Developing an interpretability scale for motion imagery

John M. Irvine, MEMBER SPIE
Ana Ivelisse Aviles
David M. Cannon
Charles Fenimore
Donna S. Haverkamp
Steven A. Israel
Gary O'Brien
John Roberts, MEMBER SPIE
Science Applications National Corp.
20 Burlington Mall Road-Suite 130
Burlington, Massachusetts 01803
E-mail: john.m.irvine@saic.com

Abstract. The motion imagery community would benefit from standard measures for assessing image interpretability. The National Imagery Interpretability Rating Scale (NIIRS) has served as a community standard for still imagery, but no comparable scale exists for motion imagery. Several considerations unique to motion imagery indicate that the standard methodology employed in the past for NIIRS development may not be applicable or, at a minimum, requires modifications. The dynamic nature of motion imagery introduces a number of factors that do not affect the perceived interpretability of still imagery—namely target motion and camera motion. We conducted a series of evaluations to understand and quantify the effects of critical factors. This paper presents key findings about the relationship of perceived interpretability to ground sample distance, target motion, camera motion, and frame rate. Based on these findings, we modified the scale development methodology and validated the approach. The methodology adapts the standard NIIRS development procedures to the softcopy exploitation environment and focuses on image interpretation tasks that target the dynamic nature of motion imagery. This paper describes the proposed methodology, presents the findings from a methodology assessment evaluation, and offers recommendations for the full development of a scale for motion imagery.
[DOI: 10.1117/1.2801504]

Subject terms: NIIRS; image and video quality; image interpretability; frame rate; full motion imagery.

Paper 070118R received Feb. 9, 2007; revised manuscript received May 20, 2007; accepted for publication May 24, 2007; published online Nov. 16, 2007.

1 Introduction

The National Geospatial-Intelligence Agency has conducted research and development into the feasibility of developing an interpretability scale for motion imagery. The National Imagery Interpretability Rating Scale (NIIRS) is a quantification of image interpretability that has been embraced by the intelligence community for still imagery.[1–5] Each NIIRS level indicates the types of exploitation tasks an image can support based on the expert judgments of experienced analysts. Development of a NIIRS for a specific imaging modality rests on a perception-based approach.[1,5] Additional research has verified the relationship between NIIRS and performance of target detection tasks.[6,7] Accurate methods for predicting NIIRS from the sensor parameters and image acquisition conditions have been developed empirically and substantially increase the utility of NIIRS.[2,5,8] In exploring avenues for development of a similar metric for motion imagery, a clearer understanding of the factors that affect the perceived quality of motion imagery was needed.

Several studies have explored specific aspects of this problem, such as target motion, camera motion, and frame rate.[9–11] Each study involved imagery analysts performing specific tasks with motion imagery in accordance with a designed experimental plan that varied the factors of interest while controlling for other effects. The first study addressed target motion, camera motion, scene complexity and ground sample distance (GSD). The second evaluation explored performance of a range of image exploitation tasks and their relationship to GSD and frame rate. The third evaluation also considered performance of specific image exploitation tasks, concentrating on consistency across the analysts.

Based on the findings from these studies, we conducted an evaluation to test a specific evaluation methodology for the development of a NIIRS-like scale for motion imagery. The methodology proposed for the development of a motion imagery scale is based on the standard NIIRS development approach, but has been adapted to the combination of the softcopy environment and motion imagery. The findings of this evaluation are guiding the overall plan for developing a motion imagery quality metric. The methodology appears to be viable, as the findings presented here will show. Full scale implementation of this approach involves an ambitious program that will produce a NIIRS-like scale for motion imagery.

2 Background

The NIIRS provides a common framework for discussing the interpretability, or information potential, of imagery. NIIRS serves as a standardized indicator of image interpretability within the national security community. An image quality equation (IQE) provides a method for predicting the

0091-3286/2007/$25.00 © 2007 SPIE

NIIRS of an image based on sensor characteristics and the image acquisition conditions.[2,5,8] Together, the NIIRS and IQE are useful for:

- Communicating the relative usefulness of the imagery,
- Documenting requirements for imagery,
- Managing the tasking and collection of imagery,
- Assisting in the design and assessment of future imaging systems, and
- Measuring the performance of sensor systems and imagery exploitation devices.

The foundation for the NIIRS is that trained imagery analysts have consistent and repeatable perceptions about the interpretability of imagery. Development of a NIIRS-like scale for motion imagery depends on a similar demonstration of consistency among analysts. Furthermore, the original NIIRS methodology was implemented for still imagery in hardcopy. Modifications to handle softcopy display and the dynamic nature of motion imagery are necessary.

3 Perceptual Studies of Motion Imagery

A series of recent studies provide a basic understanding of the critical factors affecting perceived interpretability of motion imagery. The critical factors identified in our preliminary investigations are: the motion of the targets, motion of the camera, GSD (spatial resolution), frame rate (temporal resolution), and scene complexity. These factors have been explored and characterized in three evaluations.

We found that the major factors affecting the perceived interpretability are spatial resolution (GSD) and temporal resolution (frame rate). The analysts were consistent in their perceptions of motion imagery and the perceptions of performing common image exploitation tasks. This finding holds for both "static" tasks that might be performed with still imagery and for "dynamic" tasks that involve detection and recognition of activities and rely on the temporal information provided by motion imagery.

3.1 *Motion and Complexity*

This evaluation assessed the effects of target motion, camera motion, scene complexity, and possible interactions among these factors.[9,10] The objective of this investigation is to develop an understanding of the effects of target motion, camera motion, and scene complexity in the perception of image quality and interpretability. The evaluation tasks were:

- Ratings of each motion imagery clip and corresponding frames of still imagery using the Visible NIIRS
- Paired comparisons of motion imagery clips to still images extracted from the same clips
- A set of paired comparisons of diverse motion imagery clips to assess the effects of target motion, scene complexity, and the interactions between the two.

The concept underlying the Visible NIIRS is that imagery analysts should be able to perform more demanding interpretation tasks on imagery that is of higher interpretability. The NIIRS consists of ten graduated levels (0–9), with several interpretation tasks or criteria forming each level. These criteria indicate the amount of information that can be extracted from an image at a given interpretability level. With a Visible NIIRS 2 (panchromatic) image, for example, analysts should just be able to detect *large buildings*, while on NIIRS 6 imagery they should just be able to *identify automobiles as sedans or station wagons*.[1–3]

Table 1 Characteristics represented in each GSD bin.

Target motion	Scene complexity	Camera motion
Low	Low	Low
Low	High	Low
High	Low	Low
High	High	Low
High	High	High

We conducted an evaluation with imagery analysts to address these fundamental issues using a set of 35 motion imagery clips. For each pairwise comparison, the analyst was asked to indicate the relative image interpretability for the two clips using a ratio scale. In addition, analysts were asked to rate each video clip and corresponding still images using the current Visible NIIRS.

The image set for the evaluation was populated with existing holdings at the National Geospatial-Intelligence Agency's Persistent Surveillance Office and special collections by the National Institute of Standards and Technology. The imagery used in this evaluation was high definition television data collected from a single 720 × 1280 progressive scan camera system. While the ultimate development of a motion imagery quality metric must embrace a range of camera systems and imaging conditions, this effort focuses on understanding specific effects related to perceptions of motion imagery. Consequently, the relatively limited range of image conditions effectively controls for a number of factors that might otherwise confound the effects of interest in this study.

All imagery was characterized with respect to target motion, camera motion, and scene complexity. Each clip was rated from 1 (low) to 5 (high) with respect to each of these factors. The ratings are subjective, based on the following definitions:

- Target Motion: The targets (usually vehicle or people in the scene) are moving with respect to the background and/or the raster
- Camera Motion: The camera is moving with respect to the background
- Scene Complexity: High complexity scenes include diverse clutter, multiple independent motions, higher spatial frequency information with their distribution across the image plane, target confusers, partial obscuration, or other features that make it difficult for an observer to detect and track the targets.

In addition, the ground sample distance (GSD) was estimated via mensuration of known objects in the scene and, where possible, validated by comparison to metadata. From

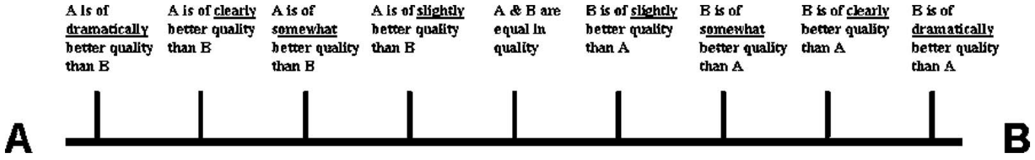

A is of dramatically better quality than B
A is of clearly better quality than B
A is of somewhat better quality than B
A is of slightly better quality than B
A & B are equal in quality
B is of slightly better quality than A
B is of somewhat better quality than A
B is of clearly better quality than A
B is of dramatically better quality than A
A
B

Fig. 1 Format for capturing paired comparison quality ratings.

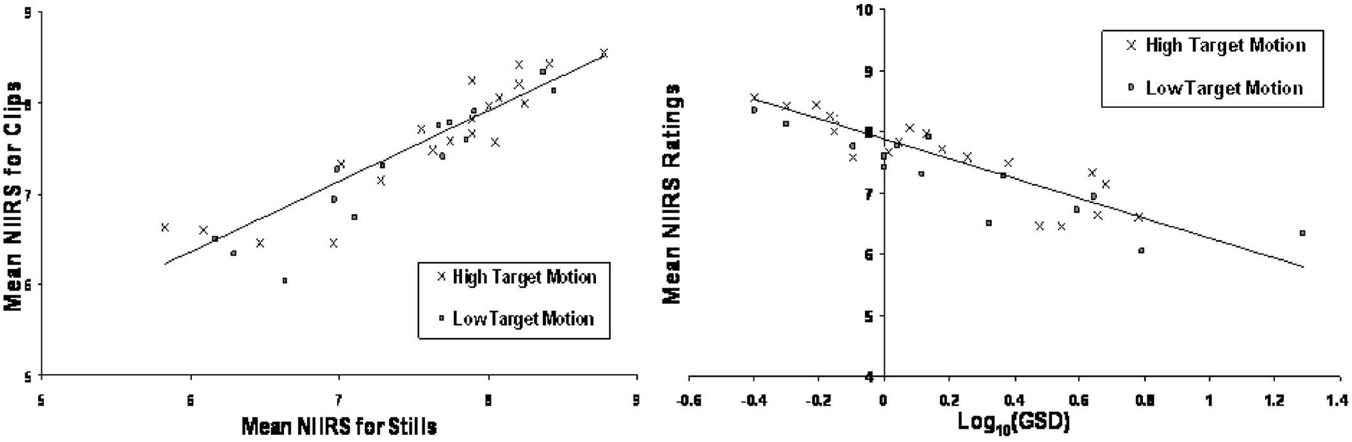
a full database of several hundred motion imagery clips, a set of 35 clips was selected for the evaluation. These clips were grouped into bins of similar GSD, where each grouping spanned five combinations of conditions (Table 1). Since camera zoom radically affects interpretability, clip selection required that the zoom remains essentially constant over the entire duration of the clip. Under this condition, a clip can then be assigned a GSD which varies minimally across frames. The unbalanced design arose from the limitations of the available imagery. From each clip, a high quality still image was extracted by applying a super-resolution process to five consecutive frames. The motivation was to correct for possible noise effects present in a single frame. However, visual inspection indicated that for these clips, the super-resolution product was almost indistinguishable from a single frame. Thus, the full set of imagery consisted of 35 video clips of approximately 5 s in length and 35 corresponding still images.

Twelve image analysts (IAs) participated in the evaluation. All of the analysts had experience with operational exploitation of imagery and were NIIRS certified. Experience levels spanned a range from junior analysts to long-tenured senior ones. Following the initial introduction, each IA worked through the evaluation at his/her own pace, taking breaks as needed. All imagery was viewed on calibrated monitors under controlled lighting conditions. To facilitate display of motion imagery for paired comparisons, the setup used two PCs, each with a high-quality color monitor. All responses were recorded in hardcopy. At the end of the evaluation, each IA completed an exit questionnaire to provide subjective feedback. The four steps in the evaluation were:

1. Visible NIIRS ratings of still images that were extracted from each motion imagery clip
2. Visible NIIRS ratings of the motion imagery clips
3. Paired comparisons of the motion imagery clip to a single frame from the clip sequence using the rating scale shown in Fig. 1.
4. Paired comparisons between various pairs of motion imagery clips, also using the scale shown in Fig. 1.

Throughout the evaluation, target motion has a significant effect on perceived image quality, in terms of both NIIRS ratings and paired comparisons. Motion imagery clips in which the targets are moving are consistently rated higher. This result is not surprising, since motion increases target salience. It is interesting to note, however, that the effects due to camera motion were not statistically significant and there are only weak indications of an interaction effect involving target motion and scene complexity.

Steps 1 and 2 of the evaluation demonstrate that trained IAs are capable of providing consistent NIIRS ratings for motion imagery. On average, the NIIRS ratings for the motion imagery clips are slightly (0.28 NIIRS units) higher than for the corresponding still image [Fig. 2(a)]. This difference approaches statistical significance (t-statistic=1.91, p-value=0.066). Both the NIIRS ratings and the paired



(a) (b)

Fig. 2 (a) The graph on the left depicts the mean visible NIIRS ratings for the motion imagery clips compared to the mean visible NIIRS ratings for the corresponding still images, with the least squares line shown; (b) the graph on the right shows the relationship between the mean visible NIIRS rating of the motions imagery clip and the estimated $\log_{10}$(GSD) for the same clip, with the least squares line shown.

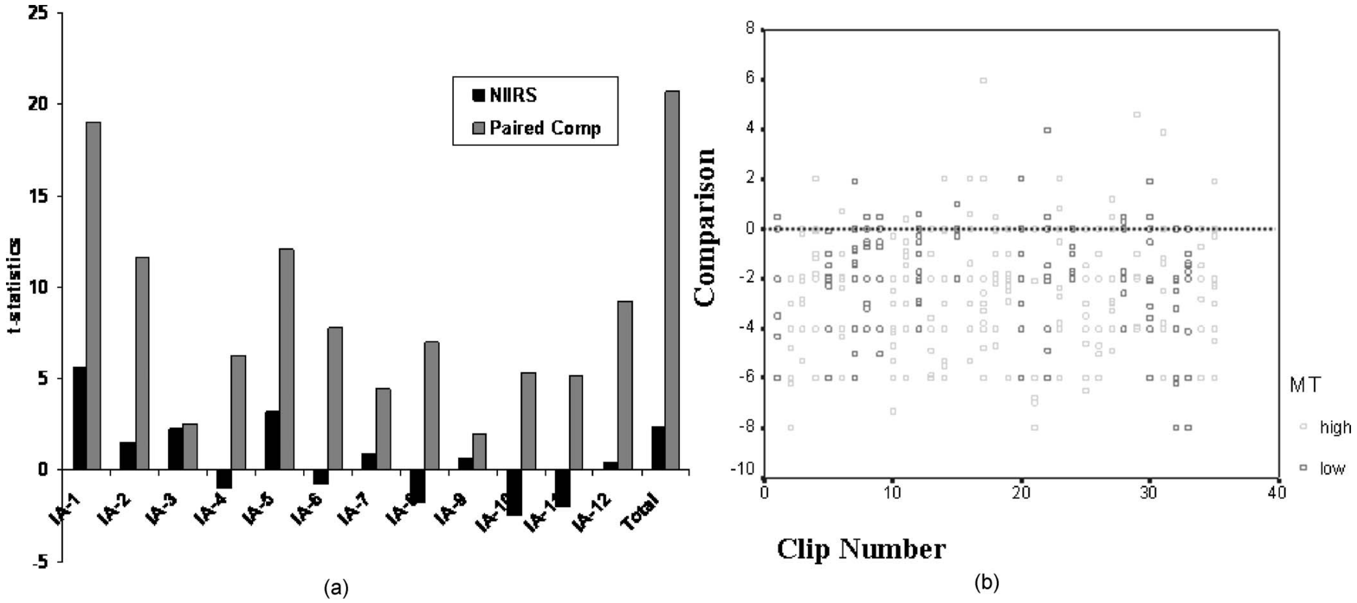


Fig. 3 (a) The *t*-statistics for testing a difference due to motion and (b) raw ratings from the paired comparisons of motion imagery clips to still images.

comparisons indicate that image interpretability is inversely related to $\log_{10}(\mathrm{GSD})$. In the regression analysis, $\log_{10}(\mathrm{GSD})$ accounts for more than 80% of the variance and the *t*-statistic on the regression coefficient is 10.71—very highly significant. While the relationship is linear, the slope is much lower than expected [Fig. 2(b)]. Historically, a doubling or halving of GSD produces a one NIIRS unit shift. These ratings exhibit about a half NIIRS unit shift when GSD varies by a factor of 2. This flatter relationship may be due to changes in the image associated with softcopy display or because the color imagery provides better target contrast than for panchromatic imagery.

Visible NIIRS ratings for the motion imagery clips are slightly, but statistically significantly, higher than for the corresponding NIIRS ratings of still images. The paired comparisons suggest that the perceived interpretability of motion imagery is considerably higher than for still images, but the Visible NIIRS is not sensitive to all the factors influencing the perceived interpretability of motion imagery. Figure 3 illustrates this point. On the left side, the bars represent the values of *t*-statistics to test for significant differences between motion imagery and still imagery. One set of bars are computed from the NIIRS ratings, i.e., the *t*-statistic arises from the paired test of NIIRS ratings for still images versus NIIRS ratings for the corresponding video clips. The other bars are the *t*-statistics computed from the paired comparisons of stills to video clips (step 3) and test for a significant difference from zero. Note that the bars based on the paired comparisons show a much stronger difference, indicating the motion imagery has much higher interpretability than the corresponding still frames.

3.2 *Evaluation of Image Exploitation Tasks*

Two evaluations assessed the ability of imagery analysts to perform various image exploitation tasks with motion imagery. The tasks included detection and recognition of objects, as might be performed with still imagery, and the detection and recognition of activities, which relies on the dynamic nature of motion imagery.[11]

In these two evaluations, trained imagery analysts rated their confidence in performing specific image exploitation tasks on a set of motion imagery clips. The tasks included things that could, in principle, be done with still imagery, such as detection and recognition of various targets or objects. Other tasks included in the evaluation were specific to motion imagery, focusing on detection and recognition of activities, e.g., loading versus unloading of cargo. In the remainder of this paper we use the term *static tasks* to refer to detection and recognition of objects and *dynamic tasks* to refer to detection and recognition of activities.

Two evaluations provided the data to address the objectives described above. The first evaluation varied a number of factors to provide an initial assessment of each of the study objectives. The second evaluation held frame rate constant at 30 frames per second (fps) in order to develop a more statistically robust assessment of consistency and the relationship between task satisfaction and GSD.

The initial image exploitation tasks were drawn from the set of tasks comprising the Visible NIIRS. These and other tasks used in previous NIIRS development efforts provided the static tasks. Development of dynamic tasks required identifying activities of interest on motion imagery. A list of approximately 50 dynamic tasks was compiled. These candidate tasks, both static and dynamic, were reviewed to select specific tasks for the evaluation that would span the range from easy to difficult. In addition, these tasks referenced common objects and activities that would be familiar to the analysts.

We used image clips that cover a range of GSDs and exhibit both low and high motion. To explore the interaction of these factors with frame rate, it is necessary to also cover a range of frame rates. We achieved this by perform-

Table 2 Imagery data for the evaluation

	Resolution		
	Coarse (60−100″ GSD)	Medium (10−60″ GSD)	Fine (1−10″ GSD)
Low target motion	30, 15, 1, and 0 fps	30, 15, 1, and 0 fps	30, 15, 1, and 0 fps
High target motion	30, 15, 1, and 0 fps	30, 15, 1, and 0 fps	30, 15, 1, and 0 fps

ing frame sampling on the selected clips. A large set of imagery (approximately 500 image clips) was reviewed with respect to motion content, complexity, and quality. The imagery selected for the evaluation ranged in GSD from 200 in. to approximately 1 in.. Each clip was 15 s long. Each of these primary clips was then sampled to produce 15 s renditions at 30, 15, and 1 fps. A single frame was selected from the middle of each 15 s clip to produce a still image for comparison, as well.

The first evaluation consisted of the assessment of each clip, at each of a number of frame rates and for a single frame, according to each task. For every clip at each frame rate, each analyst rated how effectively he/she believed each exploitation task could be performed. The ratings were on a scale of 0 (no confidence) to 100 (very high confidence). The work flow allowed the analyst to view the clip as many times as needed. The analyst would then review the first task make an assessment relative to that clip. After rating the first task, the analyst would go on to the second task and so on. Once all of the tasks were rated with respect to the first clip, the analyst would bring up the second clip and repeat the process, continuing until all tasks and been rated relative to all clips. The order of presentation of the clips was grouped into blocks by frame rate and randomized within each block. The order of the frame rates was counterbalanced across analysts and the order of tasks within each clip was randomized.

In addition, each full-frame rate (30 fps) clip and each still image (single frame) was rated with the Visible NIIRS. This allowed us to assess any correlation between NIIRS rating by the analyst and the confidence assigned to the tasks. Because both the clip and the still frame were NIIRS rated, we could examine the effects of motion on NIIRS. Thus, the evaluation itself consists of three distinct steps:

1. Assessment of all clips at each frame rate and still frame according to each criterion.
2. NIIRS ratings of the 30 fps clips.
3. NIIRS ratings of the still frames.

Because each task is rated relative to each clip, the number of ratings grows quickly with the number of tasks and clips. To keep the size of the evaluation manageable and to avoid fatigue, we constrained the evaluation to a relatively small number of clips and tasks. Six parent clips were used to generate renditions at 30, 15, and 1 fps, and the still image. Thus, 24 clips (six parent clips times four frame rates) were presented to the analysts. Three of the parent clips had high target motion and three had little or no target motion (Table 2). The spatial resolutions included coarse, medium, and fine GSDs. Fourteen exploitation tasks were used—seven static tasks and seven dynamic ones.

Evaluation 2 followed a very similar structure, but did not explore variations in frame rate. The goal for the second evaluation was to expand the number of criteria and the number of clips to provide a larger sample for assessments of rater consistency. The larger pool of data also supported more extensive comparisons of task satisfaction to GSD. This evaluation included 20 image clips, all viewed at 30 fps, and 20 image exploitation tasks. Eleven analysts participated in the evaluation.

The primary measure of performance was the confidence assigned to each task relative to each clip. The first evalu-

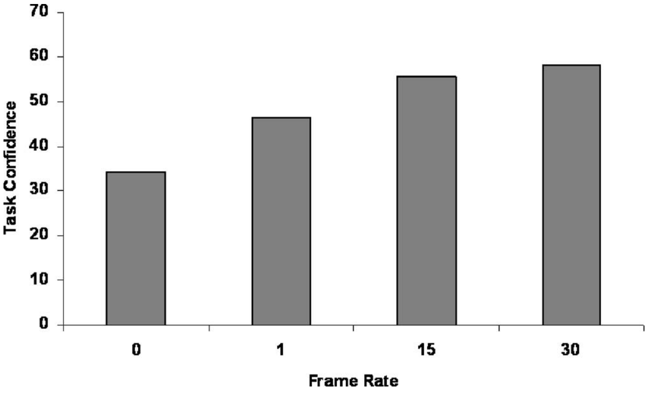


Fig. 4 Mean confidence ratings across all tasks and clip, by frame rate.

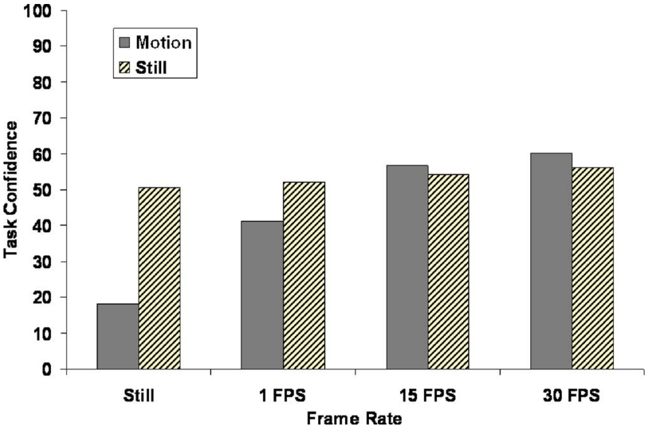


Fig. 5 Mean confidence ratings across all tasks and clip, by frame rate and task type.

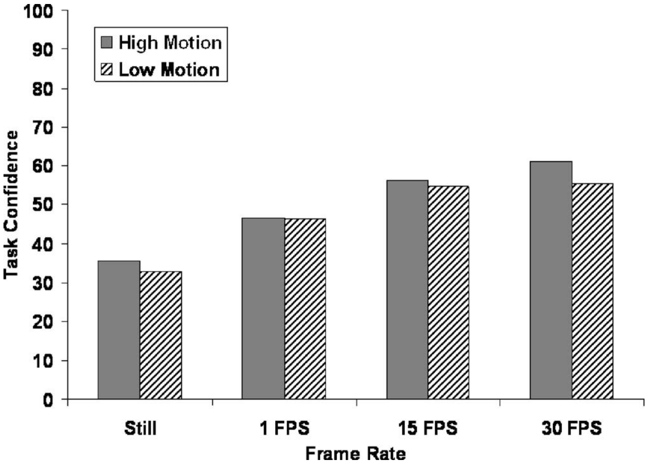


Fig. 6 Mean confidence ratings across all tasks and clip, by frame rate and target motion.

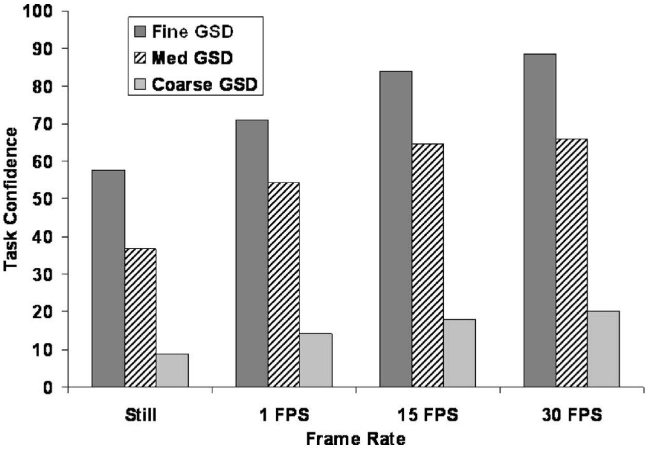


Fig. 7 Mean confidence ratings across all tasks and clip, by frame rate and GSD.

ation permitted an initial investigation of the effects of frame rate, scene content (high versus low target motion), and the nature of the task (static versus dynamic). The Visible NIIRS ratings provided additional information about perceptions of the motion imagery.

The results show that task confidence does vary with frame rate (Fig. 4). When the tasks are broken out by static and dynamic, the expected pattern emerges. Static tasks, which could in principle be performed with still imagery of sufficient quality, are insensitive to frame rate. The dynamic tasks, however, are sensitive to frame rate and confidence drops dramatically with reduced frame rate (Fig. 5). The pattern is not sensitive to the scene content in the sense of target motion. Task confidence exhibits the same pattern for both high-target motion and low-target motion clips (Fig. 6). This is a promising result in terms of scale development, because it suggests that analysts can assess information potential from a clip independent of the level of target motion. The pattern is also consistent across the three GSD bins represented in the data (Fig. 7) and consistent with other frame rate studies.[11,12] The analysis of variance (Table 3) demonstrates that GSD bin and frame have very highly statistically significant effects on the analysts' confidence in performing the image exploitation tasks. Furthermore, the frame rate by task type interaction term is also highly significant. The dynamic tasks are highly sensitive to frame rate, while static tasks are not.

Analysis of the Visible NIIRS ratings revealed the expected results. Visible NIIRS varies inversely with Log_{10}(GSD), as is also the case with still imagery (Figs. 8 and 9). The pattern is independent of the level of target motion in the clip (Fig. 8). It is also independent of frame rate (Fig. 9). Both of these findings should be expected, since the Visible NIIRS addresses exploitation tasks for still imagery.

The second evaluation provided a closer look at rater consistency and the relationship between task confidence and the properties of the image clip. In general, the analysts showed good agreement in their confidence ratings (Fig. 10). Correlations between an individual's ratings and the mean of the group were roughly 0.9. In addition, the confidence ratings exhibit the desired relationship with GSD. As Figs. 11 and 12 suggest, an "easy" task can be performed on most clips regardless of GSD.

These two evaluations provide valuable insight into analysts' perception of the interpretability of motion imagery. Following the approach used in the development of NIIRS, the interpretability can be defined by the types of image

Table 3 Analysis of variance for confidence ratings.

Source	Sum of squares	Degrees of freedom	Mean square	F-statistic	P value
Frame rate	204 248.5	3	68 082.8	56.54	<0.000005
Task type	49 711.3	1	49 711.3	41.29	<0.000005
GSD	1 543 782.1	1	1 543 782.1	1282.16	<0.000005
Frame rate by task type	123 801.4	3	41 267.1	34.27	<0.000005
Error	2 821 092.3	2343	1204.1		

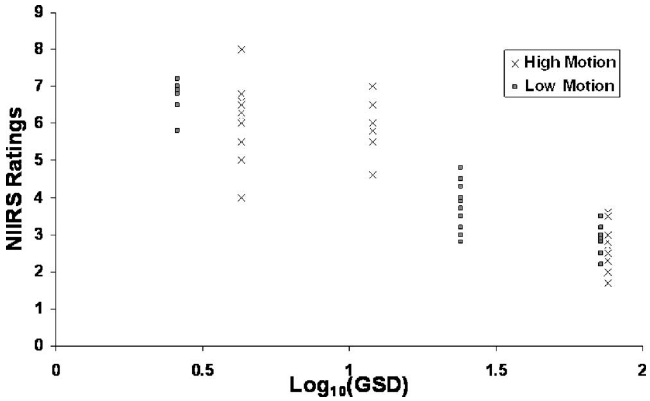


Fig. 8 NIIRS ratings by $\log_{10}$(GSD) and target motion (high vs low).

exploitation tasks that can be performed on a given clip. The findings from these two evaluations are consistent with this approach and point to three major conclusions:

1. The perceived ability to perform exploitation tasks depends primarily on the spatial resolution of the imagery, as measured by GSD. Log_{10}(GSD) accounts for about 80% of the variance in the ratings.
2. The ability to perform dynamic tasks, which involve detection and recognition of activities, depend on the temporal resolution, as measured by frame rate. Static tasks, however, are not sensitive to frame rate.
3. Confidence in performance exploitation tasks, whether static or dynamic, does not depend on the level of target motion in the scene.

These findings are consistent with the premise that development of a task-based scale for motion imagery is feasible. Based on these findings, we have conducted a pilot investigation of the scale development methodology before embarking on the full scale development effort.[13] Based on the experience to date, we conclude that development of a NIIRS-like scale for motion imagery is feasible and we hope to proceed with full scale development in the near future.

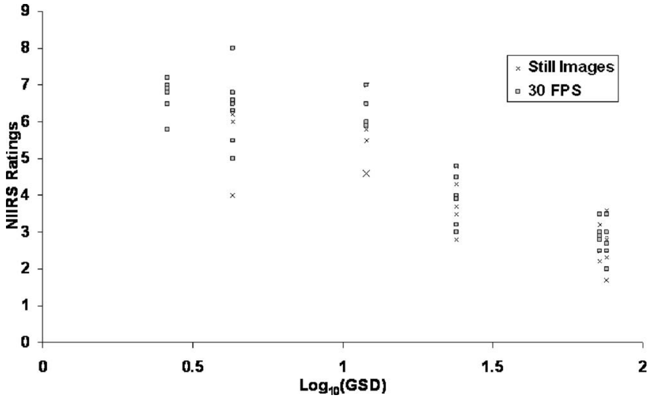


Fig. 9 NIIRS ratings by $\log_{10}$(GSD) and frame rate (30 fps vs still).

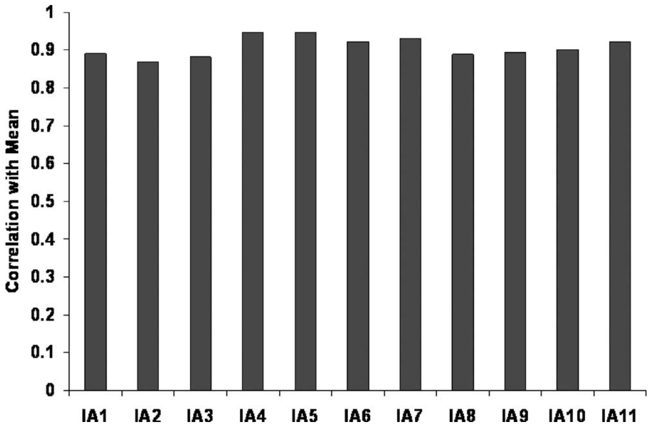


Fig. 10 Task confidence ratings for each analyst: Correlation with the mean. The high correlations (values close to one) demonstrate the consistency among the analysts.

4 Scale Development Methodology

The methodology for developing the NIIRS has been applied with multiple types of imagery and offers a robust approach to developing a scale.[1–5] The basis for the scale is that image exploitation tasks indicate the level of interpretability for imagery. If more challenging tasks can be performed with a given image, then the image is deemed to be of higher interpretability. A set of standard image exploitation tasks or "criteria" defines the levels of the scale. The purpose of the scale development methodology is to select "good" criteria to form the scale and to associate these criteria with the appropriate levels of image interpretability. Historically, the methodology has been performed with hardcopy image transparencies. The NIIRS development process involves five major steps:

- Image Scaling Evaluation: Analysts rate imagery of varying scene content and quality with respect to subjective image interpretability. The analysis of these ratings determines a set of marker images against which a set of image exploitation tasks will be rated.

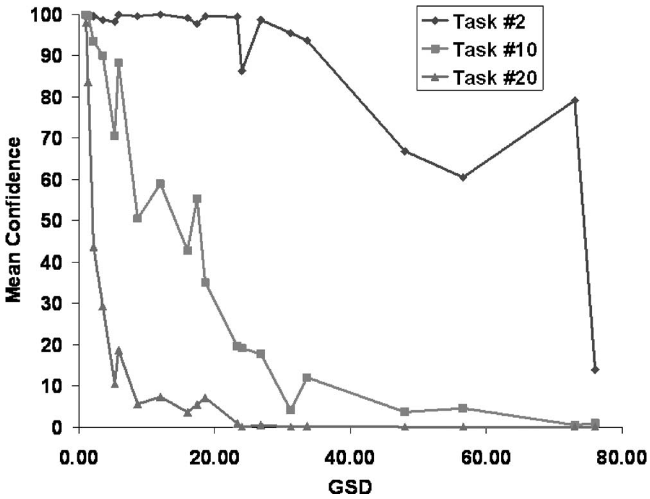


Fig. 11 Task confidence ratings for three (3) tasks relative to GSD for the image clips.

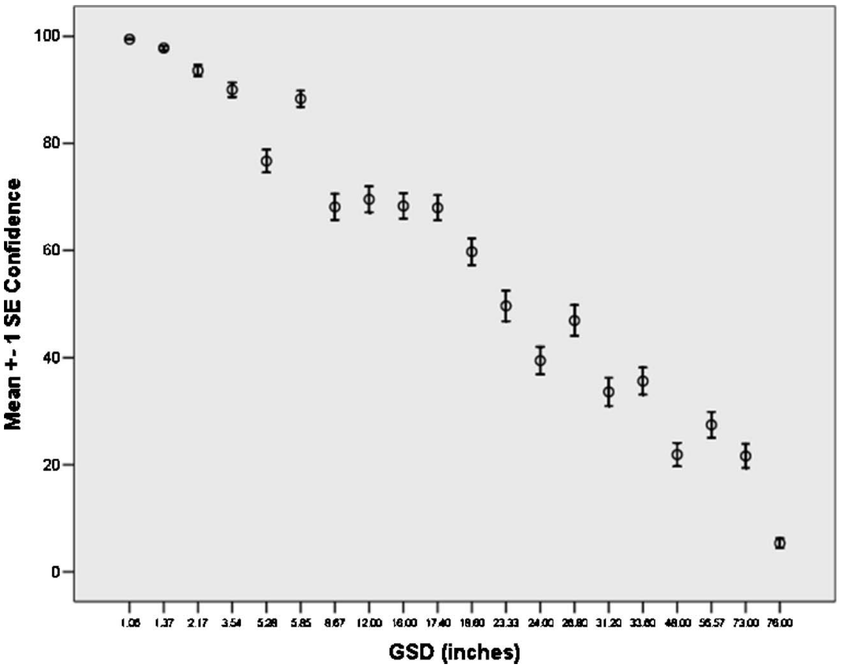


Fig. 12 Task confidence ratings across tasks relative to GSD for the image clips.

- Development of candidate criteria: Criteria are simple image exploitation tasks that are relevant to the analysts working with this type of imagery.
- Criteria Scaling Evaluation: Analysts rate the exploitation criteria relative to marker images that were selected based on analysis of the rating in the image scaling evaluation. This step links the criteria to the underlying perceptual quality scale that was implicitly defined by the analysts' ratings in the Image Scaling Evaluation.
- Construction of the actual scale: Using the data from the image and criteria scaling evaluations, specific criteria are selected to form each level of the scale.
- Scale Validation Evaluation: Analysts use the scale constructed from the criteria to rate imagery, in order to assess the properties of the scale.

4.1 *Design of the Evaluation*

To adapt the NIIRS development methodology to the softcopy environment with motion imagery, two modifications were necessary. First, the imagery was viewed in softcopy using a controlled viewing environment and standard image display software. Second, the criteria rating process has been modified slightly, because the previous approach would be unnecessarily cumbersome to employ with motion imagery. This evaluation used a limited set of motion imagery clips and criteria to assess the basic approach to scale development. The evaluation consisted of two steps:

1. Subjective ratings of the interpretability of motion imagery clips.
2. Subjective ratings of the difficulty associated with specific image exploitation tasks (criteria).

In the first step, two clips were designated as references. One clip was assigned a subjective interpretability rating of 0 (zero) and the other clip was assigned a subjective interpretability rating of 100. Each IA first reviewed these two reference clips to become familiar with their level of interpretability. The IA then reviewed the remaining clips and rated each one on the 0–100 scale. The order of presentation was randomized. The IA rated each clip according to its interpretability, keeping in mind clips that define 0 and 100 on this subjective scale. Ratings below 0 or above 100 were permitted if the IA felt that the clips were worse (of lower interpretability) than the "0" marker or better (of higher interpretability) than the "100" marker. After the IAs assigned their initial ratings, they were permitted to cycle through the clips a second time and change any ratings they thought were inappropriate.

The second step in the evaluation was a set of criteria ratings. In this step, five imagery clips served as the markers. The IA began by reviewing the five marker clips to become familiar with them. Then the criteria were presented in a randomized order. The IA assessed whether the exploitation task described by each criterion was achievable on a clip with an interpretability level comparable to each marker clip. The IA identified the two marker clips that "bounded" the criteria, i.e., the highest marker on which the task could be performed and the lowest marker on which the task cannot be performed. They noted the two markers and then rated the position of the criteria relative to the two markers (Fig. 13).

4.2 *Analysis and Results*

The goal of the data analysis was to verify that the methodology employed in this evaluation is viable for construct-

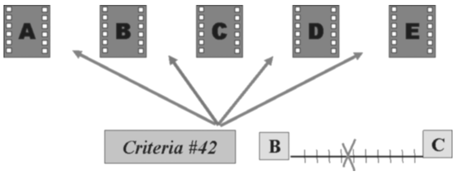


Fig. 13 Depiction of the criteria rating process.

ing a NIIRS-like scale for motion imagery.[13] The analysis addresses both steps in the evaluation—image ratings and criteria ratings.

4.2.1 *Imagery ratings*

For the imagery ratings, the fundamental question is whether analysts perceive the interpretability of the imagery clips to be well defined along a single primary dimension. The primary indicator will be the consistency of ratings across analysts and indications that the ratings are dominated by a single dimension which defines the underlying perceptual scale. The analysis will assess the following issues:

- Presence of outliers in the ratings
- Variance across IAs of the ratings of individual clips
- Principle component analysis of the ratings
- Relationship between ratings and GSD of the clips
- Effects due to the experience and backgrounds of the IAs, i.e., level of experience with motion imagery
- Effects due to target motion in the clip.

The initial data screening revealed no outliers or anomalous data. The analysts exhibited good rater consistency, with correlations between each rater and the mean of the group typically around 0.9–0.95 (Fig. 14). Principal component analysis also confirmed that the ratings fall along a single primary dimension. The first principal component accounts for more that 86 percent of the variance.

The ratings also show a strong linear relationship with $\mathrm{Log}_{10}(\mathrm{GSD})$ (Fig. 15). Regression analysis provided an estimated relationship that is consistent with other NIIRS investigations. $\mathrm{Log}_{10}(\mathrm{GSD})$ accounts for about 90 percent of the variance in the ratings.

Previous evaluations[9,10] have indicated small, but significant perceptual effects due to target motion. In previous studies, clips with high target motion have been rated slightly higher in interpretability, compared to clips with little or no target motion. We investigated the same issue with the current data set and found no significant effects. The levels of target and camera motion were rated (high, medium, and low) for each clip and the ratings were incorporated into a stepwise regression analysis. The dependent variable was the mean rating for the clip and the candidate independent variables were $\mathrm{Log}_{10}(\mathrm{GSD})$, the rating of target motion, and the rating of camera motion. Only $\mathrm{Log}_{10}(\mathrm{GSD})$ provided an explanatory power; the ratings of target and camera motion were not statistically significant.

4.2.2 *Criteria ratings*

The criteria ratings provide the link between perceived interpretability of motion imagery and the ability to perform specific image exploitation tasks. To form a scale, the ratings must be consistent across analysts, indicating that IAs share a common understanding of the tasks and how these tasks relate to the interpretability of the clips. The analysis will assess the following issues:

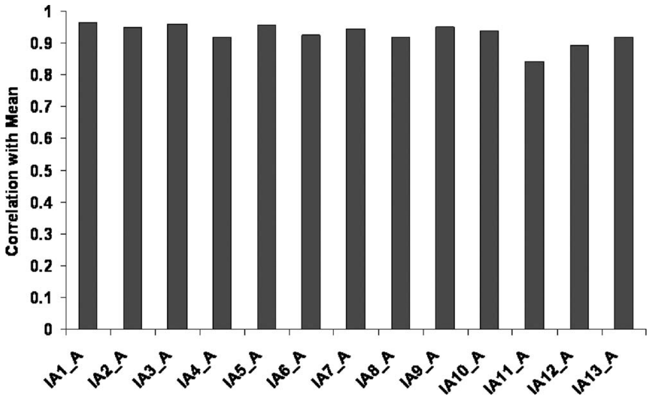


Fig. 14 Consistency of imagery ratings. The high correlations (values close to one) demonstrate the consistency among the analysts.

- Presence of outliers in the ratings
- Variance across IAs of the ratings of individual criteria
- Principle component analysis of the ratings
- Effects due to the experience and backgrounds of the IAs, i.e., level of experience with motion imagery
- Effects due to the dynamic nature of the criteria, i.e., “static” versus “dynamic” tasks.

The first step in the analysis was to convert the analysts’ responses to numerical ratings. Each marker clip was assigned a nominal position on the 0–100 scale (Table 4). The rating assigned by each analyst indicated the clips that “bounded” the exploitation task and the relative position between the marker clips. The numerical values were derived by interpolating between the corresponding marker values. A task, for example, that was judged to be midway between markers B and C would receive a numerical rating of 37.5.

Once the responses were converted to the numerical scale, analysis proceeded in a manner similar to Step 1. As with Step 1, the data screening revealed no outliers or anomalies. Subsequent analysis shows that the ratings were consistent (Fig. 16). Correlations between individuals and the overall mean were, once again, in the 0.9–0.95 range.

Table 4 Scale values assigned to marker clips

Marker clip	Scale value
A	0
B	25
C	50
D	75
E	100

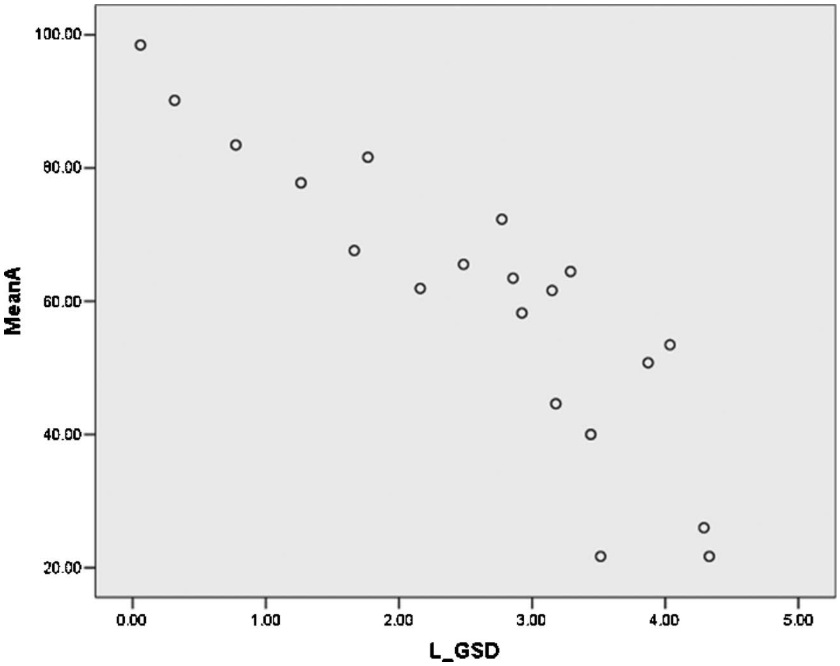


Fig. 15 Relationship between imagery ratings and $\log_{10}$(GSD).

Principal component analysis yielded similar results with the first principal component accounting for about 86 percent of the overall variance in the data.

5 Conclusions and Future Directions

The results of these investigations indicate that the proposed methodology is appropriate for development of a NIIRS-like scale for motion imagery. Imagery analysts perceive interpretability of motion imagery as one dimensional. This underlying perception of interpretability for motion imagery also correlates closely with GSD for the clip, which is consistent with previous NIIRS development efforts. Rater variability also was relatively low, indicating the IAs are consistent in their assessments of the imagery.

The results also imply that the scale development methodology is viable. IAs were consistent in their ratings of the exploitation tasks and rater variability was small enough to support construction of a scale with perceptually separable levels. The image exploitation tasks for the methodology validation evaluation focused almost exclusively on detection and recognition of activities, which would be sensitive to the dynamic nature of motion imagery. The earlier evaluations demonstrated consistent ratings of such tasks, but show that these types of tasks are sensitive to frame rate.

The findings form the basis for plans to develop a NIIRS-like scale for motion imagery. The tasks that comprise the scale will address detection and recognition of activities and, hence, will exploit the dynamic nature of motion imagery. Based on the results presented here we expect the initial scale to be closely related to Log_{10}(GSD) and to be sensitive to frame rate. In this manner, the major drivers of perceived interpretability are expected to be spatial and temporal resolution. Other studies[8,9] also suggest the target contrast will play an important role. The next steps in the scale development process are to conduct large image rating and task rating evaluations along the lines used in this study. These two evaluations will provide the basic data from which a draft scale can be constructed. The final step, of course, will be validation of the scale.

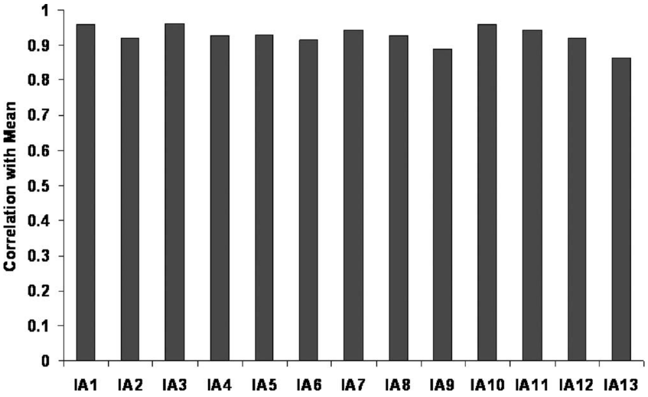
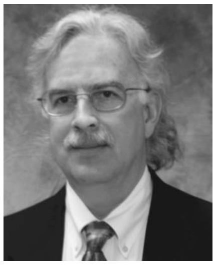


Fig. 16 Consistency of the criteria ratings. The high correlations (values close to one) demonstrate the consistency among the analysts.

References

1. J. M. Irvine, “National imagery interpretability rating scales (NIIRS): Overview and methodology,” *Proc. SPIE* **3128**, 93–103 (July 1997).
2. J. M. Irvine, National imagery intelligence rating scale (NIIRS) in *The Encyclopedia of Optical Engineering*, R. G. Driggers, Ed., Dekker, New York (2003).
3. J. C. Leachtenauer, “National imagery interpretability rating scales: overview and product description,” *Proceedings of the American Society of Photogrammetry and Remote Sensing Annual Meetings*, Baltimore, MD (April 1996).
4. L. A. Maver, C. D. Erdman, and K. Riehl, “Imagery interpretability rating scales,” *Digest of Technical Papers, International Symposium Society for Information Display*, Santa Ana, CA, Vol. **XXVI**, pp. 117–120, (May 1995).
5. J. C. Leachtenauer and R. G. Driggers, *Surveillance and Reconnaissance Systems: Modeling and Performance Prediction*, Artech House,

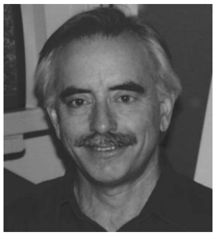
Norwood, MA (2001).
6. R. G. Driggers, P. G. Cox, and M. Kelley, "National imagery interpretation rating system and the probabilities of detection, recognition, and identification," *Opt. Eng.* **36**(7), 1952–1959 (1997).
7. R. G. Driggers, P. G. Cox, J. Leachtenauer, R. Vollmerhausen, and D. A. Scribner, "Targeting and intelligence electro-optical recognition and modeling: A juxtaposition of the probabilities of discrimination and the general image quality equation," *Opt. Eng.* **37**(3), 789–797 (1998).
8. J. C. Leachtenauer, W. Malila, J. M. Irvine, L. Colburn, and N. Salvaggio, "General image-quality equation: GIQE," *Appl. Opt.* **36**, 8322–8328 (1997).
9. J. M. Irvine, C. Fenimore, D. Cannon, J. Roberts, S. A. Israel, L. Siman, C. Watts, J. D. Miller, A. Ivelisse Avilés, P. F. Tighe, R. J. Behrens, M. Brennan, and D. S. Haverkamp, "Factors affecting development of a motion imagery quality metric," *SPIE Defense and Security Symposium*, Orlando, FL (Mar. 2005).
10. J. M. Irvine, C. Fenimore, D. Cannon, J. Roberts, S. A. Israel, L. Simon, C. Watts, J. D. Miller, A. Ivelisse Avilés, P. F. Tighe, and R. J. Behrens, "Feasibility study for the development of a motion imagery quality metric," *33rd Applied Imagery and Pattern Recognition Workshop: Image and Data Fusion*, IEEE Computer Society, Washington (Oct. 2004).
11. C. Fenimore, J. Irvine, D. Cannon, J. Roberts, I. Aviles, S. Israel, M. Brennan, L. Simon, J. Miller, D. Haverkamp, P. F. Tighe, and M. Gross, "Perceptual study of the impact of varying frame rate on motion imagery interpretability," *SPIE Conference on Human Vision and Electronic Imaging XI*, San Jose, CA, SPIE 6057–17 (Jan. 2006).
12. J. M. Irvine, C. Fenimore, D. Cannon, D. Haverkamp, J. Roberts, S. A. Israel, L. Simon, J. Miller, A. I. Avilés, and M. Brennan, "Development of a motion imagery quality metric," *Proceedings of the American Society for Photogrammetry and Remote Sensing (ASPRS) Annual Meeting*, Reno, NV (May 2006).
13. J. M. Irvine, D. Cannon, J. Miller, J. Bartolucci, L. Gibson, C. Fenimore, J. Roberts, I. Aviles, M. Brennan, A. Bozell, L. Simon, and S. A. Israel, "Methodology study for development of a motion imagery quality metric," *SPIE Defense and Security Symposium*, Orlando, FL (April 2006).



John M. Irvine is a technical fellow and the deputy division manager for Systems and Technology at Science Applications International Corporation (SAIC). He serves as a senior scientist for the National Geospatial-Intelligence Agency (NGA), where he is responsible for development, evaluation, and prototyping of technology for automated and semi-automated target recognition and other tools for image exploitation. He was the principal investigator for the development of novel biometrics under DARPA's HumanID Program. He has managed and provided senior technical guidance for projects on data fusion, signal and image processing, and automated image exploitation for a range of customers. His areas of specialization include evaluation of assisted image exploitation systems, development and evaluation of ATR and image understanding technology, image interpretability scales, and image chain optimization for intelligence, surveillance, and reconnaissance (ISR) missions. Dr. Irvine holds a PhD in Mathematical Statistics from Yale University.

Ana Ivelisse Aviles is a mathematical statistician in the Information Technology Laboratory at the National Institute of Standards and Technology, Gaithersburg, MD, since 2001. Previously with McNeil Consumer Products Co. and LifeScan, Inc. National Science Foundation graduate fellow, 1997; Grant M. Mack Memorial Scholarship, American Council of the Blind-National Industries for the Blind, 1998; Floyd R. Cargill scholar, Illinois Council of the Blind, Facilitation Award, National Science Foundation, and R. A. Freund international scholar, American Society for Quality, 1999; Mary Natrella scholar, Quality and Productivity Section, American Statistical Association, Ellis R. Ott scholar, Statistics Div., ASQ, and Ford Foundation dissertation fellow, 2000; Summer Research Opportunities Program Alumni Achievement Award, Committee on Institutional Cooperation, 2004. She currently is an associate editor for the *Journal of the American Statistical Association* (JASA). She has served as awards chair, Section on Physical and Engineering Sciences, American Statistical Association (ASA) and ASA representative to the American Association for the Advancement of Science (AAAS), Section on Industrial Science and Technology, 2005 to 2008. Dr. Aviles holds a PhD in Industrial Engineering and Management Sciences from Northwestern University.

David M. Cannon is a technical director for Exploitation Technology for Systems and Technology at Science Applications International Corporation (SAIC). He led SAIC Motion Imagery (video) Research activities at the National Geospatial-Intelligence Agency (NGA). Activities included prototyping airborne video systems; video data collection; video, image processing and exploitation tool development and evaluation; and assessing motion imagery user interpretability in support of the intelligence, surveillance and reconnaissance (ISR) community. Mr. Cannon holds degrees in Physics, BS from Brigham Young University; Electrical Engineering, BSEE from the University of New Mexico; and Electro-Optics, MSEO from the University of Dayton.



Charles Fenimore leads the Motion Imagery Quality Lab at the National Institute of Standards and Technology (NIST). He is involved in the development of quality metrics, test methods and materials, and standards for motion imagery, moving pictures, and medical imagery. He has contributed to the development of standards and test materials for motion imagery processing and presentation, through the Motion Picture Experts Group (MPEG), the Advanced TV Systems Committee, the Society of Motion Picture and TV Engineers (SMPTE), and the U.S. government's Motion Imagery Standards Board. Fenimore is a leader in international tests of quality for MPEG's Advanced Video Codec. He has worked to catalyze the emerging digital cinema and HD-DVD industries by bringing together industry stakeholders at NIST, SMPTE, and SPIE Conferences. He is a recipient of the SMPTE Journal Best Paper Award and is a recent member of Sigma Xi's College of Distinguished Lecturers. Dr. Fenimore holds a PhD in Mathematics from Berkeley.

Donna S. Haverkamp received the PhD degree in Electrical Engineering from the University of Kansas, Lawrence, in 1997. Following graduation, she took a position with Harris Corporation in Melbourne, Florida, where she participated in research toward the development of 3-D site models and later served as principal investigator on a contract under DARPA's Dynamic DataBase (DDB) program. After joining Space Imaging in Thornton, Colorado, she conducted research resulting in commercial feature extraction software. She continued to work in industry, performing research and development in intelligent image understanding technologies for SAIC, as well, with whom she continues a professional relationship. After a number of years in the commercial world, she has returned to academia and is an assistant professor on the faculty of the Department of Electrical Engineering and Computer Science at the University of Kansas. Her research interests continue in image and video processing, computer vision, and artificial intelligence. Dr. Haverkamp is a member of Tau Beta Pi and Eta Kappa Nu.

Steven A. Israel is a Senior Image and Pattern Recognition Scientist at Science Applications International Corporation (SAIC). Dr. Israel analyzes non-traditional datasets for a number of government, military, and academic organizations. He received a BS (1987) and MS (1991) from the State University of New York-College of Environmental Science and Forestry. His PhD (1999) was granted from the Departments of Information Science and Surveying at the University of Otago, New Zealand. Dr. Israel's interests include image processing, biometrics, photogrammetry, rugby, and pattern recognition.

Gary O'Brien is a mathematician with Science Applications International Corporation (SAIC). He has taught at University of Texas at Austin and George Mason University. He spent 11 years as the U.S. Army's Night Vision and Electronic Sensors Directorate (the Night Vision Lab). While there he was involved with the development and evaluation of automatic target recognition algorithms. His area of interest is image understanding from both a human and automatic perspectives. Dr. O'Brien holds a PhD in Mathematics from SUNY at Binghamton.

John Roberts has conducted research on displays and display systems within the Information Technology Laboratory at the National Institute of Standards and Technology (NIST) since 1993, observing the ways that display system hardware and playback performance affect display performance, and the aspects of human perception that are important for design of displays and content. He is currently working on ways to characterize and measure the aspects of moving images that are important for human perception, including collaborative research with the National Geospatial-Intelligence Agency (NGA). His previous work includes the development and testing of hardware and software tools for DARPA for performance measurement of multiprocessor computer systems. He was also principal inventor of three tactile display technologies for the blind.