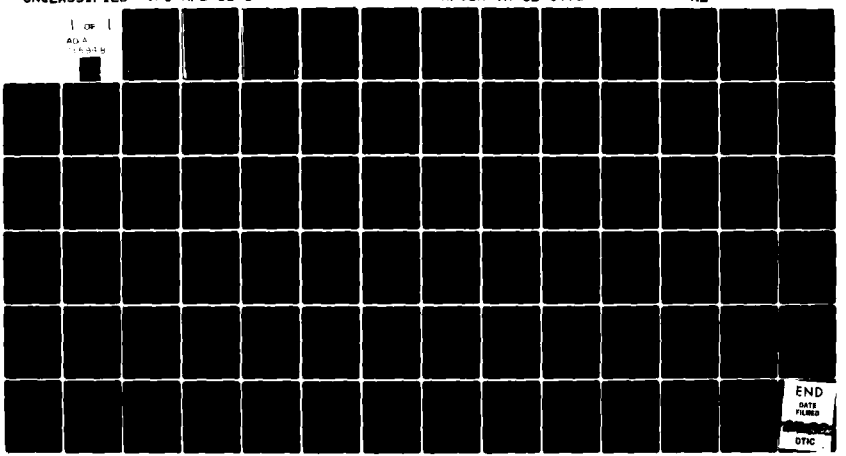
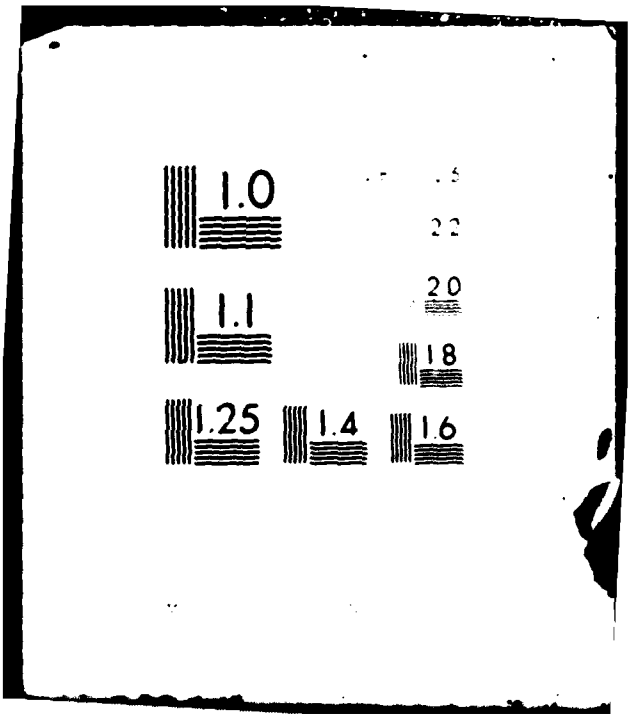
AD-A115 948 VIRGINIA POLYTECHNIC INST AND STATE UNIV BLACKSBURG --ETC F/G 14/5
QUALITY METRICS OF DIGITALLY DERIVED IMAGERY AND THEIR RELATION--ETC(U)
FEB 82 H L SNYDER, D I SHEDIVY, M E MADDOX F49620-80-C-0057
UNCLASSIFIED VPI-HFL-81-3 AFOSR-TR-82-0475 NL




1.0
1.1
1.25
1.4
1.6
2.2
2.0
1.8

AFOSR-TR- 82-0475

(13)

HFL-81-3

Quality Metrics of Digitally Derived Imagery and Their Relation to Interpreter Performance: III. Subjective Scaling of Hard-Copy Digital Imagery

FINAL TECHNICAL REPORT

Harry L. Snyder, Ph.D.
David I. Shedivy, M.S.
Michael E. Maddox, Ph.D.

Department of Industrial Engineering
and Operations Research
Virginia Polytechnic Institute and State University
Blacksburg, Virginia 24061

Contract F49620-80-C-0057
Life Sciences Directorate
U.S. Air Force Office of Scientific Research
Bolling Air Force Base, D.C. 20332

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1 REPORT NUMBER **AFOSR-TR- 82-0475** 2 GOVT ACCESSION NO AD-A115 948	3 RECIPIENT'S CATALOG NUMBER
4 TITLE (and Subtitle) QUALITY METRICS OF DIGITALLY DERIVED IMAGERY AND THEIR RELATION TO INTERPRETER PERFORMANCE: III. SUBJECTIVE SCALING OF HARD-COPY DIGITAL IMAGERY.	5 TYPE OF REPORT & PERIOD COVERED Final Report 1 Oct. 79 - 30 Sept. 80
	6 PERFORMING ORG. REPORT NUMBER HFL 81-3
7 AUTHOR(s) Harry L. Snyder, Ph.D. David I. Shedivy, M.S. Michael E. Maddox, Ph.D.	8 CONTRACT OR GRANT NUMBER(s) F49620-80-C-0057
9 PERFORMING ORGANIZATION NAME AND ADDRESS Dept. of Industrial Engineering and Operations Res. Virginia Polytechnic Institute and State University Blacksburg, Virginia 24061	10 PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 61102F 2313/A4
11 CONTROLLING OFFICE NAME AND ADDRESS Air Force Office of Scientific Research Bolling Air Force Base, D.C. 20332	12 REPORT DATE February, 1982
	13 NUMBER OF PAGES 72
14 MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)	15 SECURITY CLASS. (of this report) Unclassified
	15a DECLASSIFICATION/DOWNGRADING SCHEDULE

16 DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17 DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18 SUPPLEMENTARY NOTES

19 KEY WORDS (Continue on reverse side if necessary and identify by block number)

Digital imagery
Human factors
Image quality
Image interpretation
Image processing

20 ABSTRACT (Continue on reverse side if necessary and identify by block number)

Hard-copy digital imagery was studied with respect to subjective image quality. Trained Air Force photointerpreters judged the interpretability of 250 military scenes. The scenes varied in noise, blur, and scene content.

The results showed that noise, blur, and scene content produce differential perceptions of interpretability. Many interactions were significant.

(cont.)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Other analyses performed on the data showed that at least 62 categories should be used to scale interpretability, the correlation between information extraction performance and scale values for digital imagery is high, and multidimensional scaling can be used with limited utility in studying image quality.

In general, from a review of the literature, digital imagery did not appear to be greatly different from standard analog imagery in terms of subjective quality or interpretability.

HFL-81-3

Quality Metrics of Digitally Derived Imagery and Their Relation to Interpreter Performance: III. Subjective Scaling of Hard-Copy Digital Imagery

FINAL TECHNICAL REPORT

Harry L. Snyder, Ph.D.
David I. Shedivy, M.S.
Michael E. Maddox, Ph.D.

Department of Industrial Engineering
and Operations Research
Virginia Polytechnic Institute and State University
Blacksburg, Virginia 24061

Contract F49620-80 C-0057
Life Sciences Directorate
U.S. Air Force Office of Scientific Research
Bolling Air Force Base, D C 20332

PREFACE

This research was sponsored by Contract No. F49620-80-C-0057, from the Air Force Office of Scientific Research, under the monitorship of Dr. Alfred R. Fregly. Dr. Harry L. Snyder is the principal investigator.

The imagery used in this research was supplied by the Environmental Research Institute of Michigan. Dr. James J. Burke and his staff at the University of Arizona (UA), in cooperation with Mr. Ray Schmidt of the Image Processing Institute, University of Southern California, and Mr. H. D. Fisher of UA, degraded and digitized the original negative transparencies. Mr. Gilbert Kuperman, AFAMRL, provided technical advice throughout the effort. The staff and photointerpreters of the 548th Reconnaissance Technical Group graciously and professionally supported the data collection effort.

CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

I. INTRODUCTION

Recent technological developments have resulted in a wide variety of imaging systems and subsystems. The flexibility and technologies available to the designer include various means for collection, coding, transmitting, decoding, analog and digital processing, and analog and digital display. The applications of such systems and subsystems are myriad, ranging from static and dynamic military photointerpretive functions, through commercial and closed-circuit television and facsimile systems, to diagnostic radiological instrumentation and earth resources applications. The scientific world is quite familiar with some of the techniques which can be used to "improve" the nature of any such image, and the non-scientific world has equally seen examples of such processing effectiveness, such as the Zapruder and Hughes films of the Kennedy assassination. In many cases, it is clear that such processing and display techniques can extract information in the original image which is otherwise well below the threshold capacity of the human visual system, whereas in other cases it is clear that processing techniques can often serve either to hide existing, and important, image detail or to "create" image detail which is perhaps not present in the original image or

in the "real world". Heretofore, most of these areas of image system and subsystem development have plainly suffered from their inattention to human observer requirements. This is particularly true of the extensive effort in digital image processing, especially that part devoted to the improvement ("enhancement", "restoration") of images for purposes of human information extraction. In nearly all of the work performed in laboratories around the country that are pursuing this type of research, the necessary evaluative efforts to determine the utility of processing and display techniques have not been conducted. Rather, reports and publications of this work typically take the form of "before and after" pairs of images, where the reader is left to estimate the utility of such images either by visual inspection of these published (second- or third-generation) photographs or by the subjective opinions offered in the text by the author.

Because the intent of such image processing techniques is to improve the information extraction capabilities of the human observer, it is clearly appropriate and mandatory that evaluative techniques include objective measurement of human information extraction from such images, in addition to subjective estimates of the overall quality or utility of the image. Unfortunately, the human factors experiments required to produce quantitative and objective assessment of image quality have rarely been conducted in image processing

laboratories or in conjunction with image processing programs.

In view of the many millions of dollars being devoted to image collection, processing, and display systems for the military and civilian use of digitized images, it is quite clear that an assessment program is urgently needed to devise procedures, techniques, and metrics of digital image quality. Such a program requires the establishment of a standardized set of procedures for obtaining human observer information extraction performance; relating this performance, in a quantitative manner, to the various collection, processing, and display techniques and algorithms; and devising a quantitative relationship for the multi-dimensional scaling of the various collection, processing, and display techniques in "performance space".

Only through such an integrated program of research can the system and subsystem designer have meaningful data for cost-benefit analyses of future system development, be such systems intended either for military or for non-military applications. The image collection, processing, and display technology is now at a point whereby such evaluative research is sorely needed. Fortunately, microphotometric, microdensitometric, and human performance measurement techniques have been evolved during the past several years to relate human information extraction performance to the various physical characteristics of both electro-optical and

photographic image displays. The present research program is designed to extend these recently developed techniques into the arena of digital images, emphasizing derivation of metrics of image quality appropriate to digitized images, and providing quantitative cost-benefit data which will permit the designer and system developer to plan his developmental effort as well as to specify optimum system components for particular image acquisition and display requirements.

OVERVIEW OF THE RESEARCH PLAN

The research plan is laid out schematically in Figure 1. Each small, solid-lined box, with the exception of the uppermost, indicates a separate task to be conducted during the course of the four-year effort. The two large, broken-lined boxes delineate the specific display formats that will be studied and compared during this initial program: black and white hard-copy transparencies and electronic displays. The small, broken-lined box at the bottom illustrates important extensions of this research to be pursued in the future, namely interactive digital displays in both black and white and full color. The present report describes in detail the hard-copy subjective scaling experiment.

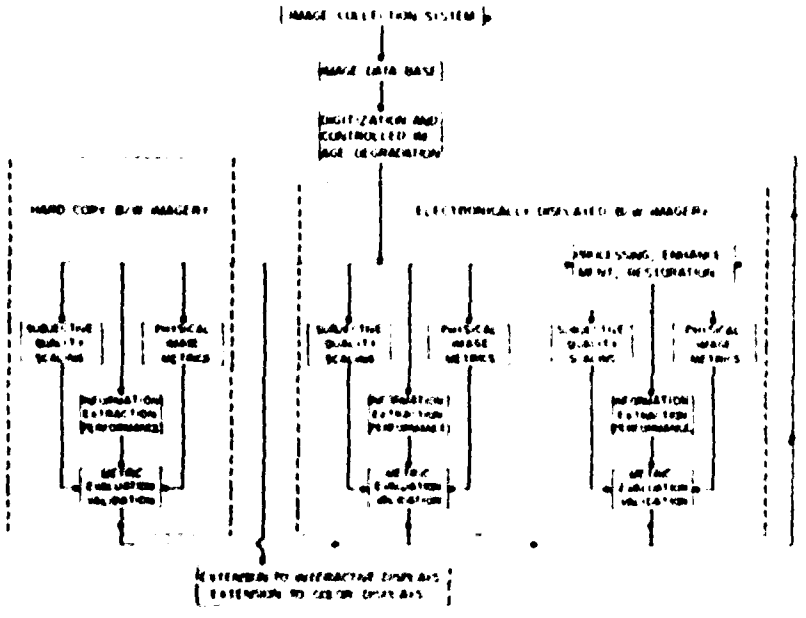


Figure 1: Schematic diagram of proposed research

RESEARCH OBJECTIVES

The overall research objectives of this program are as follows:

1. Develop standardized procedures and techniques to evaluate hard-copy (film) and soft-copy (CRT) digital image quality.
2. Compare candidate physical metrics of image quality.
3. Compare hard-copy with soft-copy displays for image interpretation.
4. Evaluate candidate processing, enhancement, and restoration algorithms for improvement of image interpretation on soft-copy displays.

SPECIFIC RESEARCH TASKS

In keeping with the general goals described above, the specific research tasks are as follows:

1. Develop an imagery database and image interpretation scenarios from high quality aerial photography relevant to the image interpretation task.
2. Select and purchase display and interface hardware to present the image database on soft-copy (CRT) displays.
3. Develop image manipulation software for soft-copy and hard-copy experiments.
4. Develop and standardize observer data collection procedures for hard-copy and soft-copy experiments.
5. Develop and standardize procedures for obtaining physical image metrics from hard-copy and soft-copy displays.
6. Digitize and degrade database imagery and record images on hard-copy and magnetic tapes for soft-copy display.
7. Obtain physical image metric data for hard-copy and soft-copy displays.
8. Conduct subjective quality scaling and information extraction studies on hard-copy images.
9. Conduct subjective scaling and information extraction studies on soft-copy displays.

10. Evaluate the utility of image quality metrics for both hard-copy and soft-copy imagery.
11. Conduct subjective scaling and information extraction studies on processed soft-copy imagery.
12. Analyze the utility of image quality metrics for processed soft-copy imagery.
13. Compare image quality metrics for hard-copy and soft-copy (processed and nonprocessed) images. Relate these results to concepts and models of human visual performance and to imaging system design variables.

This present report relates to Objective 8 above. It describes the results of that part of program dealing with subjective scaling of the hard-copy imagery. It also addresses the question of how subjective image quality is affected by measurable physical properties of digitally derived imagery. Specifically, trained photointerpreters performed a subjective scaling task using images which were degraded by two known physical characteristics common to digitized aerial imagery, blur and noise. A parallel experiment assessing information extraction performance with the same images is reported by Snyder, Turpin, and Maddox (1981).

In addition to obtaining these important baseline scaling data, the experiment also served to evaluate the scaling methodology to be used in the subsequent soft-copy phases of

the research program. Objectives of this methodology are described later.

BACKGROUND

Often, PIs are not able to inspect in detail all images brought to them. Hence, the photographs that are used are those that possess high content and technical quality. Even in this age of high speed computer technology, the calculable physical metrics are not completely reliable in determining which photographs will provide the most information. Thus, PIs typically make a cursory inspection of imagery to determine which frames they feel are most interpretable. PIs often use a standard scale (e.g., NATO scale) as a reference for selection/rejection decisions, although the literature contains other scales and experiments that purport to measure subjective image quality, and some relate noise and blur content to subjective image quality for nondigital imagery. Some of this more pertinent literature is summarized below.

For the most part, this present research is unique in its use of digital images. Few laboratories have the facilities and personnel to play an active role in the generation of digital images and evaluate the effects of degrading them. Perhaps the most advanced research in this area is conducted by the government laboratories, although few studies of government PIs have been reported in the literature.

Noise Effects

For example, several studies have investigated the effects of random noise upon subjective assessments of image quality (Allnatt and Prosser, 1966; Below, Huertas-Sendra, Fritze, and Samrau, 1963; Geddes, 1963; Newell and Geddes, 1963; Prosser and Allnatt, 1965; Weaver, 1959). All of these studies involved raster-scan television images. The primary aim of these studies was to develop standards for commercial television broadcasting. Without exception, descriptive statistics were the only results reported. Variations, other than noise levels, in the experimental conditions across these studies include viewing ratio, the type of observers, the number of raster lines (405-625), average luminance, ambient illuminance, color versus black and white presentation, noise frequency and waveform, and the subjective scale used for rating the images. The type of scales used by these investigators is called a grading scale (Prosser, Allnatt, and Lewis, 1964). The scales consist of points (typically 1-7) with descriptive adjectives corresponding to each point. Some of the scales measure the amount of image impairment while others measure the subjective quality of the image. In these studies, each observer assigned a number to an image, a technique known as direct magnitude estimation (Stevens, 1975).

The authors' conclusions from these studies are consistent: noise degrades image quality. It is

infeasible to eliminate all noise from a television image. However, in developing television standards, mark points have been reported. A mark point refers to the noise level that intersects a particular subjective score. These points can be determined most easily by graphic methods. For a scale with five points (bad, poor, fair, good, excellent, one-five, respectively), a noise level is associated with each scale category. Therefore, a standard might be based on the noise level (or signal-to-noise ratio) yielding a rating of four or good. In general, the middle scale category (average quality, good to poor, marginal, obviously impaired but not objectionable) has been associated with a range of signal-to-noise ratios (in dB) of 23 to 35 or an average of about 30. A signal-to-noise ratio of 30 dB is equivalent to a signal-to-noise ratio of 32:1.

While these studies of the effects of noise provide some data for comparison with the present research, the problems inherent in doing so should be mentioned. Television displayed images are quite different from digitized photographs. It would be foolhardy to expect great similarity in the results. Further, in all of these studies, signal-to-noise ratios were calculated at the input to the video system. The actual viewed signal-to-noise ratios should be measured at the display by photometric means.

Rosell and Willson (1973) have done much to forward the study of noise and its role in target detection. Their experiments have involved many aspects of visual performance. Much of the preliminary psychophysical work by these authors involved the detection of displayed stripes (called bar patterns) on photographs mixed with noise and displayed on a television monitor. After determining threshold signal-to-noise levels in bar pattern detection tasks, experiments were conducted in recognizing and identifying military targets. These threshold data were then compared with the bar pattern studies and it was found that the results of the bar studies could predict recognition/identification performance, and vice versa. By applying a scale factor of eight for bar pattern detection data and matching bar area to target area, the authors reported a high degree of similarity in the data. No statistics were reported.

Finally, Humes and Bauerschmidt (1968) took a different approach to the study of noise degraded imagery. Using the method of limits and five standard signal-to-noise ratios from 2 to 29.1 dB, their judges reported whether or not the test images were equal to or more or less "noisy" than the standards. The 29 images used in their study had previously been used in a target recognition study in which the standard signal-to-noise ratios were effective in showing a performance difference. The psychophysical study found an

uncertainty range of 1 dB at the 2 dB standard and about 2.5 dB at the 29.1 dB standard. In particular, this relationship between range of uncertainty and the standard signal-to-noise ratio was linear above the 7 dB standard.

Blur Effects

Blur has not been studied as extensively as noise with respect to subjective scaling. In fact, one must draw upon related research in order to gain any insight into the effects of blur. Film degraded by blur is generally discarded immediately, unless new pictures cannot be obtained.

Blur in raster-scan images is similar to echo or delay, which is manifested in "ghost" images about the desired signal. Several studies have addressed this type of impairment (Allnatt and Prosser, 1965; Cavanaugh and Lessman, 1971; Lessman, 1972; Weave 1968). These studies are very similar to those of noise-impaired television images. Whereas noise is characterized by frequency and amplitude, echo can be defined in terms of the delay (spread in time/space about the signal) and amplitude. Echo impairment is typically expressed as a signal-to-echo ratio, the amplitude of the signal to the amplitude of the displaced signal at a particular point before or after the desired signal. Allnatt and Prosser (1965) and others have shown quite convincingly that a delay of two microseconds is most objectionable subjectively.

The results of studies in this area are in general agreement. The middle scale value is associated with a range of signal-to-echo ratios (in dB) from 10 to 25 with an average of about 20 dB. A signal-to-echo ratio of 10:1 is equivalent to 20 dB. These numbers provide another means of comparing an analog mode of presentation with the present data.

Subjective Scaling and Information Extraction

Several studies of photo-interpretation by PIs have been published, primarily in the technical report literature. A few of these studies have investigated the relationship between scale values and information extraction performance (Brainard, Sadacca, Lopez, and Ornstein, 1966; Klingberg, Elworth, and Filleau, 1970; Sadacca and Schwartz, 1963). (Information extraction performance data were available for a subset of the images used in the present research. Performance data allowed for an evaluation of the relationship between performance and scale values in this study.)

Sadacca and Schwartz (1963) used a ranking technique to scale 72 images. Several performance measures were also collected, including the number of correct target identifications, the number of wrong target identifications, and an overall accuracy score. The ranks were correlated with performance for each of three scenes and the three performance measures. The correlations between the ranks

and the number of correct identifications ranged from .39 to .59, between ranks and incorrect identifications from -.30 to .00, and between ranks and overall accuracy from .14 to .61, for the three scenes. It should be noted that different PIs were used in the performance and scaling studies.

Brainard et al. (1966) had PIs make relative judgements in scaling image quality. Each of 36 images (3 scenes) was rated by comparing each image with a catalog of 36 degraded images and assigning each test image a catalog image number. Two performance measures were collected: number of correct target identifications and number of correct identifications of target area. The same PIs participated in both phases of the experiment. The correlations between catalog numbers and target identification performance ranged from .59 to .70, and for area identification from .56 to .79.

Klingberg et al. (1970) used a standard ranking procedure to scale 32 images. A target identification measure was again collected. Subjective rank correlated very highly, .92, with the performance measure.

It is obviously the case that PIs were able to predict the interpretability of images through a variety of scaling procedures in the studies reviewed here. The extent of the predictability was not consistent.

Multidimensional Scaling

Two technical reports have been published in which multidimensional scaling (MDS) was applied to subjective image quality (Marmolin and Nyberg, 1978; Sadacca and Schwartz, 1963). Sadacca and Schwartz (1963) used 12 images varying in scene content, ground scale, sharpness, and contrast. Twenty-one PIs judged the similarity, based on interpretability of all pairs of images (66). Another group of PIs performed in an information extraction study with this imagery database. The authors chose six dimensions for their MDS spatial configuration but could only explain four of the six. The four dimensions were interpreted by averaging the projections over the levels of the imagery parameters and relating the direction of these means to the logical subjective effect of the parameters. The means indicated the following: dimension one was related to ground scale, dimension two was related to sharpness, dimension three to scene content, and dimension four to contrast. No other analysis was attempted using the projections.

Marmolin and Nyberg (1978) took a similar approach using 24 degraded images and four untrained judges. Physical quality metrics were calculated for each image. Four dimensions were chosen to represent the data. While six dimensions produced minimum stress, the authors chose the four-dimension model because these dimensions could be interpreted. The four dimensions were interpreted from

correlations obtained between the image parameters (including physical metrics) and the projections on each dimension. The four dimensions were interpreted as sharpness, noise, contrast, and the bandpass by contrast by noise interaction, respectively. Since no information extraction performance measures were collected, no analyses relating the projections to performance were conducted.

These MDS analyses can be compared with the present analysis. To the extent that the data collected in this research may differ from other MDS analyses of subjective image quality, a basis for comparing digital versus analog photographs presents itself. While a well-mapped subjective quality space may serve as a device for screening the interpretability of photographs in the future, this will be useful only if a spatial configuration can be located that is highly related to performance (i.e., projections must correlate with performance). This consideration has not been addressed in past research.

Rating Scales

Almost every rating scale used by experimenters in scaling studies has a different number of categories. The choice as to the number of categories to include in a scale is generally arbitrary, depending on the resolution desired in the obtained response. Why the existing NATO scale and perhaps other similar scales used by PIs to rate the interpretability of photographs has 10 categories is

unknown. There are no data to indicate this number to be optimal in rating ground resolved distance (GRD) or image quality in general.

Without justification, Guilford (1954) suggested that about 20 categories are optimal in general uses of rating scales. The optimal number of categories is dependent on the number of stimulus categories (and stimulus variation) to be scaled (Erikson and Hake, 1955). The basic issue is the number of absolute judgements that can be made in a particular stimulus dimension. In this research, unlike basic research in auditory or visual perception, the stimulus property to be scaled (e.g., frequency, luminance, chrominance) is not easily or objectively measureable. In perception research it has been shown that about seven categories are sufficient for a variety of stimulus dimensions (Miller, 1956). However, Muller, Sidorsky, Slivinske, Alluisi, and Fitts (1955) have shown that many more categories (24) can be used efficiently when the judges are allowed to practice the task. Therefore, it would seem reasonable to conclude that a scale with more than 10 categories would better provide the response resolution needed by a well-practiced PI. For these reasons, and because the subjective scaling technique to be used in subsequent studies in this program should use the same scale (for comparability of results), careful attention was given to the scale selection.

II. SELECTION OF THE SCALING PROCEDURE

A large part of this research involved the determination of the most efficient scaling procedure, given the constraints. Because of the cost involved in generating the imagery and collecting the data, considerable time and effort were spent in evaluating alternative scaling approaches. It was known during this period of evaluation that 15 PIs would be available for about four hours each. It was also known that an imagery database consisting of 250 images would be available. Because of the professional attitude held by most PIs, the experimental task would have to be parsimonious and, in general, similar to the typical practices of the PI. Consequently, an unclassified NATO scale was found that would serve as a reference for scaling image quality (Appendix A). The NATO scale was optimal in that it possessed a high degree of similarity to the scales used operationally by PIs. It is also presumed to relate directly to GRD. The only remaining problem, and perhaps the most difficult to resolve, was to decide how to use the scale to collect the most meaningful data without creating an operationally meaningless task for the PIs.

THE ALTERNATIVES

Two broad classes of subjective scaling methods exist: direct and indirect. Each class encompasses numerous, but varied, approaches. The major difference between the two classes is the specificity of the obtained response information. When the response given by the judge (or PI) provides a direct quantization (numerical representation) of the subjective effect of the stimulus, then the procedure can be said to be direct. Direct methods of scaling can result in interval or ratio scale values; judges simply report the scale values. Types of direct methods include ratio estimation, magnitude estimation, fractionation, and cross-modality matching. On the other hand, the indirect methods of scaling result in ordinal scale values that can be transformed to an interval or ratio scale. The indirect methods are typified by the pair comparison approach, where all of the experimental stimuli are paired and compared on the basis of some attribute. Other indirect methods include the methods of triads and rank ordering. In most cases, both direct and indirect methods will produce satisfactory scale values; however, experimental constraints often dictate the most feasible approach to take.

THE MOST FEASIBLE ALTERNATIVE

The direct methods of scaling are very easy to apply. A seemingly acceptable task entailed each PI making 250 ratio estimates in four hours. As PIs, in their daily work, use an internalized concept of GRD to draw inferences from imagery, the NATO scale seemed to be an ideal instrument to use in the study of the subjective interpretability of images. The NATO scale, in conjunction with ratio estimation, satisfied the needs of this research (e.g., a task similar to that of the PIs' daily practices, a scaling method that would provide scale values for subsequent MDS analysis, and a method that could be used in light of the constraints on time and the number of available PIs).

The indirect methods, particularly pair comparisons, are perhaps more reliable than the direct methods because of the procedure involved (relative judgements), but require a large amount of time in data collection. Though there are special pair comparison techniques for reducing the number of pairs to be judged, with limitations on the number of judges available, pair comparisons could not be used in this study without reducing the range of experimental parameters currently existing in the imagery database.

Consequently, ratio estimation using the NATO scale as the reference appeared to be the most reliable and time efficient means of collecting subjective data.

III. PURPOSE OF THIS EXPERIMENT

The purpose of this experiment was to investigate the psychophysically scaled effects of blur and noise on digital image quality. To maintain an experimental environment in which PIs are most accustomed to working, the NATO scale was chosen as a reference for rating image quality. Because of the inclusion of the NATO scale, issues related to the use of scales were also studied.

The specific objectives/hypotheses of this research are as follows: (1) digital images degraded by blur and/or noise appear less interpretable as the degree of degradation increases; (2) the optimal number of categories in the NATO scale is greater than 10; (3) an MDS analysis can be used to map the subjective, spatial dimensionality of the imagery database and predict information extraction performance; (4) the correlation between information extraction performance and ratings of apparent interpretability is high for trained PIs, and (5) the scaling technique used will produce data sufficiently consistent and meaningful such that the same scaling technique can be used in the subsequent soft-copy studies in this program.

IV. METHOD

PHOTOINTERPRETERS

The PIs used in this experiment were 14 NATO-scale trained military PIs who were stationed at Hickam Air Force Base, Hawaii. One of fifteen PIs scheduled to participate in the study declined because of the importance and urgency of her regular work. The average age of the PIs was about 25 years and the average experience level about four years. The study was conducted at Hickam Air Force Base during normal working hours. The PIs were not paid extra for their participation.

APPARATUS

A Perkin-Elmer microdensitometer was used to digitize 10 standard photographs consisting of three orders of battle: air, electronic, and sea (i.e., four airfield scenes, two scenes of typical research and development installations, and four quay and shipyard scenes). The images were digitized in a 4096 x 4096 picture element format. The complete set of digitized images was planned to represent all combinations of five levels of noise, 10, 20, 40, 80, and 160 digital units (signal-to-noise ratios of 200, 100, 50, 25, and 12.5), five levels of blur, 20, 40, 80, 160, and

320 micrometers (μm), and 10 scenes. Blur was produced in software by multiplying the frequency spectrum of each digitized image by an appropriate Gaussian filter function. A Fast Fourier Transform yielded the frequency spectra. The image matrices were then reconstructed by the inverse Fourier transform. As blur was added to each image, high frequency detail was removed. Noise was added to the images by multiplying each picture element by a value randomly selected from an appropriate Gaussian random noise function. The scenes and amounts of degradation were scrutinized and, in part, chosen by a senior PI who provided the scores for the information extraction study. For a more thorough description of the imagery database and its development, see Burke and Strickland (1982). For reasons given in that report, the final hardcopy SNR levels were 75, 60, 42, 24, and 12, while the final blur levels were 40, 52, 84, 162, and 322 micrometers.

The 250 images were shown to the PIs in the form of positive transparencies, 7.6 x 7.6 cm. A light table (Richards Model 33H100) with binocular zoom stereo optics and hand held tube magnifiers was available to all PIs during data collection.

As mentioned, the existing NATO scale that was used to scale the transparencies is based on interpretability and GRD. Each increasing whole number, 0 to 9, represents a 50% reduction of GRD, beginning with "useless for

interpretation", scale score 0, greater than 9 cm, scale score 1, and ending with less than 10 cm for a scale score of 9. (See Appendix A for the scale.) In this experiment, the PIs were asked to report not only a whole number scale value but also a decimal. In other words, because each whole number represents a range of interpretability and MGRD, the PIs had to interpolate to the nearest 0.1 over the range to report the decimal portion of their scale value. Thus, the existing 10-point NATO scale was transformed to a 100-point scale.

PROCEDURE

Each PI was allowed approximately four hours to scale all 250 transparencies. Rest periods were allowed as needed. The 250 transparencies were administered to each PI in a different random order. The randomization scheme was obtained using the Statistical Analysis System (SAS) Plan Procedure (Barr, Goodnight, Sall, and Helwig, 1976). The scale values were reported verbally by the PI and recorded manually by the experimenter. The instructions were administered to each PI as follows:

> This experiment will involve your rating numerous transparencies on the basis of interpretability. It is assumed that you have had some experience with rating scales; if you have not, you should inform the experimenter at this time.
>
> The rating scale that will be used here is a 0-9 imagery interpretability rating scale. This scale is probably similar to those you are familiar with. Please look over the scale (attached) at this time. As you can see, larger

scale values represent a greater interpretability. The range of some interpretation capabilities that fall into that range are given under each Rating Category, 0-9. Your task will be to rate each transparency using this scale.

However, in an attempt to obtain more information from your ratings, we would like you to report your ratings to the nearest tenth. That is, your ratings should take the form 3.9, 7.2, or 5.0, rather than simply 0, 1, 2, 3, 4, 5, 6, 7, 8, or 9. It may be a good strategy to determine the range in which a particular transparency falls (say between 6 and 7) and then try to refine your rating. Please remember to be explicit in reporting both portions of your rating, a whole number and a decimal. We hope to show that PIs can rate transparencies with more resolution than conventional scales afford. If you have any questions regarding the scale or how we would like you to use it, please ask the experimenter at this time.

In this phase of the experiment, you will see 250 transparencies. The 250 transparencies represent 10 scenes that have differing amounts of noise and blur. Your task is to rate each transparency using the scale we have just discussed. The scale will be available for use during the experiment, and you may use any equipment you feel would be helpful. However, you will only have about 4 hours to rate the entire set of 250 transparencies; therefore, you should spend about 1 minute on each. The experimenter will be in the room with you while you rate the transparencies. He will hand you each transparency, and after you arrive at a rating, please hand back the transparency and verbally report your rating.

Before we begin, do you have any questions regarding this phase of the experiment?

The instructions given to and the procedure carried out by the experimenter were consistent with the above instructions.

Prior to participation in this experiment, each PI was asked to read and sign an informed consent form to insure that the rights of the participant were known and upheld.

V. RESULTS

ANALYSIS OF VARIANCE (ANOVA) ON RATINGS

The overall ANOVA of the scaling data includes the three parameters of the imagery database, noise, blur, and scenes, in addition to order of battle and PI. As indicated earlier, the imagery database consists of scenes that were nested in various orders of battle. The summary table for the overall ANOVA indicates that almost every source of variation, including interactions, was significant (Table 1).

Another factor was separately analyzed but does not appear in the summary table. Two of the fourteen PIs who participated in the study were actually Army PIs while the other twelve were in the Air Force. Therefore, the main effect of Service and the Service x Scene interaction were analyzed. Both effects were non-significant ($F_{1,12} = .26$, $p = .6184$ and $F_{7,48} = 1.93$, $p = 0.75$, respectively) and deleted from the overall analysis.

TABLE 1

Summary of Analysis of Variance on NATO Scale Scores

SOURCE	df	MS	F	p
Blur (B)	4	832.01	105.63	<.0001
Noise (N)	4	104.20	82.59	<.0001
Order of Battle (OB)	2	57.36	16.01	<.0001
Photointerpreter (PI)	13	259.24		
Scene/OB (S/OB)	7	28.80	17.49	<.0001
B x N	16	5.96	14.06	<.0001
B x S/OB	28	2.30	5.14	<.0001
B x PI	52	7.88		
B x OB	8	5.40	8.03	<.0001
N x S/OB	28	.57	1.40	.0882
PIx N	52	1.26		
N x OB	8	1.41	3.09	.0036
PIx S/OB	91	1.65		
PIx OB	26	3.39		
B x N x S/OB	112	.42	1.26	.0378
B x N x PI	208	.42		
B x N x OB	32	.41	1.09	.3349
B x PIx S/OB	364	.45		
B x PIx OB	104	.67		
PIx N x S/OB	364	.41		
PIx N x OB	104	.46		
B x PIx N x S/OB	1456	.34		
B x PIx N x OB	416	.37		
Total	(3499)			

Blur

Increasing blur results in decreasing scale values, as plotted in Figure 2. A Newman-Keuls multiple comparisons test (MCT) showed that all comparisons were significant ($p < .05$) except between blurs of 40 and 52 m and between 52 and 84 m. The trend was monotonically decreasing with increasing blur degradation. The linearity of this effect is indicated by a linear correlation of $r = .975$, $p < .0001$.

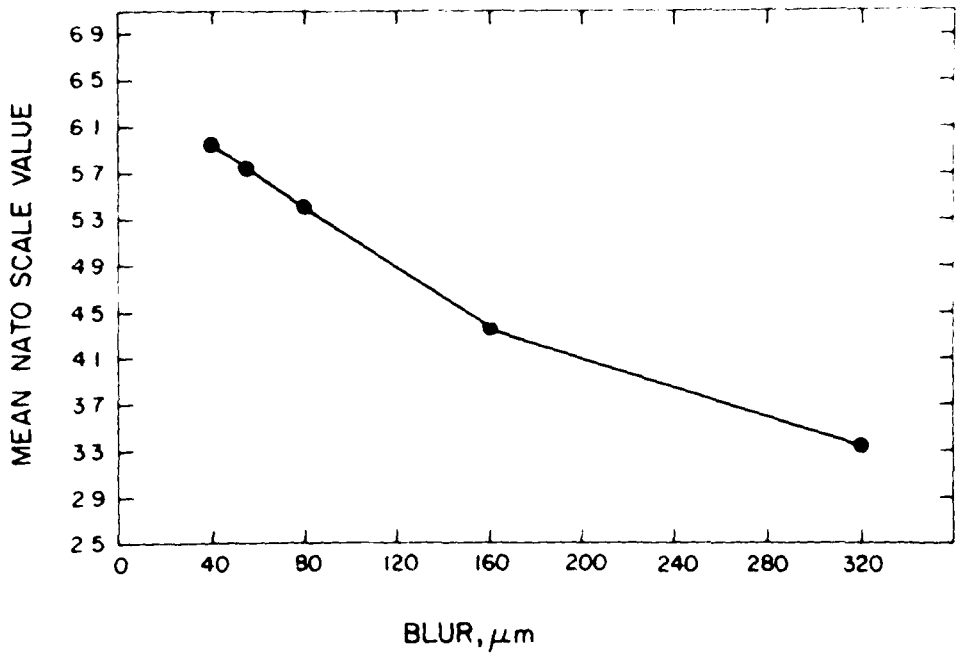


Figure 2: The effect of Blur on NATO scale value

Noise

The main effect of Noise was also highly significant, as shown in Figure 3. Increases in signal-to-noise ratio generally increased scale values. The MCT showed that all comparisons were significant except between signal-to-noise ratios 75 and 60. This trend was also monotonically decreasing and highly linear with degradation, $r = .946$, $p < .0001$.

Order of Battle

The main effect of Order of Battle (OB) is illustrated in Figure 4. The MCT showed that both air and sea orders of battle were rated more interpretable than electronic scenes ($p < .05$), although air and sea OBs did not differ from one another ($p > .05$).

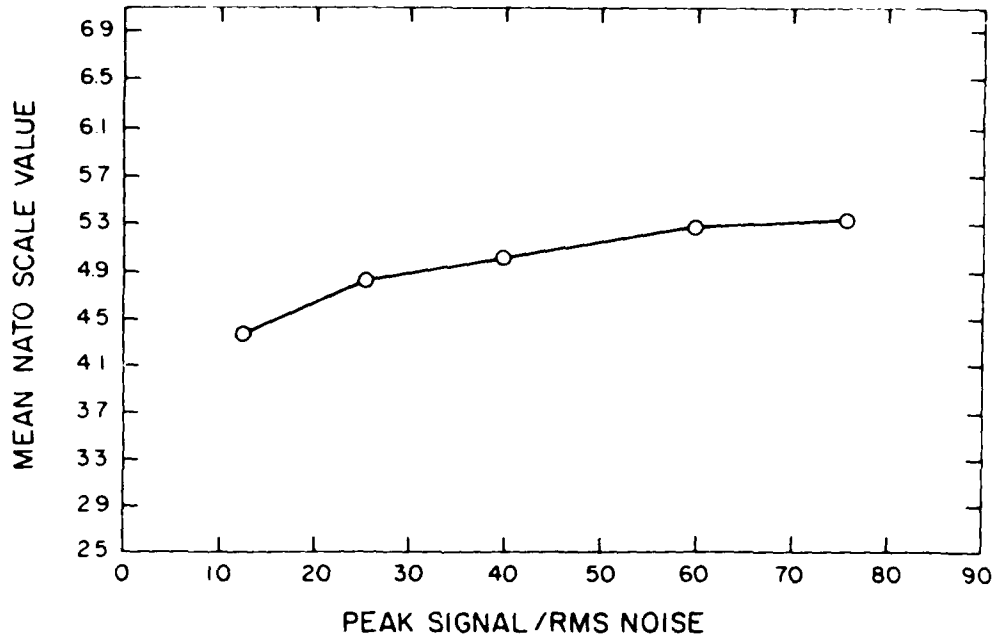


Figure 3: The effect of Signal-to-Noise ratio on NATO scale value

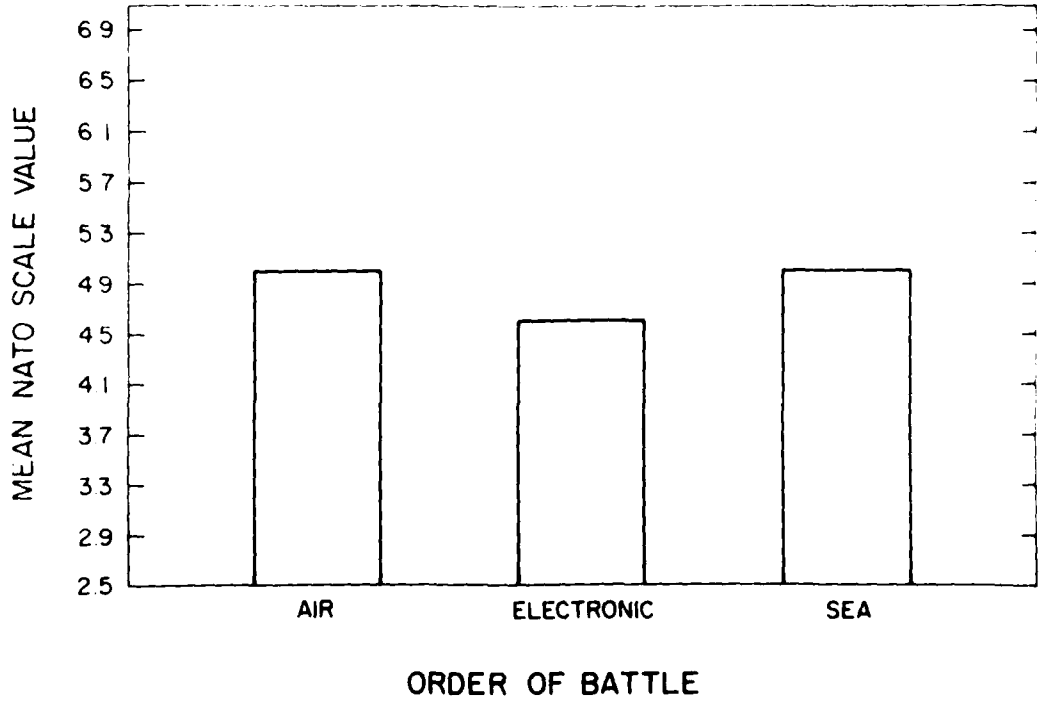


Figure 4: The effect of Order of Battle on NATO scale value

Scene

Differences among the 10 Scenes were also significant, as seen in Figure 5. Appendix C shows the differences among Scenes located by the MCT.

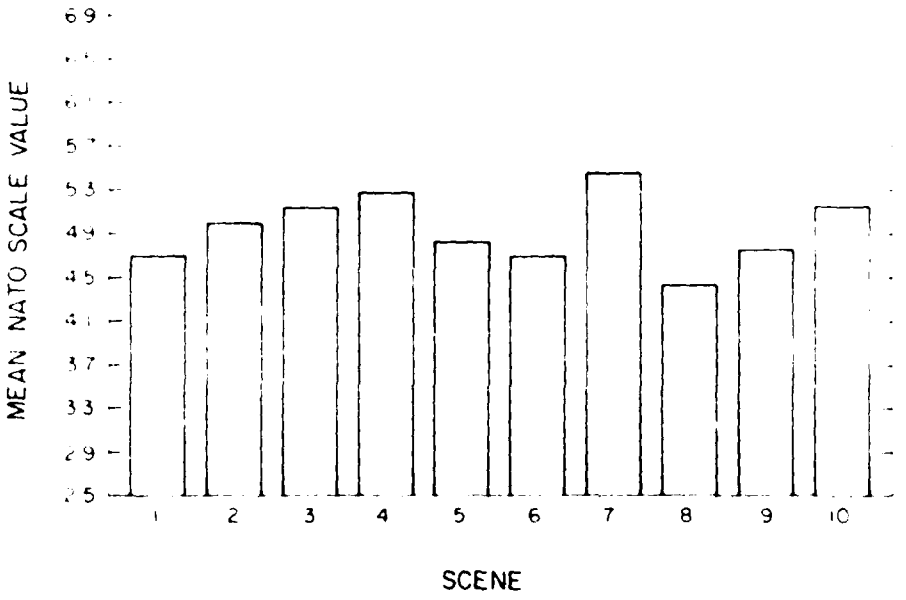


Figure 5: The effect of Scene on NATO scale value

Blur x Noise

The Blur x Noise interaction is shown in Figure 6. In general, the signal-to-noise effect is reduced as blur increases. With reduced blur, the decrease in NATO scale value with decreasing signal-to-noise ratio becomes more pronounced. Simple F ratios were calculated for the Noise effect at each Blur level. All of these Noise simple effects were highly significant, $p < .0001$, with the exception of the 322 μm blur, $p = .003$. Subsequent MCTs

showed that fewer and fewer Noise levels differed from one another with increasing blur. Appendix D gives the details of the MCTs at each level of blur.

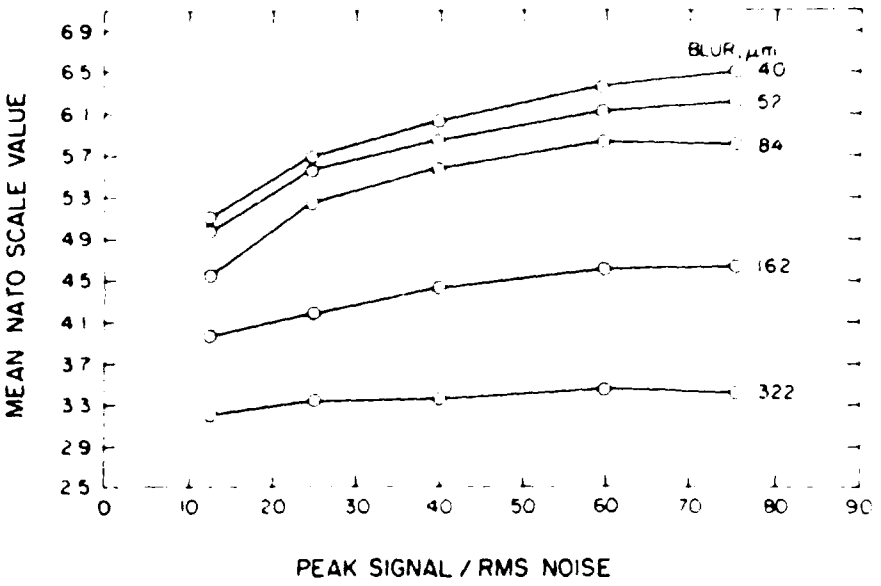


Figure 6: The effect of Blur x Noise on NATO scale value

OB x Blur.

The interaction between OB and Blur is shown in Figure 7. Generally, the effect of Blur was similar for all OBs, although a few crossover data points exist to cause statistical significance of the interaction. Simple F ratios were calculated for OB at each Blur level. The OB simple effect was significant at all levels of Blur except at a blur of 40 μm. A blur of 320 or 322 μm produced differences among all OBs, while at the intermediate levels

of Blur (52, 84, and 162 m), there was no difference between air and sea OBs. Air and electronic OBs did not differ at 84 m.

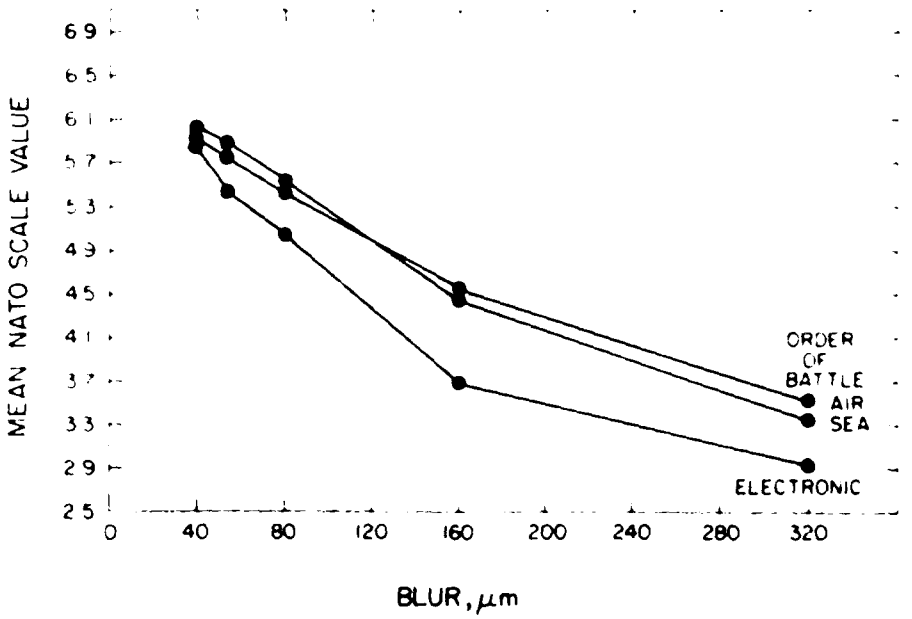


Figure 7: The effect of Blur x Order of Battle on NATO scale value

OB x Noise

The interaction between OB and Noise is shown in Figure 8. The difference in scale values between air and sea OBs becomes greater with decreasing signal-to-noise ratios. Simple effect F-ratios calculated at each Noise level were all significant. Subsequent MCTs showed that the only nonsignificant differences in OBs were between air and sea, except at a signal-to-noise ratio of 12.5, where all OBs differed.

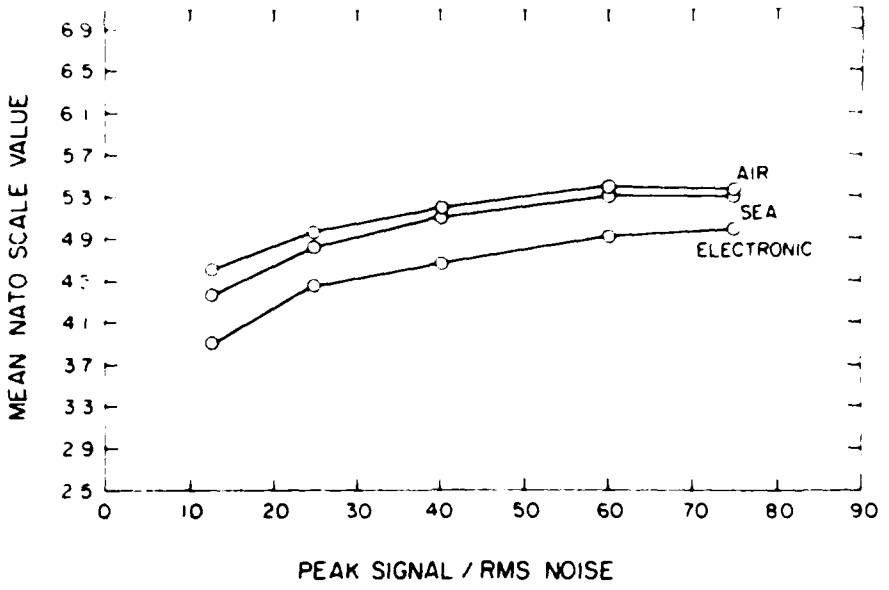


Figure 8: The effect of Noise x Order of Battle on NATO scale value

Scene x Blur

The interaction between Scene and Blur was statistically significant, Figure 9, but is difficult to interpret logically. Simple effect F-ratios were calculated for Scenes at each Blur level. All Scene simple effects were highly significant, $p < .0001$. MCTs showed various differences among Scenes at each Blur level. While the ordered Scene means at each Blur level were similar to the order shown by the overall Scene main effect, each Blur level produced unique differences among Scenes. Appendices D-H depict the Scene differences for Blur levels 40 to 322 m, respectively. MCTs were conducted but provide little in the way of clarification of the interaction; apparently,

some Scenes were affected more by Blur than were other Scenes.

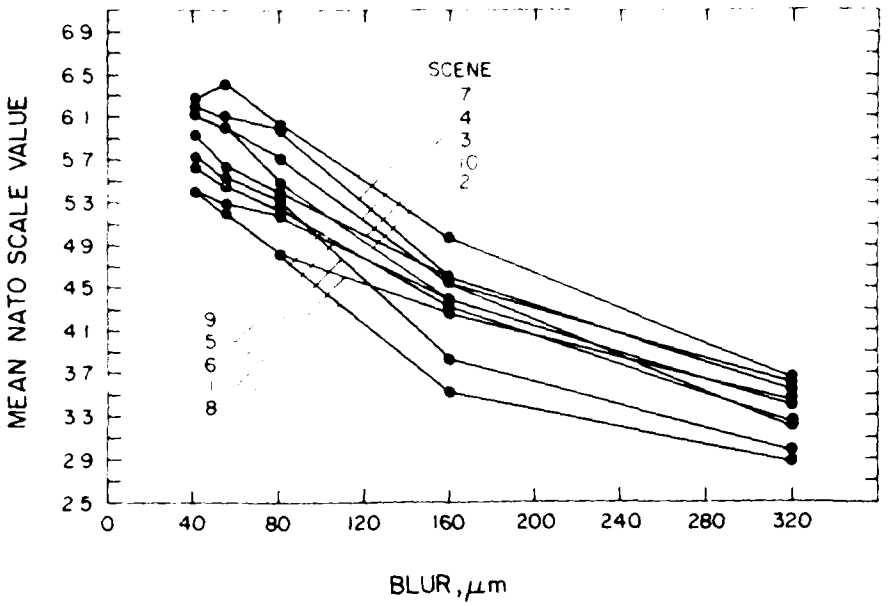


Figure 9: The effect of Blur x Scene on NATO scale value

Scene x Noise

The Scene x Noise interaction did not reach statistical significance. The means were plotted and appear in Figure 10. Obviously, the effect of noise was fairly constant across all Scenes.

Scene x Blur x Noise

The Scene x Blur x Noise interaction was statistically significant, and is plotted in Figures 11-20. Each figure represents the Blur x Noise interaction for a different Scene: Figure 11, Scene 1, Figure 12, Scene 2, Figure 13,

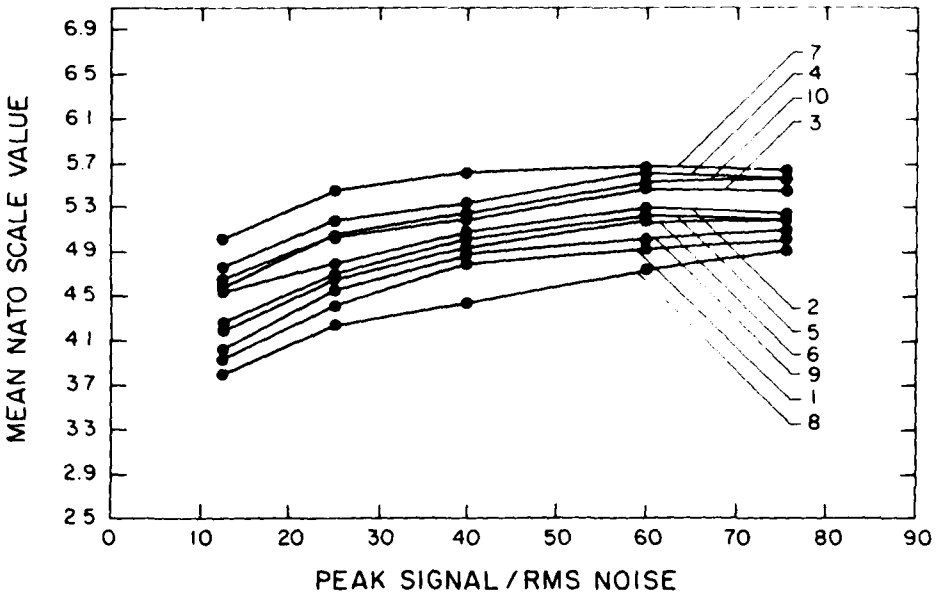


Figure 10: The effect of Noise x Scene on NATO scale values

Scene 3, Figure 14, Scene 4, Figure 15, Scene 5, Figure 16, Scene 6, Figure 17, Scene 7, Figure 18, Scene 8, Figure 19, Scene 9, and Figure 20, Scene 10. Simple F ratios were calculated for each Blur x Noise interaction at each Scene. Only the Blur x Noise interaction for Scene 4 failed to reach significance ($p > .05$). While further simple-effect tests were conducted on each Blur x Noise interaction for each Scene, they tended to repeat findings previously reported here and provided no further understanding of the three-way interaction.

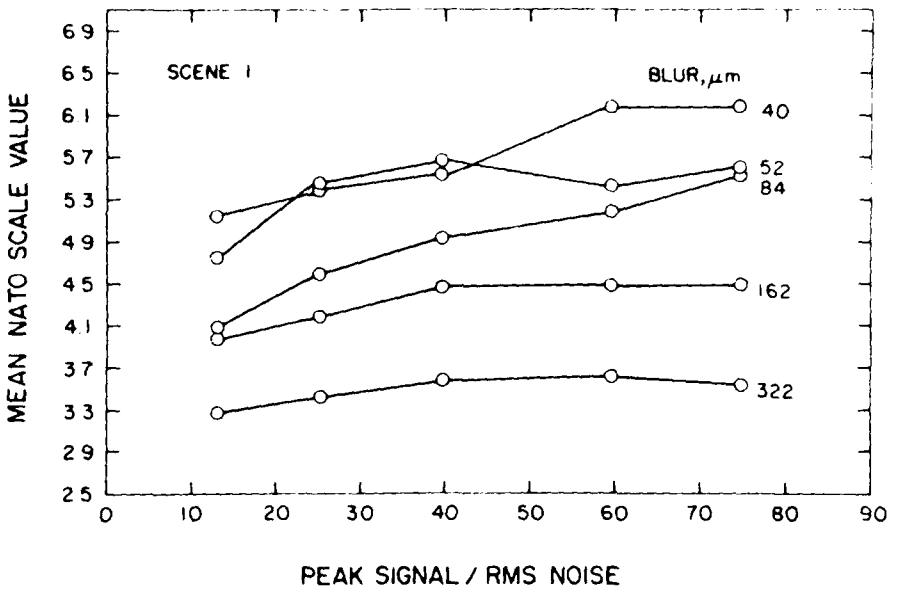


Figure 11: The effect of Blur x Noise on NATO scale value, Scene 1

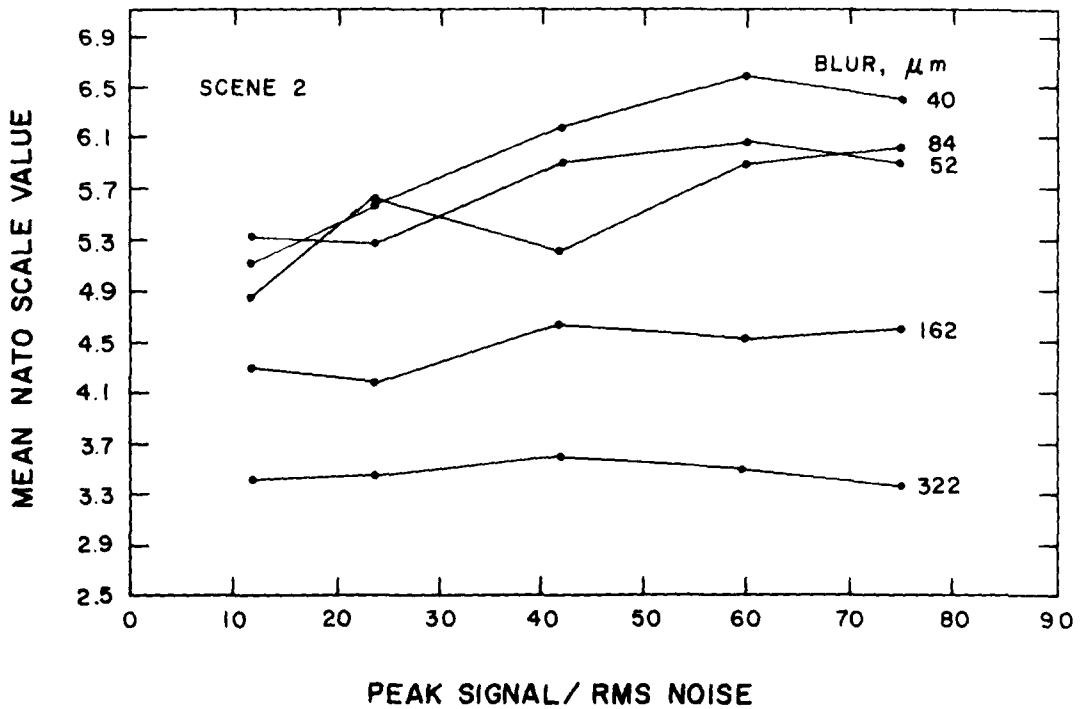


Figure 12: The effect of Blur x Noise on NATO scale value, Scene 2

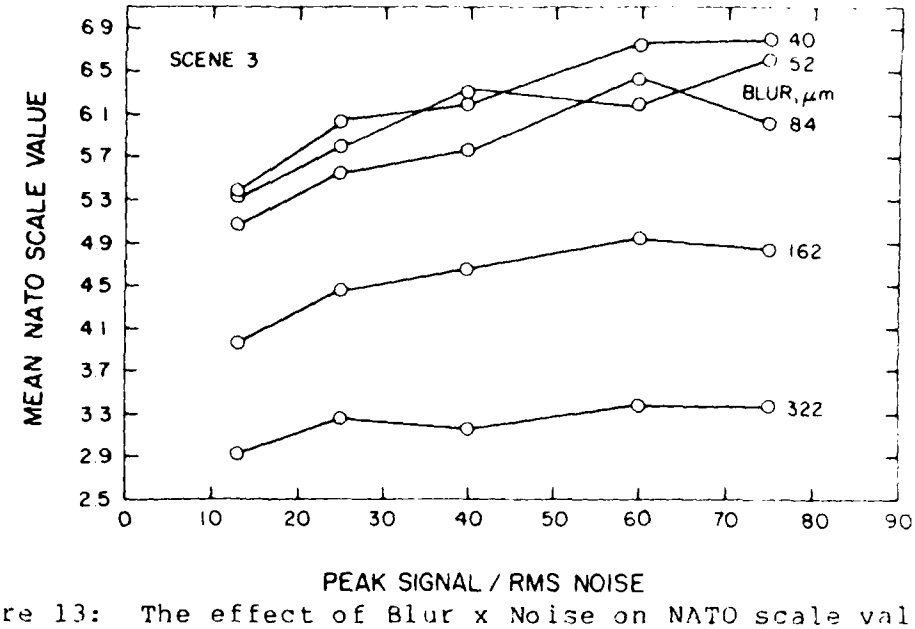


Figure 13: The effect of Blur x Noise on NATO scale value, Scene 3

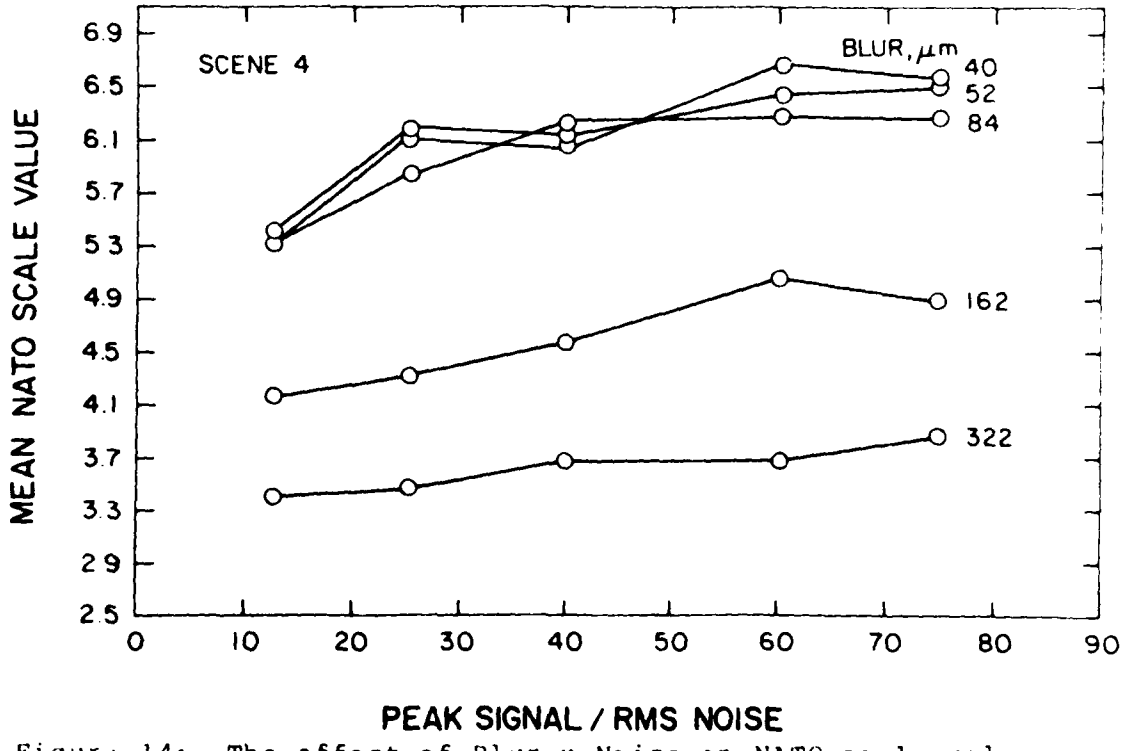


Figure 14: The effect of Blur x Noise on NATO scale value, Scene 4

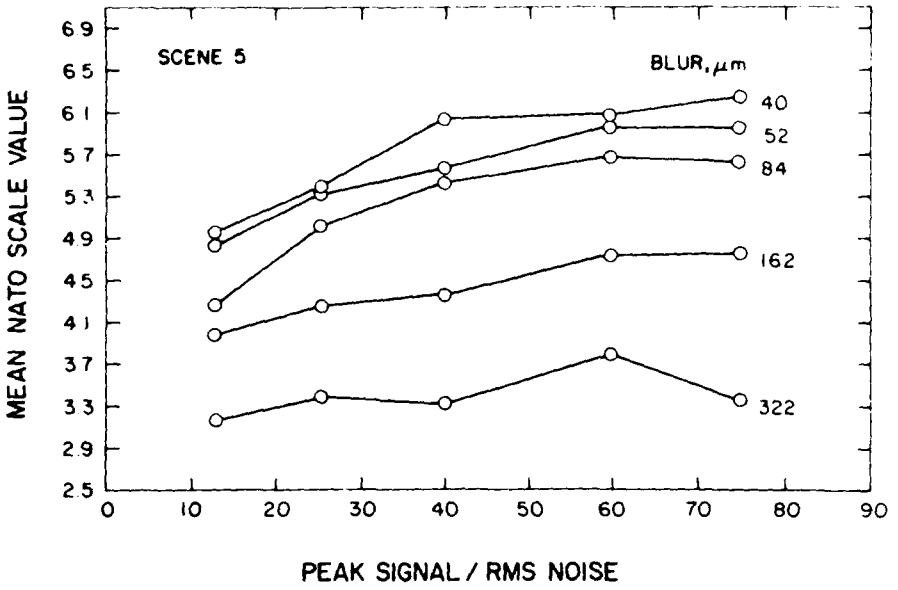


Figure 15: The effect of Blur x Noise on NATO scale value, Scene 5

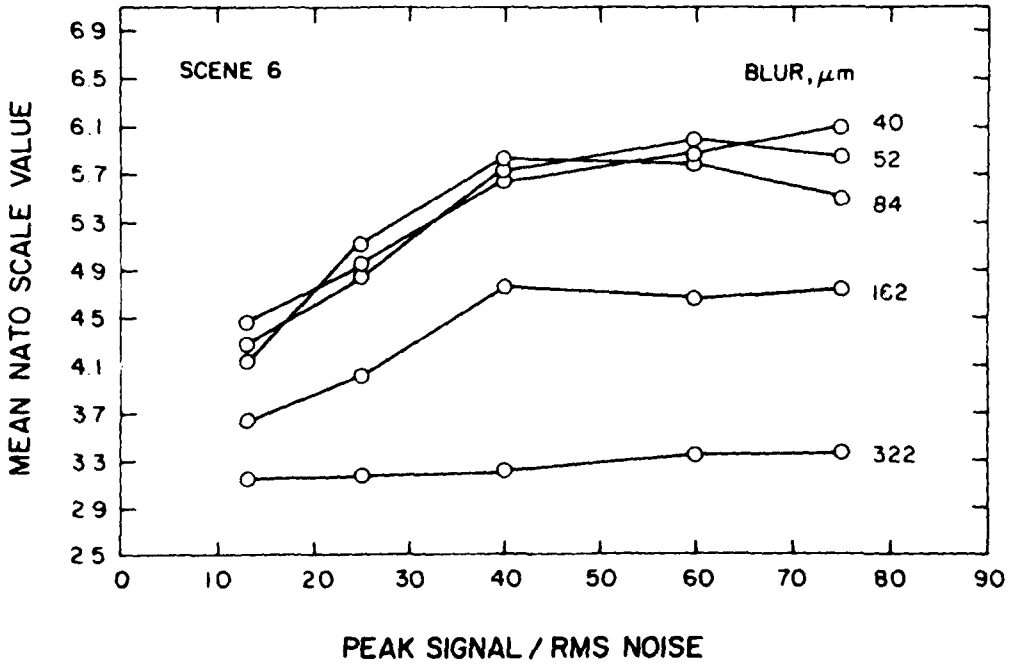


Figure 16: The effect of Blur x Noise on NATO scale value, Scene 6

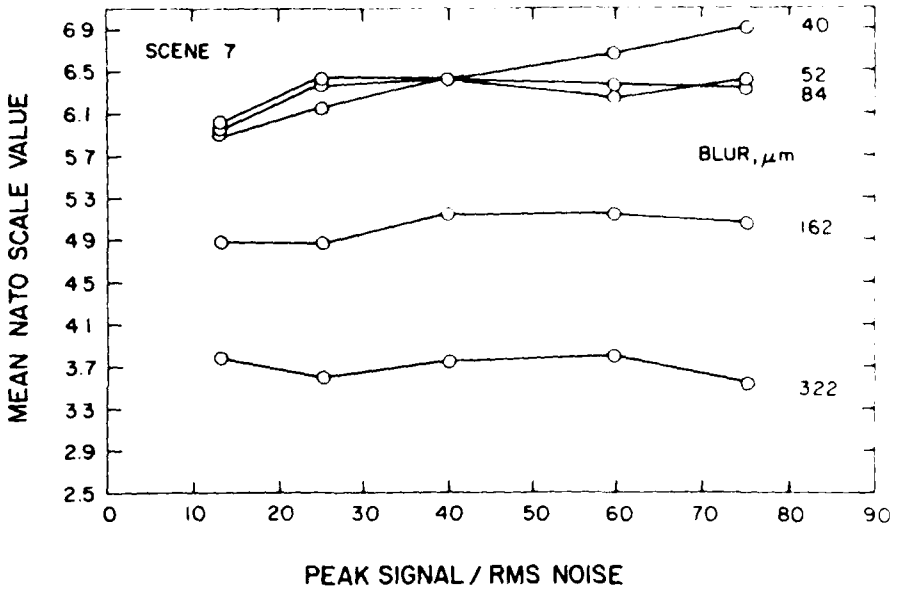


Figure 17: The effect of Blur x Noise on NATO scale value, Scene 7

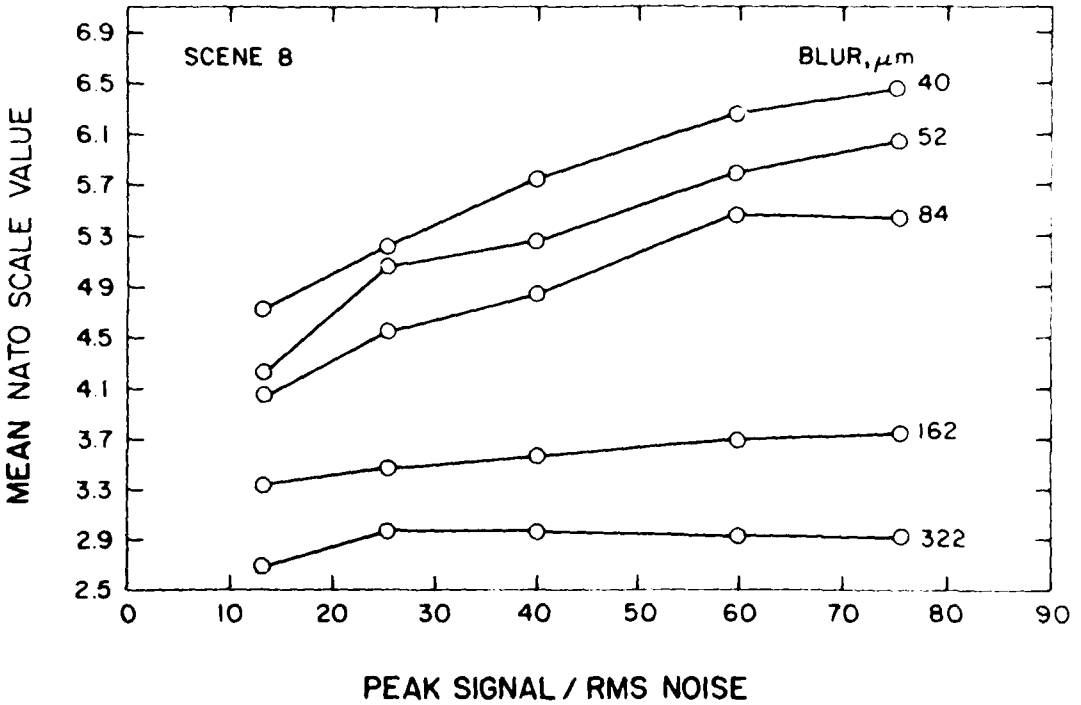


Figure 18: The effect of Blur x Noise on NATO scale value, Scene 8

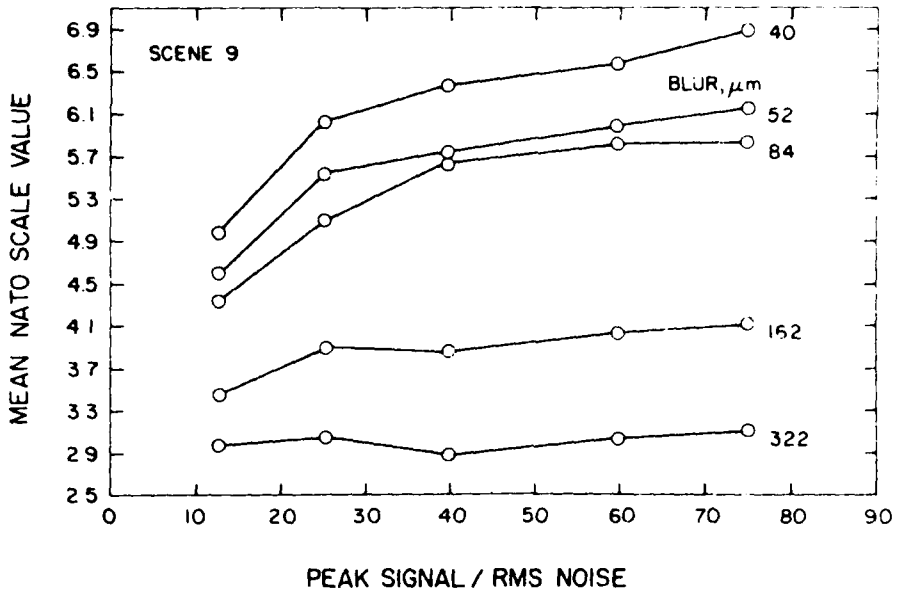


Figure 19: The effect of Blur x Noise on NATO scale value, Scene 9

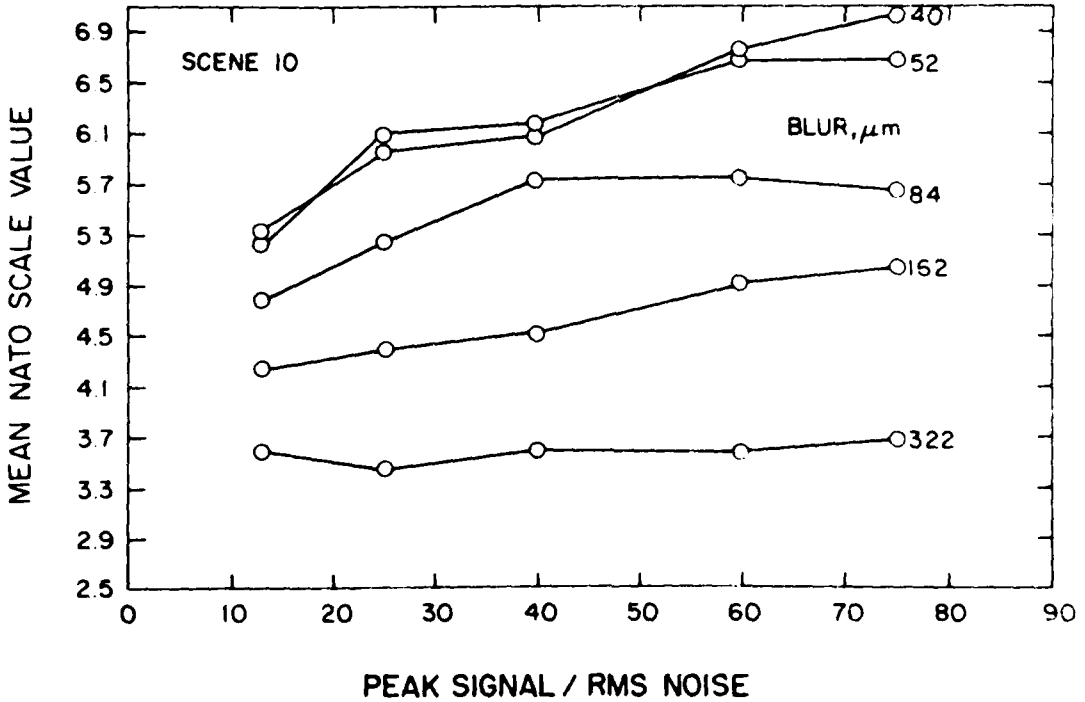


Figure 20: The effect of Blur x Noise on NATO scale value, Scene 10

OPTIMAL NUMBER OF RESPONSE CATEGORIES

When it was decided that the NATO scale would be used in this research, the only concern was that the original scale might have too few response categories (10) to test thoroughly the subjective resolution of the PIs. Therefore, as noted earlier, the normal use of the scale was modified to accommodate 100 response categories. Though 100 categories were believed to be more than necessary, it was felt that the PIs could use the whole number/decimal scale with less difficulty than any other interpolation scheme.

The Shannon-Wiener measure of information (see Equation 1)

$$H = -\sum_{i=1}^{n} p_i \log_2 p_i \qquad (1)$$

was used to calculate the average informational value of the 100 scale alternatives (Attneave, 1959). Eighty-eight of the 100 possible categories were used by the PIs. The Shannon-Wiener formula indicated that 5.95 bits (average) of information were required to arrive at each scaling decision. Therefore, about 62 ($2^{5.95}$) categories would theoretically suffice in scaling this imagery database. The 100-point scale, although larger than necessary and different from the standard 10-point scale, appeared to be easily understood and highly functional in scaling interpretability.

MDS ANALYSIS

Unfortunately, the Bell Laboratories KYST2A program purchased for the MDS analysis was not capable of analyzing all of the data at one time. In fact, a maximum of 60 images had to be selected from the imagery database (250 images) to comply with the limitations of the routine. The selection was done as follows. Because Blur levels of 40 and 52 m and S/N levels of 75 and 60dB did not differ from one another, these lowest levels of degradation were discarded. While 10 Scenes in the imagery database were particularly responsible for its size, three of the more consistently scaled Scenes, one from each OB, were chosen. Scale value data for those Scenes (5, 8, and 10) are plotted in Figures 15, 18, and 20, respectively. Thus, 48 images were submitted to the MDS analysis, 3 Scenes x 4 Blur levels x 4 Noise levels. A dissimilarity matrix was calculated for an input to MDS by taking the absolute difference between each pair of the 48 images; the actual input to the MDS analysis was a triangular matrix without diagonal elements, or 48(47)/2 = 1128 data points.

Several runs of the Bell Laboratories routine were made before the best model was selected. The best model is shown in Equation 2a. The general form of the MDS model is given in Equation 2b.

$$d_{ij} = \left(\sum_{i=1}^{n} x_{ik} - x_{jk}^{2}\right)^{1/2} \qquad (2a)$$

$$d_{ij} = \left(\sum_{i=1}^{n} x_{ik} - x_{jk}^{r}\right)^{1/r} \qquad (2b)$$

The configuration was rotated to principal components. The Minkowski metric associated with the best model was 2.0 (see Equations 2a and 2b). This model is known as the Euclidean model and represents the data in a geometric spatial configuration. The optimal number of dimensions found in fitting the data to the model was determined by successive runs of the program. Figure 21 shows the relationship between stress (inversely related to the fit) and the number of dimensions in the model. Because stress is minimized by a good fit (see Equation 3 for the stress formula), Figure 21 indicates that the model with five dimensions provided the best fit. Values of d represent

$$\text{stress} = \sum_{i<j} \sum_{i<j} (d_{ij} - \hat{d}_{ij})^2 \qquad (3)$$

spatial distances predicted by the model. Consequently, the MDS analysis gave a projection for each of the 48 images on each of the five dimensions. The projections are listed in Appendix I. This configuration resulted in minimum stress, .01, as defined by the authors of the Bell Laboratories program. No other MDS run produced minimum stress.

For each dimension, means were calculated for the levels of the effects found significant in the overall ANOVA. Scene and OB were represented by the same three scenes. Dimensions one and four were apparently unrelated to any of the effects analyzed in the overall ANOVA. However,

dimensions two, three, and five corresponded to noise, ...ne or GB, and blur, respectively.

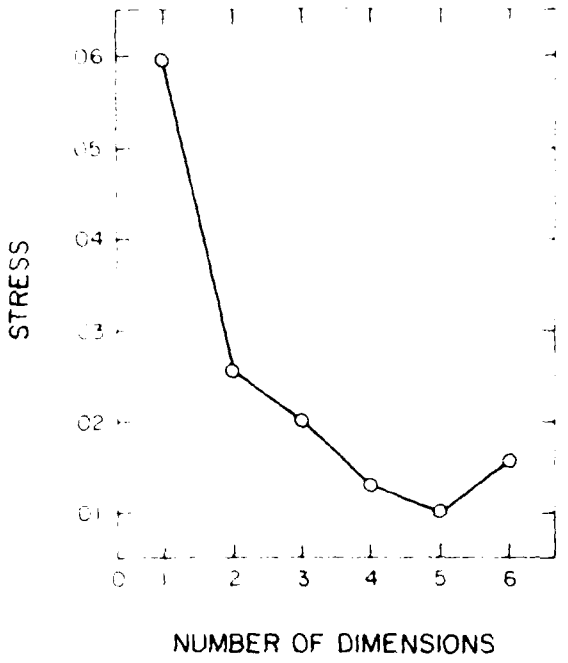


Figure 21: The relationship between stress and the number of dimensions

The MDS projections, all five dimensions, and mean NATO scale values were used to predict the previously obtained information extraction performance in a multiple regression equation. Of the 48 images analyzed by the MDS program, only 24 were applicable because Blur levels of 52 and 162 m were not exploited by the PIs. Of course, all 24 mean scale values, averaged across judges, were available. Multiple regression is tantamount to performing a multiple correlation, except that a slope or weighting factor is provided for each predictor that allows for an evaluation of

the contribution of each predictor toward the overall correlation. The regression model did not approach significance, $R = .41$, $p = .747$, with less than 17% of the variance accounted for. Further, none of the partial F ratios (unique variance in performance accounted for by each predictor) were significant, $F_{1,17} = .15$, $p = .7037$, $F_{1,17} = .33$, $p = .5719$, $F_{1,17} = .48$, $p = .4964$, $F_{1,17} = 1.20$, $p = .2887$, $F_{1,17} = .76$, $p = .3968$, for dimension one ($r = .086$), dimension two ($r = .214$), dimension three ($r = .154$), dimension four ($r = .243$), and dimension five ($r = .193$), respectively. Mean scale value ($r = .042$) also failed to reach significance, $F_{1,17} = .17$, $p = .6819$. Obviously, the regression equation did not predict information extraction performance very well. This fact was true for all predictors, MDS projections and raw rating scale alike.

CORRELATION WITH PERFORMANCE

The [illegible] information extraction [illegible] were correlated with their [illegible] values. A Pearson product-moment [illegible] calculated showed a significant positive [illegible] $r = .47$, $p < .0001$. Averaging over the two [illegible] in each of the Blur x Noise cells, the correlation was improved, $r = .52$, $p < .0001$. (However, the [illegible] not significant, $z = .61$, $p = .275$.) Further, averaging over the five judges in each Blur condition [illegible]

the correlation, $r = .70$, $p = .003$. (Statistically, this correlation was not a significant improvement over .47 or .52, $z = 1.22$, $p = .110$ and $z = .96$, $p = .170$, respectively.)

Finally, overall mean performance scores (averaged over scenes and PIs) were correlated with corresponding mean scale scores for each of the 15 Blur and Noise combinations. The correlation was improved and significant, $r = .898$, $p = .0001$. (This correlation was a statistical improvement above .47, .52, and .70, $z = 3.22$; $p = .0007$, $z = 2.93$; $p = .0017$, and $z = 2.11$; $p = .017$, respectively.)

VI. DISCUSSION

The NATO scale and its usage here presented no problems for the PIs. In general, both the imagery database and the scaling task were well received by all of the PIs. While PIs, by reputation, are very scrupulous of photographic quality and, for security reasons, sensitive about giving insight into their work procedure, this study was not hindered by the use of these highly trained judges--except for the one case of attrition which was due to workload.

NOISE AND BLUR EFFECTS

The overall statistical analysis confirmed the first hypothesis. Both noise and blur degradations reduced the judged interpretability of the images. In general, the greater the physical image degradation, the greater was the reduction in interpretability. However, for both noise and blur, there were no differences in interpretability between the two lowest levels of degradation. One or both of two conditions may have been responsible for this result. First, images assumed to be free of noise had somewhat higher base levels of noise even before any degradation was effected. The second possible explanation regards the PIs' sensitivity to noise and blur. If the true relationship

between noise or blur and interpretability is sigmoid-shaped, then judges are psychophysically insensitive to small changes in noise or blur level when the extent of degradation is either very slight or very great. Stevens (1975, p. 186) has suggested that when the full range of a stimulus parameter is used experimentally, the psychophysical function tends to be sigmoid-shaped.

The combined effect of blur and noise indicated that these degradations are interactive. The MCT for the Blur x Noise interaction indicated that noise has little or no effect on a highly blurred image. But, blur did have an effect on a very "noisy" image. It appears, then, that PIs react differently to noise degradation than they do to blur. Subjective comments by PIs substantiate this notion; it is common for PIs to state that they can "look through" noise but cannot disregard blur.

NUMBER OF RESPONSE CATEGORIES

The determination of the optimal number of response categories took an absolute judgement or information processing approach. Given that the information processing analysis is so easy to perform, future research might attempt to determine the conditions under which the technique and these results are valid. It should be kept in mind that the optimal number of response categories is affected by numerous factors, such as the psychophysical

range and spacing of the stimuli (Alluisi, 1957). This imagery database was previewed by senior PIs and deemed representative of the range of quality of imagery actually interpreted by PIs in practice. This analysis showed that about 62 categories would be required to allow adequately for the subjective quality differences in the present imagery database. Other imagery databases might well require more than or less than 62 categories. However, because this 100-point scale, while larger than needed, was easy to use by the PIs, it is strongly recommended as a replacement for the current 10-point scale NATO scale.

MDS ANALYSIS

MDS analysis fits the data to a multidimensional spatial configuration and calculates a projection on each dimension for each cell in the experimental design. While MDS analysis is simply a mathematical curve fitting procedure, similar to factor analysis, difficulty arises in the interpretation of the dimensions of the preferred configuration. It is only through the repeated replication of an experiment that an investigator can, with some degree of confidence, begin to understand the meaning of the MDS dimensions. In light of the numerous significant effects reported in the overall ANOVA, it is beyond the present data to interpret the five dimensions of the reported configuration. It does appear that at least three of the

dimensions map well onto known database variables of noise, scene (or OB), and blur. Therefore, the MDS analysis, while not conclusive in its defined dimensions, appears to be consistent with known facts about the imagery database.

Minimum stress was achieved. However, because the MDS routine is limited to six dimensions in fitting the data, there is no assurance that the best model did not represent a local stress minimum rather than a global minimum. While the projections are listed in Appendix I, the usefulness of this subjective response configuration remains uncertain.

The regression equation utilizing the projections to predict performance failed to add any meaningfulness to the MDS analysis. The ultimate goal of either scale values or physical metrics is the prediction of performance. Future MDS analyses in this area of research should concentrate on validating MDS projections in terms of performance. Perhaps scale scores that correlate better with performance than these did will result in projections that better predict performance. It is possible that the predictive value of the MDS projections would have been enhanced by a complete set of performance data to accommodate the 48 images selected for MDS.

In regard to past research and MDS analysis, it is interesting that this MDS analysis showed that a similar number of dimensions were required in the spatial configuration. Also, the interpretation of the dimensions

was quite similar to past research. Blur is doubtlessly related to sharpness. Considering the data of Marmolin and Nyberg (1978), these data support the contention that the Euclidean model, r = 2.0, is more appropriate for MDS analyses of subjective image quality.

RELATIONSHIP BETWEEN INFORMATION EXTRACTION AND SCALING

In light of the correlations reported by other authors, the correlation (for individual PIs and Scenes) obtained here between information extraction performance and NATO scale values was disappointing though statistically significant. The most cogent explanation for the low correlation, and the failure of the regression analysis discussed above, is the difference between the experimental designs in the scaling and information extraction studies. The scaling study was a completely factorial design, whereas, in the exploitation study, Scenes were confounded with Noise for each PI, and E ur was treated as a between-subjects variable. Within-subject factors have been shown to be more sensitive in demonstrating main effects (Grice and Hunter, 1964). Blur, the more severe degradation, was treated as a between-subject variable because of experimental constraints and because only one of the degradations could be treated within subjects in that experiment. Unfortunately, the overall ANOVA for the information extraction experiment failed to find a

significant main effect of Blur. Noise, the less impactful parameter in this study as shown by the Blur x Noise interaction (see Figure 6), was significant in the information extraction study. This difference in experimental designs doubtlessly had an effect on the association between scale values and information extraction performance.

On the other hand, averaging across Scenes and PIs to obtain overall sample means for the 5 Noise x 3 Blur combinations yielded a very satisfactory correlation of r = .898 between NATO scale value and information extraction performance. Thus, while individual scene/PI combinations cannot be predicted very accurately, overall group performance can. And, after all, the main objective in most image interpretation research is to predict the performance of the typical, not the individual, PI.

DIGITAL IMAGERY SIMILARITIES

A final point should be made concerning a comparison between analog and digital modes of imagery presentation. From the noise studies reviewed, recall that the middle subjective scale value corresponded to a signal-to-noise ratio of about 30 dB (or 32:1). From an inspection of Figure 3, it can be seen that a scale value of five (midpoint of NATO scale) corresponds to a signal-to-noise ratio of 42:1, or about 32 dB.

The average signal-to-echo ratio at the middle scale value in the studies reviewed was 20 dB (or 10:1). In Figure 2, the NATO middle scale value of five corresponds to 115 μm of blur or a signal-to-echo ratio of about 17:1 (signal = 2000) or 25 dB. For both noise and blur, the obtained ratios, "signal-to-degradation", at the middle scale value were about the same as those reported in the analog imagery literature.

On the basis of this comparison, digital images would appear to be very similar to analog images in terms of subjective quality or interpretability. However, due to myriad differences in the experimental procedures, past and present (e.g., electronic versus photographic presentation, dynamic versus static noise), these conclusions can only be viewed as cursory pending further research.

VII. CONCLUSIONS

The main effects of Blur and Noise and the Blur x Noise interaction showed that digital images, like analog images, are poorer in subjective quality as the degree of Blur or Noise increases. The Blur x Noise interaction indicated that Blur was the more serious degradation at the levels investigated.

Scene content currently unfamiliar to PIs does not appear less interpretable than scenes commonly exploited by the Air Force. In general, PIs seem to evaluate the information available in an image in an objective manner, regardless of practice or bias.

The optimal number of response categories for an image interpretability scale is greater than 10. Because well-practiced PIs are so attuned to resolution differences in photographic imagery, as many as 62 categories may be needed to accurately judge interpretability. The scale of 100 points used in this study provided useful information.

MDS can be used to represent the subjective dimensionality of a large imagery database. The resulting spatial configuration can then be used to determine the physical parameters of the imagery that underly the perception of interpretability. However, the attempt to

predict information extraction performance from MDS projections of subjective data failed. If MDS analysis is to be used as a predictive response surface, the projections must be related, in a meaningful way, to performance.

Finally, mean scale scores correlated quite well with information extraction performance in spite of differences in experimental design. The correlation proved best when variance due to scene content and PIs was eliminated by averaging the data.

REFERENCES

Allnatt, J. W. and Prosser, R. D. Subjective quality of television pictures impaired by long-delayed echoes. Proceedings of the IEE, 1965, 112, 487-493.

Alluisi, E. A. Conditions affecting the amount of information in absolute judgements. Psychological Review, 1957, 64, 97-103.

Attneave, F. Applications of information theory to psychology: A summary of basic concepts, methods, and results. New York: Holt, Rinehart, and Winston, 1959.

Barr, A. J., Goodnight, J. H., Sall, J. P., and Helwig, J. T. A user's guide to SAS 76. Raleigh, NC: SAS Institute Inc., 1976.

Below, F., Huertas-Sendra, F., Fritze, E., and Samraj, E. The subjective disturbing effect of noise in television pictures. EBU Review, 1963, 78, 49.

Brainard, R. W., Sadacca, R., Lopez, L. W., and Ornstein, G. N. Development and evaluation of a catalog technique for measuring image quality. Washington, D.C.: U.S. Army Personnel Research Office, Technical Research Report 1150, AD 645-644, August, 1966.

Burke, J. J. and Snyder, H. L. Quality metrics of digitally derived imagery and their relation to interpreter performance: II. interpretability and judged quality of hard copy imagery. Paper presented at the Society of Photographic Scientists and Engineers Meeting, Tucson, Arizona, January, 1981.

Burke, J. J. and Strickland, R. N. Quality metrics of digitally derived imagery and their relation to interpreter performance: I. Preparation of a large-scale database. Technical Report HFL 81-1, Virginia Polytechnic Institute and State University, Blacksburg, VA, 1982.

Cavanaugh, J. R. and Lessman, A. M. Subjective effects of differential gain and differential phase distortions in NTSC color television pictures. Journal of the SMPTE, 1971, 80, 614-619.

Erikson, C. W. and Hake, H. W. Multidimensional stimulus differences and accuracy of discrimination. Journal of Experimental Psychology, 1955, 50, 153-160.

Geddes, W. K. E. The relative impairment produced by random noise in 405-line and 625-line television pictures. EBU Review, 1963, 78, 46.

Grice, G. R. and Hunter, J. J. Stimulus intensity effects depend on the type of experimental design. Psychological Review, 1964, 71, 247-256.

Guilford, J. P. Psychometric methods. New York: McGraw-Hill, 1954.

Humes, J. M. and Bauerschmidt, D. K. Low light level TV viewfinder simulation program: Phase B. Wright-Patterson Air Force Base, Ohio: Aerospace Medical Research Laboratory, Technical Report AFAL-TR-68-271, November, 1968.

Klingberg, C. L., Elworth, C. L., and Fillean, C. K. Image quality and detection performance of military interpreters. The Boeing Company, Final Report of USAFOSR Contract No. F44620-69-C-0128, April, 1970.

Lessman, A. M. The subjective effects of echoes in 525-line monochrome and NTSC color television and the resulting echo time weighting. Journal of the SMPTE, 1972, 81, 907-916.

Marmolin, H. and Nyberg, S. Multidimensional scaling of subjective image quality. Forsvarets forskningsanstalt, Huvudenhet 3, FOA Report no. C 30039-H9, Stockholm, November, 1978.

Miller, G. A. The magic number seven, plus or minus two: some limits on our capacity for processing information. Psychological Review, 1956, 63, 81-97.

Muller, P. F., Sidorsky, R. C., Slivinske, A. J., Alluisi, E. A., and Fitts, P. M. The symbolic coding of information on cathode ray tubes and similar displays. USAF, WADC Technical Report No. 55-375, 1955.

Newell, G. F. and Geddes, W. K. E. Visibility of small luminance perturbations in television displays. Proceedings of the IEE, 1963, 110, 1979.

Prosser, R. D. and Allnatt, J. W. Subjective quality of television pictures impaired by random noise. Proceedings of the IEE, 1965, 112, 1099-1102.

Prosser, R. D., Allnatt, J. W., and Lewis, N. W. Quality grading of impaired television pictures. Proceedings of the IEE, 1964, 111, 491-502.

Sadacca, R. and Schwartz, A. I. Psychophysical aspects of image quality--exploratory study. Washington, D.C.: U.S Army Personnel Research Office, Technical Research Note 136, September, 1963.

Snyder, H.L., Turpin, J.A., and Maddox, M.E. Quality metrics of digitally derived imagery and their relation to interpreter performance: II. Hard-Copy Digital Imagery Interpretability. Technical Report 81-3, Virginia Polytechnic Institute and State University, Blacksburg, VA, 1981.

Stevens, S. S. Psychophysics. New York: Wiley and Sons, 1975.

Torgerson, W. Theory and methods of scaling. New York: Wiley and Sons, 1958.

Weaver, L. E. Subjective impairment of television pictures. Electronic and Radio Engineering, 1959, 36, 170.

Weaver, L. E. The quality rating of color television pictures. Journal of the SMPTE, 1968, 77, 610-612.

Appendix A

THE NATO SCALE -- AN IMAGE INTERPRETABILITY RATING SCALE

Rating Category 0

Useless for interpretation due to cloud cover, poor resolution, etc.

Rating Category 1

Detect the presence of larger aircraft [illegible].
Detect surface ships.
Detect ports and harbors (including piers [illegible].
Detect railroad yards and shops.
Detect coasts and landing beaches.
Detect surfaced submarines.
Detect armored artillery ground force [illegible].
Recognize urban areas.
Recognize terrain.

Rating Category 2

Detect bridges.
Detect ground forces installations (including training areas, administration/barracks buildings, vehicle storage buildings, and vehicle parking areas).
Detect airfield facilities (count accurately all larger aircraft, by type, straight-wing and large delta wing).
Recognize ports and harbors (including large ships and drydocks).

Rating Category 3

Detect communications equipment (radio/radar).
Detect supply dumps (POL/ordnance).
Detect and count accurately all straight-wing aircraft, all swept-wing aircraft, and all delta-wing aircraft.
Detect command and control headquarters.
Detect surface-to-surface and surface-to-air missile sites (including vehicles and other pieces of equipment).
Detect land minefields.
Recognize bridges.
Recognize surface ships (distinguish between a cruiser and a destroyer by relative size and hull shape).
Recognize coast and landing beaches.
Recognize railroad yards and shops.
Recognize surfaced submarines.
Identify airfield facilities.
Identify urban areas.
Identify terrain.

Rating Category 4

Detect rockets and artillery.
Recognize troop units.
Recognize aircraft (such as FAGOT/MIDGET when singly deployed).
Recognize missile sites (SSM/SAM). Distinguish between missile types by the presence and relative position of wings and control fins.
Recognize nuclear weapons components.
Recognize land minefields.
Identify ports and harbors.
Identify railroad yards and shops.
Identify trucks at ground force installations as cargo, flatbed, or van.
Identify a KRESTA by the helicopter platform flush with the fantail, a KRESTA II by the raised helicopter platform (one deck level above fantail and flush with the main deck).

Rating Category 5

Detect the presence of call letters or numbers and alphabetical country designator on the wings of large commercial or cargo aircraft (where alphanumerics are three feet high or greater).
Recognize command and control headquarters.
Identify a singly deployed tank at a ground forces installation as light or medium/heavy.
Perform Technical Analysis (PTA) on airfield facilities.
PTA on urban areas and terrain.

Rating Category 6

Recognize radio/radar equipment.
Recognize supply dumps (POL/ordnance).
Recognize rockets and artillery.
Identify bridges.
Identify troop units.
Identify coast and landing beaches.
Identify a FAGOT or MIDGET by canopy configuration when singly deployed.
Identify the ground force equipment; T-54/55 tank, BTR-50 armored personnel carrier, 57 mm AA gun.
Identify by type, RBU installations (e.g., 2500 series), torpedo tubes (e.g., 21 inch/53.34 cm), and surface-to-air missile launchers on a KANIN DDG, KRIVAC DDGSP, or KRESTA II.
Identify a ROMEO-class submarine by the presence of the cowling for the snorkel induction and the snorkel exhaust.
Identify a WHISKEY-class submarine by the absence of the cowling and exhaust.

Rating Category 7

Identify radar equipment.
Identify major electronics by type on a KILDEN DDGS or KASHIN DLG.
Identify command and control headquarters.
Identify nuclear weapons components.
Identify land minefields.
Identify the general configuration of an SSBN/SSGN submarine sail, to include relative placement of bridge periscope(s) and main electronics/navigation equipment.
PTA on ports, harbors, and roads.
PTA on railroad yards and shops.

Rating Category 8

Identify supply dumps (POL/ordnance).
Identify rockets and artillery.
Identify aircraft.
Identify missile sites (SSM/SAM).
Identify surface ships.
Identify vehicles.
Identify surfaced submarines (including components such as ECHO II SSGN sail missile launcher elevator guide and major electronics/navigation equipment by type).
Identify, on a KRESTA II, the configuration of the major components of larger electronics equipment and smaller electronics by type.
Identify limbs (arms, legs) on an individual.
PTA on bridges.
PTA on troop units.
PTA on coast and landing beaches.

Rating Category 9

Identify in detail the configuration of a D-30 howitzer muzzle brake.
Identify in detail on a KILDEN DDGS the configuration of torpedo tubes and AA gun mountings (including gun details).
Identify in detail the configuration of an ECHO II SSGN sail including detailed configuration of electronics communications equipment and navigation equipment.
PTA on radio/radar equipment.
PTA on supply dumps (POL/ordnance).
PTA on missile sites.
PTA on nuclear weapons components.

Appendix B

SCENE MCT RESULTS

A connecting line indicates no differences among means, as determined by the Newman-Keuls multiple comparisons test, at the $p < .05$ level of confidence.

ORDERED SCENES

8 1 6 9 5 2 3 10 4 7

Appendix C

BLUR X NOISE MCT RESULTS

A connecting line indicates no difference among means, as determined by the Newman-Keuls multiple comparison test, at the $p < .05$ level of confidence.

40 µm BLUR

ORDERED SIGNAL-TO-NOISE RATIOS

12 24 42 60 75

52 µm BLUR

ORDERED SIGNAL-TO-NOISE RATIOS

12 24 42 60 75

84 µm BLUR

ORDERED SIGNAL-TO-NOISE RATIOS

12 24 42 60 75

162 μm BLUR

ORDERED SIGNAL-TO-NOISE RATIOS

12 24 42 60 75

322 μm BLUR

ORDERED SIGNAL-TO-NOISE RATIOS

12 24 42 60 75

Appendix D

SCENE MCT RESULTS -- BLUR LEVEL 40

A connecting line indicates no difference among means, as determined by the Newman-Keuls multiple comparisons test, at the $p < .05$ level of confidence.

ORDERED SCENES

6 1 8 5 2 4 9 10 3 7

Appendix E

SCENE MCT RESULTS -- BLUR LEVEL 52

A connecting line indicates no difference among means, as determined by the Newman-Keuls multiple comparisons test, at the $p < .05$ level of confidence.

ORDERED SCENES

8 6 1 5 9 2 3 4 10 7

Appendix F

SCENE MCT RESULTS -- BLUR LEVEL 84

A connecting line indicates no difference among means, as determined by the Newman-Keuls multiple comparisons test, at the $p < .05$ level of confidence.

ORDERED SCENES

8 1 5 6 9 10 2 3 4 7

Appendix G

SCENE MCT RESULTS -- BLUR LEVEL 162

A connecting line indicates no difference among means, as determined by the Newman-Keuls multiple comparisons test, at the $p < .05$ level of confidence.

ORDERED SCENES

8 9 1 6 2 5 3 4 10 7

Appendix H

SCENE MCT RESULTS -- BLUR LEVEL 322

A connecting line indicates no difference among means, as indicated by the Newman-Keuls multiple comparisons test, at the $p < .05$ level of confidence.

ORDERED SCENES

8 9 3 6 5 2 1 10 4 7

```
---
  -----
    ---------
        ----------
            --------
```

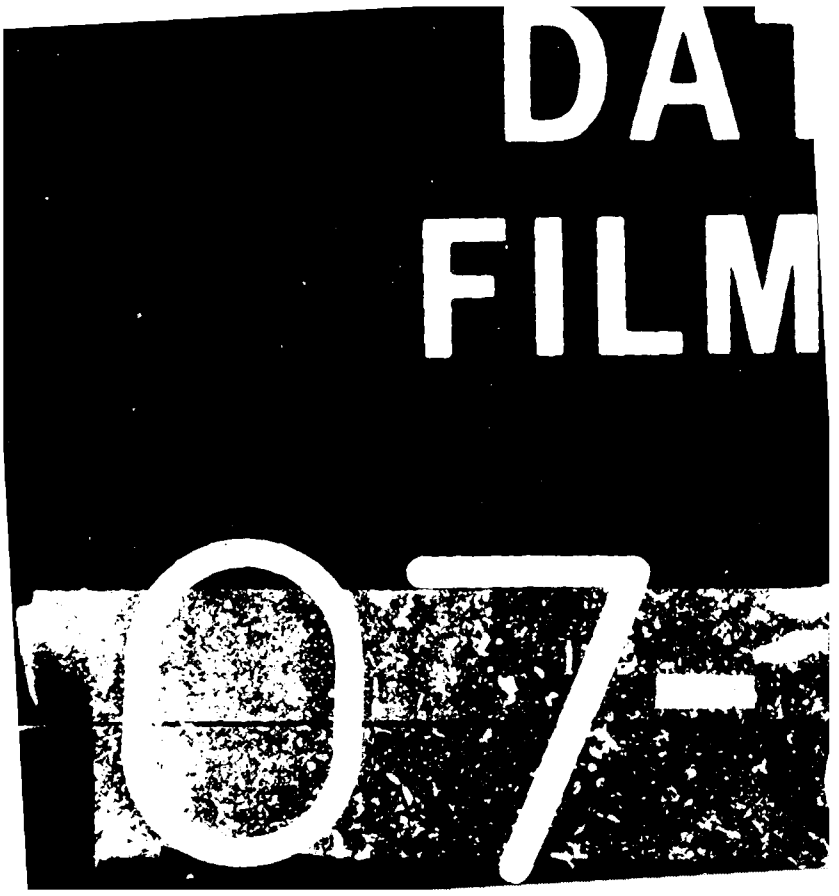
Appendix I

PROJECTIONS OF MDS ANALYSIS

Tabled values are the projections on each of the five dimensions calculated for each image in terms of the parameters, Scene, Noise, and Blur.

				DIMENSION				
				1	2	3	4	5
SCENE	NOISE	BLUR	IMAGE					
5	33	52	1	-0.631	0.768	0.582	- . [illegible]	- .[illegible]
5	33	84	2	-0.795	0.624	.[illegible]	[illegible]	-.[illegible]
5	33	162	3	-0.448	-0.708	-.[illegible]	. .64	.[illegible]
5	33	322	4	0.769	-0.226	[illegible]	[illegible]	-.[illegible]
5	48	52	5	-0.825	0.545	[illegible]	[illegible]	-[illegible]
5	48	84	6	-0.852	0.370	-[illegible]	.[illegible]	[illegible]
5	48	162	7	0.052	-0.847	[illegible]	-[illegible]	-0.145
5	48	322	8	0.907	0.387	-[illegible]	[illegible]	[illegible]
5	83	52	9	-0.848	0.185	-0.[illegible]	[illegible]	-.[illegible]
5	83	84	10	-0.722	-0.327	-0.[illegible]	[illegible]	0.[illegible]
5	83	162	11	0.210	-0.795	[illegible]	-[illegible]	-0.[illegible]
5	83	322	12	0.918	0.320	-[illegible]	[illegible]	[illegible]
5	167	52	13	-0.567	-0.591	-0.328	-0.093	0.[illegible]
5	167	84	14	0.196	-0.802	[illegible]	-[illegible]	-0.[illegible]
5	167	162	15	0.593	-0.489	0.[illegible]	0.[illegible]	-0.[illegible]
5	167	322	16	0.859	0.495	-0.[illegible]	-[illegible]	[illegible]
8	33	52	17	-0.722	0.715	[illegible]	[illegible]	-[illegible]
8	33	84	18	-0.838	0.096	-[illegible]	[illegible]	[illegible]
8	33	162	19	0.844	-0.057	-.[illegible]	[illegible]	[illegible]
8	33	322	20	0.704	0.570	-[illegible]	-[illegible]	-[illegible]
8	48	52	21	-0.839	0.092	-[illegible]	[illegible]	[illegible]
8	48	84	22	-0.577	-0.577	-[illegible]	-[illegible]	[illegible]
8	48	162	23	0.900	[illegible]	[illegible]	[illegible]	[illegible]
8	48	322	24	0.724	[illegible]	-[illegible]	-[illegible]	-[illegible]
8	83	52	25	-0.764	-0.225	-[illegible]	[illegible]	[illegible]
8	83	84	26	-0.239	-0.[illegible]	-[illegible]	-[illegible]	-[illegible]
8	83	162	27	0.915	[illegible]	[illegible]	[illegible]	[illegible]
8	83	322	28	0.740	[illegible]	-[illegible]	-[illegible]	-[illegible]
8	167	52	29	0.280	-0.[illegible]	[illegible]	-[illegible]	-[illegible]
8	167	84	30	[illegible]	-[illegible]	[illegible]	[illegible]	-.[illegible]
8	167	162	31	[illegible]	[illegible]	-[illegible]	[illegible]	[illegible]
8	167	322	32	[illegible]	[illegible]	-[illegible]	[illegible]	-[illegible]
10	33	52	33	-0.185	[illegible]	[illegible]	-[illegible]	[illegible]
10	33	84	34	-0.768	[illegible]	[illegible]	[illegible]	-[illegible]
10	33	162	35	-0.644	-0.481	-[illegible]	-[illegible]	[illegible]

10	33	322	36	0.902	0.136	0.100	0.271	0.051
10	48	52	37	-0.510	0.745	0.714	-0.425	-0.112
10	48	84	38	-0.772	0.667	0.280	0.166	-0.063
10	48	162	39	-0.151	-0.850	0.021	-0.197	-0.014
10	48	322	40	0.880	0.056	0.169	0.286	0.087
10	83	52	41	-0.544	0.756	0.688	-0.342	-0.106
10	83	84	42	-0.827	0.027	-0.337	0.224	0.095
10	83	162	43	0.021	-0.849	0.168	-0.172	-0.10-
10	83	322	44	0.920	0.263	-0.245	0.113	0.0--
10	167	52	45	-0.824	0.011	-0.350	0.221	0.08-
10	167	84	46	-0.510	-0.651	-0.293	-0.117	0.048
10	167	162	47	0.237	-0.783	0.329	-0.084	-0.067
10	167	322	48	0.889	0.123	0.141	0.325	0.120


DA
FILM
07-