AFOSR-TR- 82-0467

HFL-81-2

12

Quality Metrics Of Digitally Derived Imagery and Their Relation to Interpreter Performance: II. Effects of Blur and Noise on Hard-Copy Interpretability

ANNUAL TECHNICAL REPORT
for
Period Ending 30 September 1980

Harry L. Snyder, Ph.D.
James A. Turpin, M.S.
Michael E. Maddox, Ph.D.

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

Contract F49620-78-C-0055
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- 82-0467 2. GOVT ACCESSION NO. AD-A115160	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) QUALITY METRICS OF DIGITALLY DERIVED IMAGERY AND THEIR RELATION TO INTERPRETER PERFORMANCE: II. EFFECTS OF BLUR AND NOISE ON HARD-COPY INTERPRETABILITY.	5. TYPE OF REPORT & PERIOD COVERED ANNUAL 1 Oct. 79 - 30 Sept. 80 6. PERFORMING ORG. REPORT NUMBER HFL 81-2
7. AUTHOR(s) Harry L. Snyder, Ph.D. James A. Turpin, M.S. Michael E. Maddox, Ph.D.	8. CONTRACT OR GRANT NUMBER(s) F49620-78-C-0055
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 (NL) Bolling Air Force Base, D.C. 20332	12. REPORT DATE SEP 1980 13. NUMBER OF PAGES 42
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	Image interpretation
Human factors	Image processing
Image quality	

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

Ten medium scale (1:2700 to 1:4400) aerial photographs, typical of the imagery viewed by Air Force photointerpreters, were digitized to 4096 x 4096 files (20- μm aperture by 11 bits intensity) on a scanning microdensitometer. The image files were then multiplied by two Gaussian filter functions (Fourier domain) to yield two blurred and one "ground truth" level of each image and transformed to eight bits of intensity for output. One of four weightings of a 4096 x 4096 Gaussian noise file was added to each image file, yielding 3 Blur

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

82 06

x 5 Noise x 10 Image combinations (150 images, total). Positive transparencies then served as the database for an information extraction task.

Fifteen photointerpreters (PIs) from the 548th Reconnaisance Technical Group, Hickam AFB, Hawaii, served as subjects. Blur was a between-subjects variable with five PIs at each of three levels. Noise was a within-subjects variable (five levels). The Noise main effect was significant ($p < .01$). The Blur main effect and the Blur x Noise interaction were not found to be statistically significant, although the Blur main effect was of the expected form. The data were correlated with image quality scaling values from a separate study using the same images and PIs. A significant correlation was found ($r = .898$).

HFL-81-2

Quality Metrics Of Digitally Derived Imagery and Their Relation to Interpreter Performance: II. Effects of Blur and Noise on Hard-Copy Interpretability

ANNUAL TECHNICAL REPORT
for
Period Ending 30 September 1980

Harry L. Snyder, Ph.D.
James A. Turpin, M.S.
Michael E. Maddox, Ph.D.

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

Contract F49620-78-C-0055
Life Sciences Directorate
U.S. Air Force Office of Scientific Research
Bolling Air Force Base, D.C. 20332

PREFACE

This research was supported by Contract No. F49620-78-C-0055 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 authors gratefully acknowledge the assistance of Mr. David I. Shedivy during the data collection phase; Dr. James J. Burke and his assistants at the Optical Sciences Center, University of Arizona, for production of the database used in this research; Mr. Gilbert Kuperman, AMRL, for technical assistance; the staff and photointerpreters of the 548th Reconnaissance Technical Group during the data collection effort; and the staff of the 460th Reconnaissance Technical Squadron in image selection, development of scoring criteria, and blind scoring of the EEIs.

CONTENTS

Appendix

LIST OF TABLES

Table

LIST OF FIGURES

Figure

I. INTRODUCTION

STATEMENT OF THE PROBLEM

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 seen equal 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, that 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 that 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, rather than merely 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 that 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 information extraction experiment.

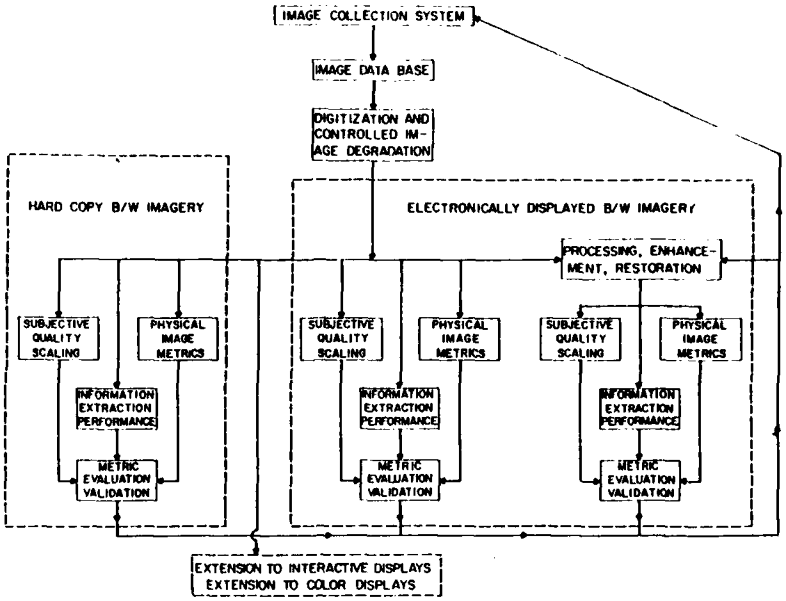
IMAGE COLLECTION SYSTEM
IMAGE DATA BASE
DIGITIZATION AND CONTROLLED IMAGE DEGRADATION
HARD COPY B/W IMAGERY
ELECTRONICALLY DISPLAYED B/W IMAGERY
PROCESSING, ENHANCEMENT, RESTORATION
SUBJECTIVE QUALITY SCALING
PHYSICAL IMAGE METRICS
INFORMATION EXTRACTION PERFORMANCE
METRIC EVALUATION VALIDATION
SUBJECTIVE QUALITY SCALING
PHYSICAL IMAGE METRICS
INFORMATION EXTRACTION PERFORMANCE
METRIC EVALUATION VALIDATION
SUBJECTIVE QUALITY SCALING
PHYSICAL IMAGE METRICS
INFORMATION EXTRACTION PERFORMANCE
METRIC EVALUATION VALIDATION
EXTENSION TO INTERACTIVE DISPLAYS
EXTENSION TO COLOR DISPLAYS

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 the program dealing with information extraction *performance from* the hard-copy imagery. It also addresses the question of how photointerpretation performance is affected by measurable physical properties of digitally derived imagery. Specifically, trained photointerpreters performed an information extraction task using images which were degraded by two known physical characteristics common to digitized aerial imagery: blur and noise. A parallel experiment assessing subjectively scaled quality of the same images is reported by Snyder, Shedivy, and Maddox (1981).

In addition to obtaining these important baseline performance data, the experiment also served to evaluate the information extraction methodology to be used in the subsequent soft-copy phases of the research program. Objectives of this methodology are described later.

II. METHOD

This report describes the first of two separate but related experiments. This first experiment provided the data necessary to determine the effects of the experimental variables, noise and blur, on the objective measure of information extraction performance. The second experiment used the same database (aerial photographs) to scale the subjective judgments of image quality by the photointerpreters. It is treated in detail in a separate report (Snyder et al., 1981).

There are three attributes of aerial photographs which are known to affect photointerpreter performance. These are blur, noise, and contrast. Blur and noise are known to degrade performance. Most photointerpreters request that their images be processed to a high, but constant, gamma. Therefore, they usually work with the same contrast. In addition, the study of contrast in a digital imaging system is not particularly meaningful, simply because digital systems have variable gain that "stretches" the contrast range in any scene to the maximum acceptable to the imaging system. Thus, contrast is typically never "lost" or attenuated by the acquisition process. For these reasons, only blur and noise are investigated in this experiment.

EXPERIMENTAL DESIGN

To examine the effects of noise and blur on photointerpreter performance, a factorial design was implemented. Noise and blur are within-subject and between-subjects variables, respectively; both are considered to be fixed effects (Figure 2).

BLUR, μm					
40	S_{1-5}	S_{1-5}	S_{1-5}	S_{1-5}	S_{1-5}
84	S_{6-10}	S_{6-10}	S_{6-10}	S_{6-10}	S_{6-10}
322	S_{11-15}	S_{11-15}	S_{11-15}	S_{11-15}	S_{11-15}
	12	24	42	60	75

SIGNAL-TO-NOISE LEVEL

Figure 2: Experimental design

Five of the fifteen subjects were randomly assigned to each of three blur levels (40, 84, and 322 μm blur). Each subject viewed two scenes at each noise level for a total of 10 scenes. The five noise levels were signal-to-noise ratios (SNRs) of 75, 60, 42, 24, and 12, where SNR is defined as the ratio of maximum intensity to RMS noise. Thus, for each of the blur levels, information was extracted from each scene at each noise level. The order of presentation of each unique scene/noise combination was randomized for each subject.

PHOTOINTERPRETERS

The 15 photointerpreters (5 female) were highly trained PIs serving with the 548th Reconnaissance Technical Group (RTG) at Hickam Air Force Base, Hawaii. Although the subjects served with the same command, 11 were U.S. Air Force interpreters, 3 were U.S. Army interpreters, and 1 was a U.S. Navy interpreter. However, all of the subjects received advanced training at the same USAF photointerpreter school.

The average age of the subjects was 24.9 years (S.D. = 2.85), with 3.9 years average experience (S.D. = 1.33).

APPARATUS

Equipment

Standard light tables (Richards Model 33H100) and binocular stereo zoom optics (Bausch and Lomb Zoom 70) were provided for the subjects. In addition, the subjects were allowed to use any other photointerpretation equipment of their choosing. (Many used hand-held magnifiers.)

Standard photointerpretation reference volumes were provided for the subjects as aids in the information extraction. Pen and paper were used to record the responses. All the equipment used in this experiment is common in this type of photointerpretation task.

Imagery

The original negatives (23 x 23 cm), from which the aerial photographic database was generated, were provided by the Environmental Research Institute of Michigan (ERIM). The original scenes were recorded by a K17 aerial camera. Three of the selected images were on Kodak Plus-X aerial film (Type 3401); the rest were on Kodak Tri-X aerographic film (Type 2403). Of the 10 scenes used in this study, 4 were of U.S. airfields; 4 more were U.S. naval installations; and the 2 remaining scenes were representative of electronic or aerospace research and development installations.

From the selected photographs, 7.6 x 7.6 cm areas were chosen as the desired targets. These areas were judged by the experimenters and senior photointerpreters at the 460th Reconnaissance Technical Squadron (RTS) at Langley AFB, Virginia, to be best suited for the information extraction task.

Personnel of the Optical Sciences Center at the University of Arizona scanned the target areas with a Perkin-Elmer PDS microdensitometer (Model No. 1010A). These scans contained 4096 x 4096 discrete points, each pixel representing a 20 μm square aperture, digitized to eight bits. The data files containing the scans constitute the uncorrupted ("ground truth") database for this study.

Measurements on the original images indicated that the scenes contained targets with variable spatial frequencies and contrast. The maximum (peak) intensity was adopted as the "signal". Noise (Gaussian) was added proportionally to the peak signal. All scenes scanned were found to yield a peak digital value of approximately 2000. Signal-to-noise ratios of 200, 100, 50, 25, and 12.5 were then selected as being representative of the range of typical operational imagery.

The original images, on computer tape, were read to a disk on a VAX 11/780 computer for further processing. The 4096 x 4096 pixel images were subjected to a Fast Fourier Transform and then multiplied by two appropriate Gaussian filter functions (same in each x,y dimension) which yielded the frequency spectrum of the blurred images. This process was then inverse Fourier-transformed to yield the data for the blurred images (i.e., removal of high frequency components is perceived as blurring). The nominal blur values are approximately equal to the width at half amplitude of the Gaussian blur distribution. For additional details see Burke and Strickland (1982).

A 4096 x 4096 noise file was created by scanning a Gaussian noise source (film) with the scanning microdensitometer. This file was then weighted appropriately and added to the scene and blur image combinations to produce the 150 images used for the experiment. This

150-image data set consisted of the factorial combination of 10 scenes by 3 blur levels (40, 84, 322 μm) by 5 signal-to-noise ratios (75, 60, 42, 24, and 12). These final signal-to-noise ratios are lower than those originally desired due to hard-copy playback difficulties, described in detail in a separate report (Burke and Strickland, 1982).

From the data files, that contained all combinations of blur-by-noise-by-scene, 150 original 7.6 x 7.6 cm images were produced with a Dicomed Model D-47 in a playback configuration using Kodak Type 2415 film processed in Kodak D-19 developer. Five reversal contact copies (Kodak Type SO-015 film) of each image (for a total of 750 images) were then produced and served as the imagery for the experiment. Figure 3 schematically depicts the image generation process that is described in greater detail by Burke and Strickland (1982).

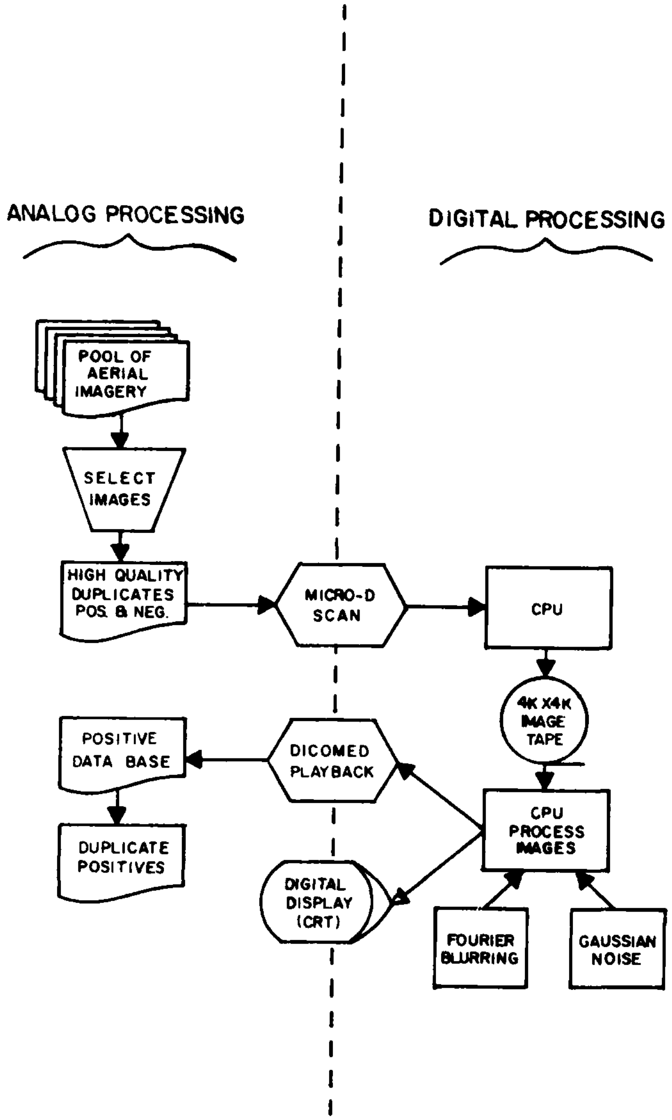

Figure 3: Generation of digitized hard-copy imagery

ESSENTIAL ELEMENTS OF INFORMATION (EEIS)

Each target area from the uncorrupted images was exploited by the panel of senior photointerpreters (460th RTS) to determine the "ground truth". That is, for each target, all Essential Elements of Information (EEIs) were determined and recorded.

The EEIs consisted of a hand written list of all information attainable from each image. The EEIs are presented in a format that includes varying degrees of image information (i.e., detection, identification, etc.), and is common to the daily task of image exploitation in the intelligence community. A sample blank EEI form is shown in Appendix A. The "ground truth" determined by the panel of interpreters serves as the key for scoring the performance measures.

PROCEDURE: DATA COLLECTION

Task

For each image, the PIs were required to complete the blank EEI form (Appendix A) by recording their data directly on the data sheets. There were no time constraints attached to this task.

Instructions

Each PI was required to read the set of instructions (Appendix B). After reading the instructions, PIs were offered the opportunity to ask the experimenter any questions and were presented an informed consent statement. Each PI understood his rights as a human subject, and indicated so by his signature. During the experiment, the PIs were allowed to ask the experimenter for any clarification of the instructions that were available throughout the task.

PROCEDURE: SCORING

All identification of PI or experimental level was removed from the data sheets. The coded EEIs were then transported to Langley AFB. At the 460th RTS, a panel of senior photointerpreters scored the data sheets. These judges also determined the percent of correct responses for each EEI.

Each response to each EEI was assigned a point total from zero (0) to nine (9). A score of "0" indicates that no information was obtained from the image. A score of "9" indicates that the photointerpreter cannot improve upon the answer. Scores between zero and nine indicate how well an interpreter satisfied the EEI based on the "known" or "ground truth" answer. It is important to point out that these scores are a reflection of how well an interpreter performed with an image as judged by senior photointerpreters. Thus, the measures of objective performance are in the context of the performance expected by the Air Force of its interpreters in their normal work setting. The points assigned to each EEI were, however, based upon a priori evaluations of each image. That is, the senior photointerpreter panel specified criteria for the 0 to 9 point scoring in advance, based on the EEI and the "ground truth". As a result, the scoring scheme is repeatable for new subjects, new answers, and new experimental conditions using the same scenes.

All point totals for the EEIs were normalized by image, and a percent correct for each image was determined. The percent correct scores provided the data for the subsequent analyses.

III. RESULTS

The arithmetic means of the percent correct scores of the two scenes (per PI per blur and noise combination) were subjected to a 3 x 5 analysis of variance. The results are shown in Table 1.

TABLE 1

Analysis of Variance Summary Table

Source	df	Mean Square	F	p
Blur (B)	2	119672.3333	2.27	0.146
Noise (N)	4	32023.8333	4.27	0.005
B x N	8	3869.8333	0.52	0.839
Subjects within Blur (S/B)	12	52833.6667		
N x S/B	48	7506.9999		
Total	74			

BLUR

The results of the analysis of variance indicated that the difference in information extraction performance attributed to blur is not statistically significant ($F_{2, 12}$= 2.27, $p < .10$). The mean scores across blur levels ranged from 53.4% (322 μm) through 63.3% (84 μm) to 66.5% (40 μm).

NOISE

The degradation of information extraction performance attributable to noise is statistically significant ($F_{4, 48}$ = 4.27, $p < .01$). In a general sense, scores collapsed across noise levels follow a meaningful ordering (Figure 4). The Newman-Keuls a posteriori test shows that the mean percent correct for the SNR of 12 was statistically less at the 5% confidence level than the percent correct for all other SNR levels, except 24. No other differences were significant ($p > .05$).

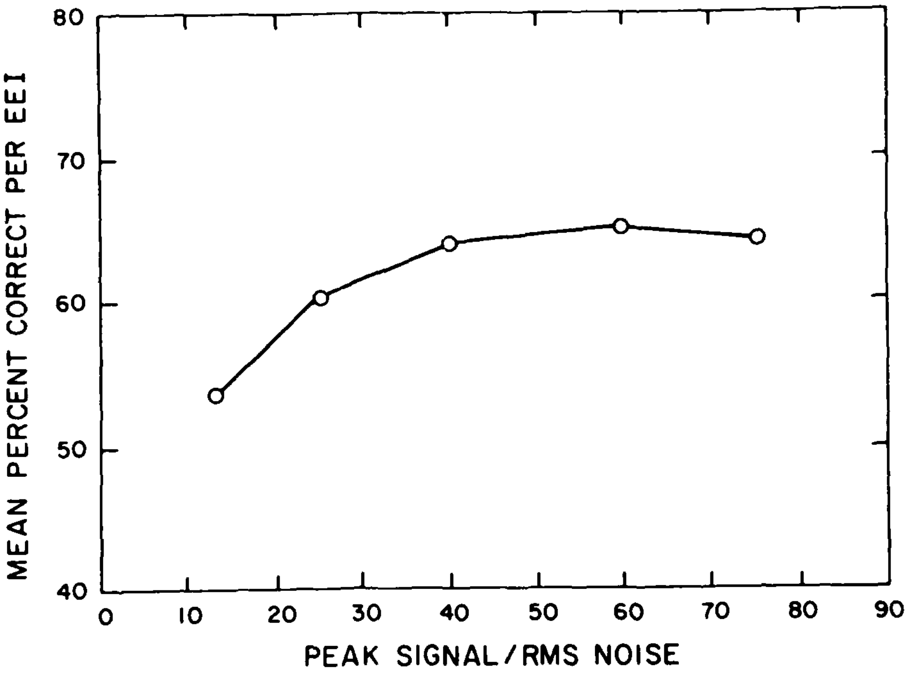

Figure 4: Effect of SNR on percent correct EEIs

BLUR X NOISE INTERACTION

The Noise x Blur interaction was not statistically significant.

CORRELATION WITH SCALING

In a separate study (Snyder et al., 1981), 14 of the 15 photointerpreters who participated in this study were subsequently asked to rate the same images on a scale of 0 to 9. The subjective scale was specifically designed to assign rating values to each image indicating the degree of perceived interpretability of that image. Figure 5 is a plot of the means for the rating scores versus the means from the EEI scores. The high linear correlation (r = .898) between the subjective scaling and the information extraction performance strongly suggests that information extraction performance can be predicted from image rating scores.

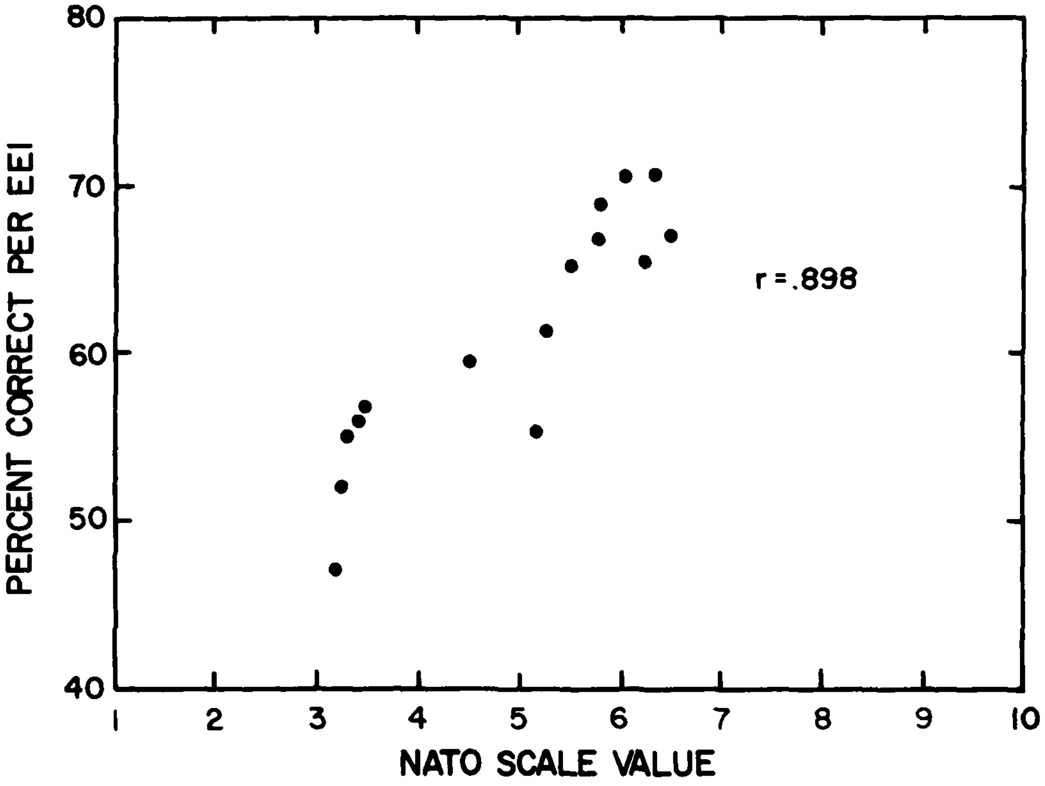

Figure 5: Correlation of scaling to EEI performance

IV. DISCUSSION

EFFECTS OF IMAGE VARIABLES

Photointerpreters are frequently confronted with imagery that, due to a loss of resolution (blurring or inappropriate scale factor, for instance), becomes difficult to exploit for specific information. However, PIs are generally well trained and highly motivated to perform to a level of successful exploitation limited only by the ground resolvable distance.

Exploitation of noisy imagery is an entirely different issue. The most common form of noise in aerial photography is excessive amounts of film grain. When photointerpreters receive images with moderate amounts of noise, they tend to look "through" the noise to detect and identify targets. Given that the noise is not extremely pronounced, its appearance is considered an annoyance but in general not a severe hindrance to sucessful information extraction. However, when the noise signal is sufficiently strong to mask targets, thus directly affecting resolution, the imagery is judged as less acceptable and often is not exploited.

Ideally, an experimental design treating noise and blur with equal emphasis would have been preferred. That is, a two-factor within-subject design would have enhanced the ability to detect performance differences attributable to the variables noise and blur, with each subject serving as his/her own control. However, a pool of 15 PIs for the experiment and 10 targets at 5 noise levels would have required 50 images for each blur level. Unlike other within-subject designs, repeated measures cannot be made on the same PI using the same target due to learning/memory effects. Thus, for each successive change in blur level, 10 new target areas would have been necessary. A within-subjects test of three blur levels would have accordingly increased the amount of well-controlled imagery by approximately 200%. The costs in time and money to maintain precision in production over 30 targets x 5 noise levels x 3 blur levels is infeasable. Further, early efforts in locating suitable high-quality imagery produced only approximately 20 useful scenes meeting the database requirements.

Considering these practical constraints of image production and availability, a mixed factors design was necessarily implemented. Allowing blur to be the between-subjects variable decreases the sensitivity of the study to small changes in performance attributable to blurring. Although it is generally accepted that more blurring in an image will cause smaller targets to become unidentifiable or even undetectable, and performance thereby degraded, a reliable decrease in performance was not detected at the 5 percent level of confidence.

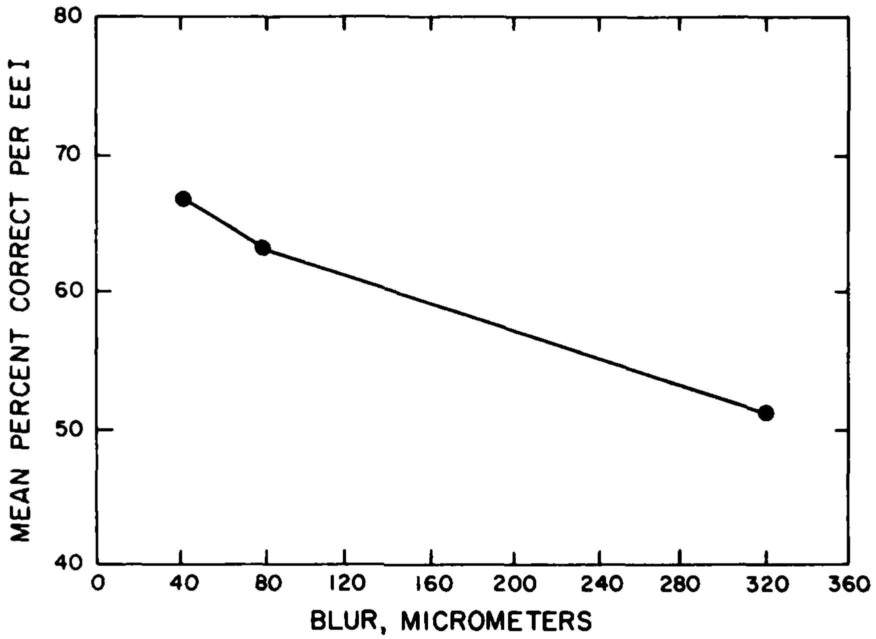

Figure 6: Effect of blur on percent correct EEIs

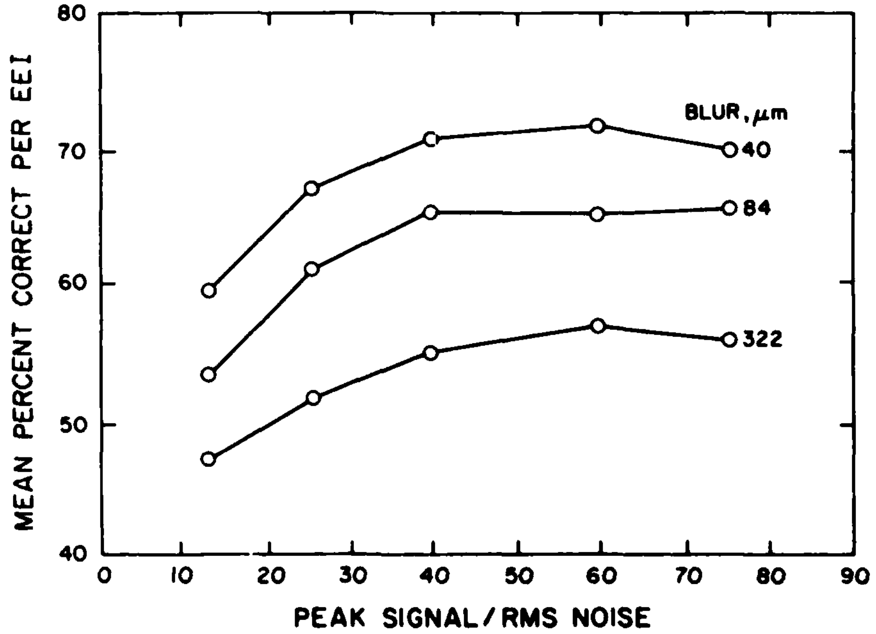

Figure 7: Effect of noise by blur interaction on percent correct EEIs

Non-significant findings are usually not reported in detail; however, a more complete discussion of the blur main effect appears warranted at this point. An examination of Figure 6 shows the expected decrease in performance with increased blurring. Plotting the Blur x Noise means also supports the argument that increases in blur degrade interpreter performance. Figure 7 graphically depicts this point. The means are well ordered except at the highest noise level (SNR of 12). At this level, scores for the 84 μm blur are higher than the scores for the 322 μm blur as expected, but also higher than the scores for the 40 μm level. It can be argued that this "medium" (84 μm) blur serves to "soften" the noise and, in fact, enhances a relatively noisy image, but that the 322 μm softens the noise too much, with an overwhelming loss of resolution.

The logical ordering of scores as a function of blur levels certainly seems to imply that the blur effect is "real", although not found to be statistically reliable. It is strongly suggested here that the relatively small number of degrees of freedom (2) used to test the blur effect, combined with high error variance between subjects, account for the failure to reject the hypothesis that blur has no effect on information extraction performance. It might also be noted that, while the blur effect variance (119672, Table 1) was 3.7 times greater than the noise effect variance (32024, Table 1), the error variance for blur was 7 times the error variance for noise (52834 and 7507 respectively, also Table 1).

SCORING

The scoring reflects the degree to which photointerpreters are expected to perform their normal duties. Among the many stringent requirements for sucessful interpretation of aerial imagery, accuracy and completeness are rarely compromised. It is generally accepted that accuracy and completeness on such a task can be measured by errors of omission and commission (Coluccio _et al_., 1969).

Such measures are easily attainable with a design which provides multiple-choice or forced-choice answers (Coluccio _et al_., 1969; Scott, 1968). However, this type of procedure is usually not practiced by military PIs. Rarely is the image content known before a given photograph is exploited. Thus, to keep the experiment within the context of the normal duties of an interpreter, satisfying the EEIs required open-ended answers typical of _operational_ photo interpretation.

The ability to assign objective performance scores becomes increasingly difficult with the degree of variability that accompanies open-ended answers. Thus, a stringent scoring scheme was developed by the panel of senior photointerpreters. The scoring scheme weighted each question in the EEIs by addressing the "most important" item. The highest weights were given to answers that satisfied the "most important" criteria. For example, counting vehicles (detection) was not weighted as heavily as the ability to recognize and record whether or not a particular vehicle was a troop carrier or a privately owned vehicle (identification). A stronger weighting for identification enhances the ability to isolate the effects of noise and blur on the full range of the photointerpretation task, yet relates to the intrinsic meaning of the PI's task.

It also follows that "careless" answers can bias a PI's score. For example, a PI may accurately identify scaffolding and engine stands adjacent to an aircraft, but when asked, "What type of maintenance is visible?" the PI may report, "None". Clearly, the PI identified the maintenance activity in a prior subsection, but failed to answer a different form of the same question. Assigning a low score to this type of mistake would not accurately reflect some aspect of image quality or performance related to the same. Such inconsistencies were also considered in the scoring procedure. Once rules were specified for scoring the data sheets, the task became somewhat simplified.

The degree to which judge biases affect scoring is controlled by removing all identification of subject and experimental conditions from the data sheets. Thus, subjective judgements made by the senior photointerpreters reflect only their expectation of normal interpretation reports that are reviewed daily. The now-existing detailed criteria per image, coupled with elaboration of the scoring "rules", will permit scoring of fudure EEI answers with the same imagery to remain comparable and useful. (It is for these reasons, including the cost of database preparation, that the EEIs and images are not contained in this or other technical reports. In a word, the database must remain secure to remain useful for future experimentation; however, it is available from the authors for use by other researchers.)

METHODOLOGY

One of the main objectives of this research was to establish a sensitive, natural, and repeatable methodology to measure the effects of digital image variables on PI performance. As described above, the task required by the methodology is quite similar to the usual tasks of the PI. No artificial questions/answers were used. Rather, the scoring technique was designed to assess those responses typical of the task and dictated by the image content. The scoring can be applied consistently and is sensitive to the image variables.

Further, it appears that the methodology is valid. PI comments on the image interpretation difficulty appeared to correlate well with the summary EEI means. In addition, the EEI means correlated well with mean NATO scale values that are accepted subjective measure of image interpretability within the PI community.

V. CONCLUSIONS

The degradation of performance attributable to the noise variable was significant. Also, the degradation of performance due to blurring was "graphically" demonstrated and believed to be more reliable than statistically shown.

Both forms of image degradation are known to affect information extraction performance with conventional photography. This study served to verify the fact that similar effects exist with digitized imagery.

The rigorous approach to developing a methodological procedure to produce these digitized images is outlined in a report by Burke and Snyder (1981). Sucessfully maintaining a secure and well defined database to be used in future studies epitomizes the ability to make sensitive performance measures with imagery. That is, future researchers who have access to this database can, after conducting appropriate studies, directly compare their performance results with the results obtained in this study. It has been the lack of secure and well-controlled databases that has prevented successful generalization from one image interpretation performance study to another. It is hoped that this study has circumvented that particular problem. In fact, there is ongoing research that will use this same database presented on soft-copy displays (CRTs). Direct comparisons will be made between the results of the studies.

The results of this study provide a valuable first step in relating blur and noise contcnt in digitally derived imagery to information extraction performance of professional PIs. Further, the high correlation between subjective ratings and PI performance suggest that much of the imagery used by the intelligence communities need only to be screened with a NATO-type scale to determine which imagery will produce the best information extraction performance.

Lastly, the data obtained in this study can and will be correlated with candidate physical metrics of image quality to permit an evaluation of alternate physical metrics. Ultimately, it is desirable to isolate or develop a physical metric of image quality that can be used for system design/evaluation as well as to predict PI performance.

REFERENCES

Burke, J.J. and Snyder, H.L. "Quality metrics of digitally derived imagery and their relation to interpreter performance: I. Preparation of a large scale database." S.P.I.E. Technical Digest, Image Analysis Techniques and Applications, January 6-9, 1981, 62-65.

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 81-2, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, 1982.

Coluccio, T.L., MacLeod, S., and Maier, J.J. "Effects of image contrast and resolution on photographic target detection and identification." Journal of the Optical Society of America, 1969, 59, 1478-1481.

Scott, F. "The search for a summary measure of image quality: A progress report." Photographic Science and Engineering, 1968, 12, 154-164.

Snyder, H.L., Shedivy, D.I., and Maddox, M.E. Quality metrics of digitally derived imagery and their relation to interpreter performance: III. Subjective scaling of hard-copy digital imagery. Technical Report HFL 81-3, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, 1981.

Appendix A

SAMPLE EEIS

This target area is part of a civilian airport used as a cargo facility. The photograph was taken from approximately 2000 ft.

1. Identify surface material of taxiway and parking areas; detect markings and repairs if any (patching or resurfacing).

2. Identify buildings by number, type, function, size (small, med, large), and serviceability.

3. Detect and identify aircraft by number, type, and state of repair.

4. Detect and identify all vehicles by number and type.

5. Determine loading activity of aircraft.

6. Detect any aircraft markings (numbers or letters).

Appendix B

INSTRUCTIONS

The study in which you are about to participate is being sponsored by the Air Force Office of Scientific Research in cooperation with the Tactical Air Command. This research is being conducted collaboratively by the Human Factors Laboratory at Virginia Tech and the Optical Sciences Center at the University of Arizona.

The overall purpose of the study is to compare the information extraction performance of Air Force photointerpreters on hard copy and soft copy (CRT) displays. We will also try to relate the PI performance to some objective and subjective measures of image quality.

The distinctive feature of the materials you will be seeing is that all the images have been generated digitally. That is, instead of continuous images, the transparencies are printed one small element at a time. These picture elements, or pixels, are printed approximately 4000 per row with about 4000 rows to yield a target area about 8 cm on a side.

During this particular phase of the research, we are interested in finding out how information extraction performance is related to certain image attributes, such as blur. We have prepared a series of transparencies from aerial photographs of various installations. All of the materials you will see have a consistent level of blur. The blur level you see may or may not be the same as the blur seen by others.

In addition to the blur present in the images, noise will be present in some of the scenes. The noise is usually caused by electrical receptor and transmission properties and it resembles very pronounced film grain. You will see only one noise level in each scene, but different scenes may have different levels.

You will see a total of ten images. The procedure for each image is identical. You will be given a 4" x 5" piece of film with the target area centered on it. You will also be given a print of the overall scene from which the target area was taken. This overall view is for reference only and should <u>not</u> be exploited for information.

Along with the transparency you will be given an information sheet with certain essential elements of information (EEIs) that are to be satisfied as completely as possible. You may use any standard interpretation equipment. The area in which we will be working will contain a light table with zoom microscope, a tube magnifier, standard target keys, and copies of Jane's Aircraft and Ship catalogs. You may use any personal equipment you wish, such as your own tube magnifiers.

Each EEI should be satisfied as accurately and completely as possible, given the content and quality of a particular target area. In the space provided on the information extraction sheet, simply fill in the elements of information as you find them in the target area. Use standard names and modifiers. No mensuration is required. If size is requested, of buildings and hangars for instance, please describe the items as small, medium, large, etc. Please list elements of information to the level of detail possible. For example, if the type number of an aircraft is discernable, e.g., 747, DC-3, then use this descriptor on the information sheet. If only general features can be seen, e.g., twin-engine, swept-wing, then use these in the EEI.

If you need more space to list particular elements of information use the back of the information sheet. Any questions that might arise during the session can be answered by the experimenter. There is no time limit for this part of the study other than the overall limit of two days to interpret the full set of 10 transparencies. The experimenter will tell you when to begin. PLEASE NOTE: It is very important that you do not discuss this task or these images with anyone else. This research is very important to the Air Force and TAC. Serious design errors can be caused if the data from this study are invalidated by discussion among participants. This requirement is very important and cannot be overemphasized.