effect would have tended to make the impaired frame stand out, rather than disappear. In particular, the artifactual high frequency information produced at block edges would have enhanced impairment visibility, rather than reducing it. The frame immediately preceding the scene cut (Backward masking condition) would have been masked by the integration process already described, but not as strongly as it was in the Forward masking condition, since it was masked from one direction only. It is argued that this provides an explanation for why masking

¹⁰ It should be noted that Breitmeyer and Ganz specify that metacontrast masking only occurs when masks and targets do not overlap which they obviously do in a sequence of video images. However, if we look more closely at how they specify the difference between metacontrast masking and integration masking it becomes reasonable to make this claim: Generally, metacontrast masking occurs when a target image is masked by a similar mask that is shifted in position in space. In other words, the mask and target are similar but are displaced in space. Integration masking, on the other hand, occurs when an image in a position in space is replaced by a different image in the same position. Though the change in image at a scene cut is obviously similar to classical integration masking definitions, the change in image from frame to frame, within the target sequence itself, produces general image displacement in space, similar to the way in which classical metacontrast masks and targets do. Thus, it is argued that a legitimate metacontrast masking effect took place in our study.

¹¹ Breitmeyer and Ganz (ibid) argue that this metacontrast activity, as it suppresses sustained activity, functions to end previous sustained-system processing. By "interrupting" any further integration, the transient-system clears the way for new sustained-system processing, allowing for the separation of visual events, as well as their integration.

effects were stronger in the first frame following the scene cut, than in the first frame preceding the scene cut.

This approach also yields a possible explanation for the residual masking effects noted in the second and third frame following the scene cut. Girod (1989) considered the possibility that noise could be hidden by movement, but determined that this was unlikely because of the visual system's well developed ability for smooth tracking. He did not, however, explore the possibility of an interaction between movement and masking at scene cuts. If sequential image displacement following a scene cut initiates the smooth tracking movement necessary to integrate displaced images over time, it must be assumed that this initiation has some latency period. It has already been noted that the first frame after the scene cut would have been masked both by the frame prior to the scene cut, and by the frame following the impaired frame. However, when the impaired image was placed in the second frame following the scene cut, the integration masking effect should have been negligible. Nonetheless, the transient visual system, triggered by the sudden decorrelation at the scene cut, would have been fully engaged just as this impaired frame occurred.

It is known that the transient-visual system is relatively insensitive to high frequency information. So, although the transient system would have responded directly to the impaired frame, it would not have distinguished it easily from the surrounding unimpaired frames. (Recall that the impaired frame contained low frequency information that was virtually identical to the low frequency information present in the surrounding unimpaired frames. The impaired frames were only distinguished from the unimpaired frames by their high frequency information.) Thus, although there was no true masking effect taking place by the second frame, the similarity in low frequency image content, coupled with the dominance of the transient system at that point, would have been expected to result in some degree of insensitivity to the impaired frame.

At this point, given the spatial translation expected from frame to frame after the scene cut (as would be expected in any "moving" image), it would be assumed that the sustained visual system would take over with a continuous smooth tracking movement. However, in order to determine the appropriate direction and speed of tracking, the visual system needs some means of performing a correlation between images.¹² A minimum of two successive frames must occur after the scene cut in order to enable the visual system to make this calculation, i.e. A reference frame and a frame that has incurred some degree of correlated displacement. Once this occurs there must be some latency period while the system calculates the speed and direction, and engages the smooth tracking system. Although this is speculative, it is possible that due to this latency period, transient-system activation would extend into the third frame, accounting for the slight residual Forward masking effects that were observed in the experiments performed here.

We will now turn back to the question of why a difference was found between thresholds, and between quality perception, for the three Images used in this study. Though coding differences may account for differences in image thresholds and ratings, there were some clear differences between images that may effected experimental results. One of the main differences between the three images was the degree of image displacement from frame to frame, i.e. image "velocity". It is not immediately clear how this may have effected making effects, however Hogben and DiLollo's (1985) study of the suppression of visual persistence in apparent motion suggests one possible explanation. They found that target velocity had a clear effect on the duration of smear, which we have already suggested is correlated with the period in time prior to the activation of transient visual processing: A target with a velocity of 15 degrees per second was measured to have a "smear" duration of between 33 and 63 ms depending

¹² See Bridgeman (1978) for a model of how this correlational process may occur.

on luminance conditions; a target with a velocity of 10 degrees per second had a "smear" duration of between 20 and 53 ms; and a target velocity of 5 degrees per second had a "smear" duration of between 10 and 33 ms. This suggests that the integration period may be longer for higher velocity targets, though the actual mechanism involved is not immediately apparent.¹³ If this were the case we would expect that a difference would be found between image thresholds and ratings, for both Backward and Forward masking conditions. Since this was found to be the case, we cannot discount this possibility.

A final issue that cannot be resolved here, is that of the differences between images in the amount of high frequency information present in the different image sequences. Given that the MPEG image compression algorithm functions as a high frequency filter, it is possible that an impaired image embedded within an image sequence, such as "Football", which naturally contained less high frequency information, was less noticeable. Alternatively, an impaired image placed within an image sequence, such as "Mobile", which had a greater higher frequency information content, may have been more noticeable. This could account in part for differences in threshold and differences in perceived quality rating. Unfortunately, without being able to determine how much of the observed difference was actually due to coding differences caused by the same variation in high frequency information, there is no way to determine whether this was the case. As well, it is difficult to speculate about how much of an effect the high frequency artifacts created at block edges (by the MPEG compression process) may have had.

¹³ Hogben and DiLollo attempt to account for this effect using an "interpolation" model, arguing that motion "smear" occurs when a spatial filter, receiving input from a temporal filter, is able to resolve separate stimuli. "At wide inter-point separations (i.e. high velocities), successive points matched or exceeded the spatial domain of the filter, thereby becoming resolvable as distinct points which were seen as smear" (p. 457).

This brings up a final point: Seyler and Budrikis (1964) in their study of Forward masking effects reported a subjective masking effect lasting much as 780 ms. This represents approximately 22 video frames, or about 7-8 times the outer limit that we were able to measure. They "compressed" their data using a transient low-pass filter which reduced image bandwidth, effectively introducing impairment through the blurring of images. This method, because it eliminated all high frequency information (unlike the MPEG compression process), was likely an effective means of hiding noise. However, it is unlikely that the extended masking duration they reported can be fully accounted for by differences in compression algorithms, image variability, viewing conditions, or subjective rating procedures. What is more likely to have changed are the viewers themselves. Adult television viewers in 1964 did not normally watch television regularly and had not logged the tens of thousands of hours of television viewing that an average 1997 audience has. Image impairment that seemed acceptable to a general audience in 1964, might be far from acceptable to an audience in 1997.¹⁴

The investigation of masking effects in moving naturalistic video images may help to extend our knowledge of the complex interlocking visual processes that traditional masking studies have only begun to understand. While the interpretations presented here are, perhaps, not as parsimonious as might be hoped, the theoretical problems they attempt to address are far from straightforward. Nonetheless, this study has been useful for confirming some of the effects, such as the greater relative strength of Forward masking using naturalistic images, that had been reported in related earlier studies.

¹⁴ With this in mind, it may be of interest to note that one of the participants who participated in our suprathreshold study complained that she couldn't see anything wrong with most of the images. She hypothesized that because her television had poor reception, she was used to a poor quality image and no longer noticed it.

As well, this study has been useful in a more immediate sense, in that it has been able to provide a set of practical recommendations for video compression. These can be used to guide the design of video coding systems around scene cuts. The recommendations, presented below, include compression limits for the three frames preceding, and the three frames following a scene cut. Included are actual visual thresholds which can be used for "equivalent to Reference" quality video transmission, and two grades of suprathreshold compression, one for high/ transparent quality, and one for moderate quality video for less stringent requirements.

Some suggestions for future research have been laid out below, following the Recommendations section.

Recommendations

Determining appropriate quantization ranges:

Thresholds estimated in Experiment 1 can be applied directly to coding design as upper quantization limits for video transmission at near transparent levels of quality. The results of Experiment 2 indicate that quantization levels can be increased above threshold, particularly in the first frame on either side of the scene cut. What is needed now is a means for translating Experiment 2 ratings into design guidelines.

In order to do this we chose to follow the recommendations set out in the Federal Communications Commission's Advisory Committee's Final Technical Report on HDTV¹⁵. The Report recommends that a difference between Reference and Test be limited to 1/4 grade (five points) on the ITU one hundred point subjective scale. This 1/4 grade criteria is appropriate for the highest quality, near transparent, video transmission. Further, the Report suggests that a difference of up to one full grade (20 points) is acceptable for moderate quality video transmission, transmission that is not required to meet such exacting standards.

Described below is how these criteria for high and moderate quality transmission apply to the results of Experiment 2. Following this is a Results summary table (Table 8), and a table of recommended quantization limits by Frame for both quality ranges (Table 9). Table 9 also includes the quantization Thresholds estimated in Experiment 1. These are suitable for visually equivalent to lossless quality video transmission.

If we look at Figure 15 a, we see that Test curve FF1 (Frame 1, Forward condition) hardly decreases at all in comparison with the Reference line. Even at the

¹⁵ High Definition Television.

maximum compression of $Q = 255$, participant ratings remain high. Using our high quality, near transparent, criterion of 1/4 grade we see that the masking effect in the FF1 is strong enough to permit maximum compression with virtually no subjective quality loss. Looking at Figure 15 b, we see that Test curve BF1 (Frame 1, Backward condition) drops slightly, indicating that even with a very high level of compression (a Q value close to 200) high quality, near transparent, video quality can be maintained. Even with the maximum possible compression of $Q = 255$, the masking effect is strong enough that scores drop less than 1/2 grade, providing for moderate to high quality video transmission.

For FF2 (Frame 2, Forward condition) in Figure 15 a values decrease more rapidly, nonetheless, though the masking effect is considerably diminished in comparison with FF1, the entire compression range still falls within our single grade criterion. Thus, even the maximum compression of $Q = 255$ produces moderate quality video transmission. Furthermore, as we see from the graph of FF2, the curve remains within 1/4 grade until a Q value of about 120. This indicates that FF2 can be compressed to as much as $Q = 120$ and still produce high quality video transmission.

BF2 (Frame 2, Backward condition) in Figure 15 b, follows a steeper curve than FF2, indicating less masking in the Backward condition. Still, considerable compression may be achieved for moderate quality video transmission. Compression of up to about $Q = 175$ may be done to meet our one grade criterion. Further compression in BF2 of up to 255, produced difference scores of slightly more than one grade, but this difference is slight.

In FF3 (Frame 3, Forward condition) the masking effect was still apparent, though less than that in FF2 or BF2. Our 1/4 grade criteria holds only to about $Q = 35$, no different from the amount of compression allowed by our threshold study. However, as we can see from the curve (Figure 15 a), moderate quality video may be obtained with compression of up to almost $Q = 255$. Again, as in BF2, even

compression of $Q = 255$ in FF3 produces a subjective rating drop of about 1 1/4 grade, still relatively slight.

The Frame 3 Backward condition curve (BF3, Figure 15 b) drops off the most quickly, indicating reduced masking. We can see that compression to $Q = 35$ will still, as in FF3, produce a near transparent quality video image. However, again, this adds nothing to the compression limits measured in our threshold study. Further compression of up to about $Q = 100$ falls within our criteria for moderate quality video, and, again, compression up to the maximum of $Q = 255$ produces a score decrease of less than 1.5 grades, still not a great quality loss.

Threshold recommendations:

I choose to use our average threshold measurements from Days 2 and 3 as our actual "threshold" recommendations for our "equivalent to reference quality" video transmission (Table 9), though there are arguments to be made both for using more conservative or less conservative estimates. A more conservative approach would have been to use the bottom boundaries of our standard error bars, whereas a less conservative approach would have been to use the top error bars. The argument for the former position was, obviously, to account for participant variation, aiming more conservatively in order to account for participants who had lower than average thresholds.

The argument for the latter position was that the extreme conditions under which participants performed the threshold study -- the short duration of the video sequences, the constant repetition, etc., -- meant that threshold estimates were by their nature very conservative, since it is highly unlikely that any viewer would consciously ignore video content and spend any amount of time fixating their eyes on a fraction of the video screen. As well, it may be noted that most participants saw almost no image degradation at any impediment level at first. Only with concentration, repetition, and

Suprathreshold Results Summary Table

The following Table lists Average Difference Scores (Reference - Test) by Frame:

FF1

Quantization level	Q = 155	Q = 180	Q = 205	Q = 230	Q = 255
Difference scores	3.0**	1.3**	3.0**	3.0**	3.6**

FF2

Quantization level	Q = 55	Q = 105	Q = 155	Q = 205	Q = 255
Difference scores	2.8**	7.2*	13.5*	16.7*	19.9*

FF3

Quantization level	Q = 35	Q = 90	Q = 145	Q = 200	Q = 255
Difference scores	0.4**	9.1*	13.9*	20.8	25.7

BF1

Quantization level	Q = 105	Q = 140	Q = 180	Q = 215	Q = 255
Difference scores	1.5**	2.2**	3.4**	5.3*	7.7*

BF2

Quantization level	Q = 40	Q = 95	Q = 150	Q = 200	Q = 255
Difference scores	5.4*	11.6*	18.6*	22.1	22.8

BF3

Quantization level	Q = 35	Q = 90	Q = 145	Q = 200	Q = 255
Difference scores	3.2**	17.3*	23.2	28.2	28.0

* indicates a Q-value which falls within one grade (Moderate quality video transmission);

** indicates a Q-value which falls within 1/4 grade (High, near transparent quality video transmission).

Table 8: Summary Table of Results: Frame X Quantization (Experiment 2).

Recommendations: Summary Table

	FF1	FF2	FF3
Threshold*	155	55	35
High**	255	80 [†]	60
Moderate***	---	255	195

	BF1	BF2	BF3
Threshold*	105	40	35
High**	215	40	35
Moderate***	255	175	115

* Threshold = actual quantization thresholds for "equivalent to Reference" quality video.

** High = maximum quantization for high, near transparent quality video.

*** Moderate = maximum quantization for moderate quality video

[†] actual recommended maximum values are interpolated from Experimental results

Table 9: General Quantization Recommendations
(Threshold and Suprathreshold studies)

increasing use of strategies like preparing for a cut by fixating the eyes on a particular area of the screen, did participants begin to recognize impairments. Some participants who were scoring correctly at low thresholds, reported that they were uncertain about whether they were guessing or not, even when the PEST algorithm settled on a threshold. In the end, having considered both these positions we chose to split the difference, and settled on average values for our thresholds.

Suprathreshold Recommendations:

Forward

Frame 1: For the first frame after the scene cut, no average difference was greater than the ITU criteria of 5, therefore the maximum quantization of $Q = 255$ was acceptable even for high, transparent quality, video transmission.

Frame 2: For the second frame after the scene cut a quantization co-efficient of up to 80 was acceptable for high quality video transmission; for moderate quality video transmission, images could be further compressed by a Q of up to 255 (scored by participants as a difference of 19.9).

Frame 3: For the third frame after the scene cut the quantization co-efficient became unacceptable at values of over 60; however for less stringent, but still acceptable standards, images could be compressed by a Q value of up to about 195.

Backward

Frame 1: We can see that though Backward masking effects are not as strong as Forward masking effects, they are still quite distinct, in Frame 1. For the last frame just prior to a scene cut, quantizations of up to 215 are acceptable for high, transparent quality, video transmission. Further compression to a maximum compression of $Q = 255$, is acceptable for moderate transmission standards.

Frames 2 and 3: For the second frame prior to the scene cut, at the measured threshold Q-value of 40, the average difference between Reference and Test ratings was 5.4 (Table 8). This is very close to the acceptance criterion for near transparent quality video transmission. For the third frame, the threshold Q-value of 35 produced a difference of 3.2, well within the acceptance criterion for near transparent quality video transmission. However, as was argued earlier, a quantization value of 35 to 40, appears to be the threshold level for single frame degradation, when no further masking can be attributed to the effect of the scene cut. Thus, it would appear that for Frames 2 and 3, in the Backward masking condition, no further compression can be added above threshold if near transparent quality standards are desired. For moderate video standards, though, further compression up to a Q- value of 150 is acceptable for Frame 2, and up to 90 for Frame 3.

Future Research

There are several directions for further research suggested by the results of the present thesis:

1. Given that our results show that the maximum amount of compression noise may be hidden in the first frame before, and the first frame after a scene cut, it is reasonable to assume that noise can be hidden on both sides of a scene cut simultaneously. If this were the case than each video segment would act as mask for the noise in the other video segment. However, it is possible that the "blocky" nature of digital video compression might act to amplify the two degraded frames, when block "edges" were identical in both frames, as they could be if the compression levels were identical in the frames on both sides of the scene cut. If this were the case, differing

compression levels might ameliorate this problem. Of the suggestions given here, this would be the simplest to test, and would be of immediate practical use for image coding and compression.

2. It would be of interest to extend the preliminary research of Tam et al. (1995), into the interactive effects of noise hidden in more than a single frame. They report additive effects of masking whereby subthreshold noise hidden by a mask becomes suprathreshold noise when it is combined with noise in a second frame. This is important with respect to the practical purpose of this thesis: We cannot assume that subthreshold noise levels in a given frame can be combined with subthreshold noise levels in another frame and still remain subthreshold. Research done by Girod (1989) also supports this concern. These additive effects should be investigated for both threshold and suprathreshold masking, and for both Forward and Backward masking.

3. It has been assumed that introduced image impairment up to a quantization level of 35-40 is indistinguishable from no impairment, even under conditions where masking is absent. This is assumed because threshold estimates converged at this level in Experiment 1. However, it has also been noted, in Experiment 2, that no unimpaired image sequence was rated more than a grade and a half below its unimpaired counterpart. Though it seems unlikely, it is unclear whether these ratings would continue to drop were the impaired frame to be placed further away from the scene cut. Furthermore, we noted that there was considerable further compression allowable in all three frames on both sides of the scene cut, provided that a moderate video quality standard was acceptable. Whether this further subjective compression has reached an asymptote, or whether it too would decrease further away from the scene cut, is not known.

The Experiments that were done, could be extended outward, away from the scene cut, by a further few frames. This would tell us whether masking, in some capacity was still operating over a longer time course than our threshold results indicated. Additionally, a degraded frame could be placed in within a single video sequence, without any scene cut.. This would give provide a clear measure of how much degradation was acceptable when scene cut masking was unequivocally absent.

4. We have discussed the potential confound between inherent differences in masking due to spatial variance, and actual differences in degree of coding artifact noise caused by differential image demand on the quantization coder. It would be appropriate to attempt to distinguish between these two by controlling for artifact differences and then rerunning part of this study. This could the first step in a more thorough investigation into potential temporal and spatial interactions at scene cuts.

Given that not all difference could be accounted for by variation in image coding difficulty, it would be interesting to examine more systematically the effects that variation in image sequences may have on masking effects. It was noted that the three images differed, both in the amount of high frequency information they contained, and in the amount of image displacement from frame to frame. Both of these factors could be varied systematically to see how masking strength was affected.

Even if all differences between images turned out to be attributable to coding variation, different types of masks, as well as mask and target interactions, might also affect masking strength at scene cuts. This study used a single mask, a panned view of a flower garden. It would be of interest to test different types of masks. Variations that might produce differences in masking strength would be similar to those already mentioned: The amount of image displacement from frame to frame would be appropriate to investigate: A masking sequence with no movement might be less effective as a mask, given that the visual system would not be engaging its smooth

tracking system and so would not be "interrupted". A panning sequence of higher speed might have a greater masking effect as the visual system would have to stop tracking the first image and begin processing a second one¹⁶. Effects of multidirectional, antagonistic movement, like that found in the "Football" sequence used in this study, may have their own unpredictable effects.

Frequency variation within masks may play a role as well. Breitmeyer and Ganz (1976) point out that masks that are more similar to their targets appear to have stronger masking effects.¹⁷ Though the low frequency information left after compression is almost identical to the low frequency information in the following frame,¹⁸ one of the artifacts of the JPEG image compression algorithm the high frequency artifacts that occur at block edges. It is possible that masks with a greater high frequency content may help to mask these artifacts.

Finally, interactions between Targets and Masks could be investigated. If a panned mask has a stronger masking effect than a still one, for reasons stated above, then we can also assume it likely that: a) A scene cut where the target and mask sequence pan in the same direction will produce less masking then when they do not; and that b) by extension, panning sequences whose movements oppose each other will have the strongest masking effect. Masks and targets containing larger quantities of more variegated displacement might, or might not strongly mask one another, as might panned sequences that are perpendicular to one another. Masks which feature spatial displacement within 100 ms of the scene cut can be expected to elicit saccadic eye

¹⁶ This is obviously only relevant for Forward masking.

¹⁷ "Flower", the image sequence used as the mask, had a mid to low frequency image content. Masks of different frequency ranges might have had differential masking effects in relation to the frequency of the Target image sequence used.

¹⁸ Unless there is a large amount of image displacement from frame to frame.

movements which will have an inhibitory effect on the visual system should have a measurable masking effect.

Though the primary focus of this thesis is a practical one, not all of these research suggestions would have immediate practical applicability. Nonetheless, they would provide us with information about how the visual system processes images in a more ecologically valid setting. Masking studies using naturalistic video image sequences can help to bridge the gap between classical masking studies and the broader study of visual processing in its interactive complexity.

Figures

- Figure 1:** SOA & ISI in backward and forward masking
- Figure 2:** Interruption and integration masking
- Figure 3:** A comparison of forward and backward masking
- Figure 4:** JPEG compression coding stages
- Figure 5:** The Discrete Cosine Transformation (DCT)
- Figure 6:** An 8 X 8 matrix, transformed to frequency space
- Figure 7:** A Group of Pictures (GoP) schematic.
- Figure 8:** A video segment example.
- Figure 9:** Double Stimulus, Continuous Quality, subjective rating scale: International Telecommunications Union (ITU) in Recommendation 500 (Sample Response sheet).
- Figure 10:** Experiment 1: Main Effect of Frame
- Figure 11:** Experiment 1: Interaction between Direction and Frame
- Figure 12:** Experiment 1: Main Effect of Image
- Figure 13:** Experiment 1: Interaction between Image and Frame
- Figure 14:** Experiment 1: Main Effect of Day
- Figure 15 a & b:** Experiment 2:
Reference vs. Test x Frame x Quantization
(a: Forward; b: Backward)
- Figure 16:** Experiment 2:
Reference vs. Test x Image (Forward vs. Backward)
- Figure 17 a - f:** Experiment 2:
Reference vs. Test x Image x Quantization
(a: FF1; b: FF2; c: FF3; d: BF1; e: BF2; f: BF3)

FIGURE 1: SOA & ISI in Forward & Backward masking

Comparison of Forward & Backward Masking

Figure 2: Interruption vs. Integration Masking (from Turvey, 1973)

Monotonic v.s. U-shaped Masking Functions / Integration v.s. Interruption Masking

Figure 3: Forward vs. Backward Masking (from Turvey, 1973)

FIGURE 4: JPEG Compression Coding Stages

$$DCT(i,j) = \frac{1}{\sqrt{2N}} C(i) C(j) \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} \text{Pixel}(x,y) \cos \left[\frac{(2x+1)i\pi}{2N} \right] \cos \left[\frac{(2y+1)j\pi}{2N} \right]$$

$$C(x) = \frac{1}{\sqrt{2}} \text{ if } x \text{ is } 0, \text{ else } 1 \text{ if } x > 0$$

The Discrete Cosine Transform

$$\text{Pixel}(x,y) = \frac{1}{\sqrt{2N}} \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} C(i) C(j) DCT(i,j) \cos \left[\frac{(2x+1)i\pi}{2N} \right] \cos \left[\frac{(2y+1)j\pi}{2N} \right]$$

$$C(x) = \frac{1}{\sqrt{2}} \text{ if } x \text{ is } 0, \text{ else } 1 \text{ if } x > 0$$

The Inverse DCT

Figure 5: Discrete Cosine Transform (DCT) & Reverse DCT

Figure 6: 8 x 8 Matrix in Frequency Space

FIGURE 8: A Video Segment Example For A Single Trial (Backward masking)

QT 3
DSCQS

SUBJECT NO. DATE / TIME SESSION / SEAT AGE / SEX CITIZENSHIP DATE ORDER

	41 A B	42 A B	43 A B	44 A B	45 A B	46 A B	47 A B	48 A B	49 A B	50 A B
EXCELLENT	+									
GOOD										
FAIR										
POOR										
BAD										
	+									+

	51 A B	52 A B	53 A B	54 A B	55 A B	56 A B	57 A B	58 A B	59 A B	60 A B
EXCELLENT	+									
GOOD										
FAIR										
POOR										
BAD										
	+									+

Figure 9: ITU-R Recommendation 500 Basic AB Quality Scale (sample page)

NOTE TO USERS

Page(s) not included in the original manuscript are unavailable from the author or university. The manuscript was microfilmed as received.

UMI

Figure 10: Main Effect of Frame

Figure 11: Interaction between Frame and Direction

Threshold values by Image

Figure 12: Main Effect of Image

Threshold Estimates by Frame (by Image)

Figure 13: Interaction between Frame & Image

Figure 14: Main Effect of Day

Figure 15 a: Three-way Interaction: Reference vs. Test x Frame x Quantization

Figure 15 b: Three-way Interaction: Reference vs. Test x Frame x Quantization

Figure 16:
3-Way Interaction: Reference v s. Test x Image sequence x Direction

Reference vs. Test Quality Ratings for Image x Quantization

Figure 17 a - f: Three way interaction: Reference vs. Test x Image x Quantization

References

- Allport, D.A. (1968). Phenomenal simultaneity and the perceptual moment hypothesis. *British Journal of Psychology*, 59-4, 395-406.
- Alpern, M. (1953). Metacontrast. *Journal of the Optical Society of America*, 43, 648-657.
- Aravind, R., Cash, G.L., Duttweiler, D.L., Hang, H., Haskell, B.G. & Puri, A. (1993). Image and video coding standards. *AT&T Technical Journal*, Jan./Feb, 67-89.
- Averbach, E. & Coriell, A.S. (1961). Short term memory in vision. *Bell Systems Technical Journal*, 40, 309-328.
- Barlow, H.B. (1958). Temporal and spatial summation in human vision at different background intensities. *Journal of Physiology*, 141, 337-350.
- Bhaskaran, V. & Konstantinides, K. (1995). *Image and Video Compression Standards: Algorithms and Architectures*. Boston: Kulwer Academic Publishers.
- Bowen, R.W., Pola, J. & Martin, L. (1974). Visual persistence: Effects of flash luminance, duration, and energy. *Vision Research*, 14, 295-303.
- Braddick, O.J. (1980). Low-level and high-level processes in apparent motion. *Philosophical Transactions of the Royal Society of London*, B209, 137-151.
- Breitmeyer, B.G. (1978). Disinhibition in metacontrast masking of vernier acuity targets: sustained channels inhibit transient channels. *Vision Research*, 18, 1401-1405
- Breitmeyer, B.G. (1984). *Visual Masking: An integrative Approach*. New York: Oxford University Press.

- Breitmeyer, B.G. & Ganz, L. (1976). Implications of sustained and transient channels for theories of visual pattern masking, saccadic suppression, and information processing. *Psychological Review*, 83-1, 1-36.
- Bridgeman, B. (1978). Distributed sensory coding applied to simulations of iconic storage and metacontrast. *Annals of Mathematical Biology*, 40, 605-623.
- Bruce, V. & Green, P. (1990). *Visual Perception: Physiology, Psychology, and Ecology*, 2nd edition. Hillsdale, U.S.A.: Lawrence Erlbaum & Associates.
- Burr, D. C. (1980). Motion Smear. *Nature*, 284, 164-165.
- Burr, D. C. (1981). Temporal summation of moving images by the human visual system. *Proceedings of the Royal Society of London*, 211B, 321-339.
- Colheart, M. (1980). Iconic memory and visible persistence. *Perception & Psychophysics*, 27, 183-228.
- DiLollo, V. & Dixon, P. (1988). Two forms of persistence in visual information processing. *Journal of Experimental Psychology: Human Perception and Performance*, 14-4, 671-681.
- DiLollo, V. & Dixon, P. (1992). Inverse duration effects in partial report. *Journal of Experimental Psychology: Human Perception and Performance*, 18, 1089-1100.
- DiLollo, V. (1977). Temporal characteristics of iconic memory. *Nature*, 267, 241-243.
- Dodwell, P.C. (1971). On perceptual clarity. *Psychological Review*, 78-4, 275-289.
- Findlay, J.M. (1980). The visual stimulus for saccadic eye movements in human observers. *Perception*, 9, 7-21.
- Girod, B. (1992). Psychophysical aspects of image communication. *Signal Processing*, 28, 239-251.

- Girod, B. (1989). The information theoretical significance of spatial and temporal masking in video signals. *Proceedings of SPIE: Human Vision, Visual Processing, and Digital Display*, 1077, 178-187.
- Howard, J.P. & Rogers, B.J. (1995). *Binocular Vision and Stereopsis*. Oxford: Oxford University Press.
- Hogben, J.H. & DiLollo, V. (1985). Suppression of visible persistence in apparent motion. *Perception & Psychophysics*, 38(5), 450-460.
- Humphries, G. W. & Bruce, V. (1989). *Visual Cognition..* Hillsdale, U.S.A.: Lawrence Erlbaum associates.
- Kohlers, P.A. (1983). Perception and Presentation. *Annual Psychological Review*, 34, 129-166.
- Lee, J. & Dickinson, B.W. (1994). Temporally adaptive motion interpolation exploiting temporal masking in visual perception. *IEEE Transactions on Image Processing*, 3-5, 513-526.
- Loftus, G.R. & Hanna, A.M. (1989). The phenomenology of spatial integration: data and models. *Cognitive Psychology*, 21, 363-397.
- Loftus, G.R. & Ruthruff, E. (1994). A theory of visual information acquisition and visual memory with special application to intensity-duration trade-offs. *Journal of Experimental Psychology: Human Perception and Performance*, 20, No. 1, 33-49.
- Nelson, M. (1992). *The Data Compression Book.: Lossy Graphics Compression*. SanMateo, California: M & T Books, 347-407.
- Scheerer, E. (1973). Integration, interruption and processing rate in visual backward masking. *Psychol. Forsch.*, Springer-Verlag, 36, 71-93.
- Seyler, A.J. & Budrikis, Z.L. (1965). Detail perception after scene change in television image presentations. *IEEE Transactions on Information Theory*, January, IT-11-1, 31-43.

- Spencer, T.J. & Shuntich, R. (1970). Evidence for an interruption theory of backward masking. *Journal of Experimental Psychology*, 85-2, 198-203.
- Stelmach, L.B., Tam, W.J. & Hearty, P.J. (1991). Static and dynamic spatial resolution in image coding: An investigation of eye movements. *SPIE*, 1453, 147-152
- Tam, W.J., Stelmach, L., Wang, L., Lauzon, D & Gray, P. (1995). Visual masking at scene cuts. *SPIE*, 2411, 111-119.
- Taylor, M.M. & Creelman, C.D. (1967). PEST: Efficient estimates on probability functions. *Journal of the Optical Society of America*, 41, 782-787.
- Townsend, V. M. (1973). Loss of spatial and identity information following a tachistoscopic exposure. *Journal of Experimental Psychology*, 98-1, 113-118.
- Turvey, M.T. (1973). On peripheral and central processes in vision: Inferences from an information-processing analysis of masking with patterned stimuli. *Psychological Review*, 80, 1-52.
- Von Grunau, M. W. (1978). Interaction between sustained and transient channels: Form inhibits motion in the human visual system. *Vision Res.* , 18, 197-201.
- Wandell, B.A. (1995). *Foundations of Vision*. Sunderland, Massachusetts: Sinauer Associates, Inc.
- Wang, L. (1995). Rate control for MPEG video coding. *SPIE, Proceedings*, 2501, 53-64.
- Zeki, S. (1990). The motion Pathways of the visual cortex. In *Vision: Coding and Efficiency*, Colin Blakemore, ed. Cambridge: Cambridge University Press, 321-345.B

IMAGE EVALUATION TEST TARGET (QA-3)

APPLIED IMAGE, Inc.
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989

© 1993, Applied Image, Inc., All Rights Reserved