Quantifying Interpretability for Motion Imagery: Applications to Image Chain Analysis

John M. Irvine, David M. Cannon, Steven A. Israel, Gary O'Brien
SAIC
Burlington, MA, U.S.A.
John.M.Irvine@saic.com

Charles Fenimore, John Roberts, Ana Ivelisse Aviles
National Institute of Standards and Technology
Gaithersburg, MD, USA
fenimore@nist.gov

Abstract - The motion imagery community will benefit from the availability of 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. We conducted a series of user evaluations to understand and quantify the effects of critical factors affecting the perceived interpretability of motion imagery. These evaluations provide the basis for relating perceived image interpretability to image parameters, including ground sample distance (GSD) and frame rate. The first section of this paper presents the key findings from these studies. The second portion is a new study applying these methods to quantifying information loss due to compression of motion imagery. We conducted an evaluation of several methods for video compression (JPEG2000, MPEG-2, and H.264) at various bitrates. A set of objective image quality metrics (structural similarity, peak SNR, an edge localization metric, and edge strength) were computed for the parent video clip and the various compressed products. In an evaluation, imagery analysts rated each clip relative to image interpretability tasks. The analysis quantifies the interpretability loss associated with the various compression methods and bitrates. We present the evaluation results and explore their relationship to the objective image quality metrics. The findings indicate the compression rates at which image interpretability declines significantly. These findings have implications for sensor system design, systems architecture, and mission planning.

Keywords: motion imagery, image interpretability, NIIRS, perception, compression

1 Introduction

A multidisciplinary team 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, 2, 3, 4, 5]. Each NIIRS level indicates the types of exploitation tasks an image can support based on the expert judgments of experienced imagery analysts (IAs). Development of a NIIRS for a specific imaging modality rests on a perception-based approach [1]. Accurate methods for predicting NIIRS from the sensor parameters and image acquisition conditions have been developed empirically and substantially increase the utility of NIIRS [3, 4]. In exploring avenues for the 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, color, and frame rate [6, 7, 8, 9, 10]. This paper begins with a summary of the recent studies on motion image interpretability. Building on that foundation, we present an evaluation of image compression which illustrates the applicability of these techniques to image chain analysis.

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 community. An image quality equation (IQE) offers a method for predicting the NIIRS of an image based on sensor characteristics and the image acquisition conditions [3, 4]. 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 IAs have consistent and repeatable perceptions about the interpretability of imagery. The methodology for developing the NIIRS has been applied with multiple types of imagery and offers a robust approach to developing a scale. 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. A recent investigation has demonstrated the extension of the methodology to motion imagery in the softcopy environment [9]. 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

- 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 ratings 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.

Extension of the NIIRS development methodology requires some adaptation due to the dynamic nature of motion imagery and the range of factors affecting perceived interpretability [9].

3 Perception Studies

A series of studies provide a basic understanding of the critical factors affecting perceived interpretability of motion imagery. Factors affecting perceived interpretability of motion imagery include 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 a series of evaluations with trained imagery analysts [6, 7, 8, 9, 10]

- Motion and Complexity: These evaluations assessed the effects on perceived image quality of target motion, camera motion, scene complexity, and possible interactions among these factors.

- Criteria Satisfaction: 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 done with still imagery and the detection and recognition of activities, which relies on the dynamic nature of motion imagery.

- Frame Rate: These evaluations assessed target detection and identification and other image exploitation tasks as a function of frame rate and contrast, using both synthetic and measured imagery

3.1   Ground Sample Distance (GSD)

All of the evaluations show that perceived interpretability has a strong linear relationship with $Log_{10}(GSD)$ (Figure 1). Regression analysis provided an estimated relationship that is consistent with other NIIRS investigations. $Log_{10}(GSD)$ accounts for about 80 percent of the variance in the ratings (Table 1). A critical issue that has not been explored, however, is the effect of grazing angle on this relationship. For both still and motion imagery, images acquired close to nadir exhibit approximately the same GSD in both the X and Y directions, but at lower grazing angles the two GSD measurements can differ substantially.

Table 1. Regression Relationship with $Log_{10}(GSD)$

|	Coefficient	t-statistic	Tail Probability
Constant	84.37	55.13	> 0.00001
$Log_{10}(GSD)$	-29.65	-26.9	> 0.00001

$R^2 = 0.8$

3.2   Frame Rate

The frame rate evaluations focused on the ability of analysts to perform specific image exploitation tasks on clips of varying frame rate and GSD (table 2). 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 analyst viewed the clip as many times as needed. The analyst would then review the first task and 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 were rated relative to all clips. The order of clip presentation 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. The results show that confidence declines as a function of frame rate for dynamic tasks related to the analysis of activity, but not for static tasks found in the current (still imagery) NIIRS (Figure 2).
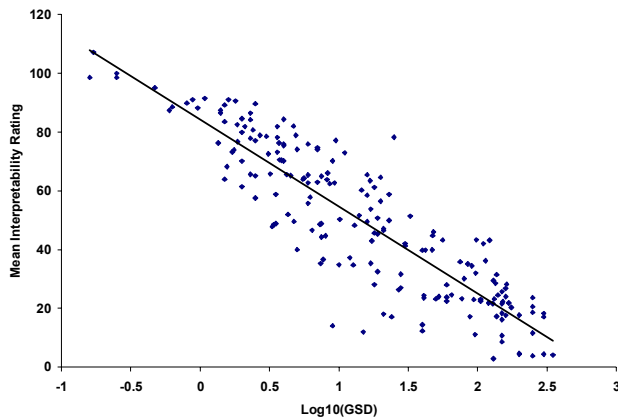
Figure 1. Comparison of Mean Ratings to Log₁₀(GSD)

Table 2. Imagery Data for Evaluation of Task Performance With Respect to Frame Rate and GSD

	Resolution		
	Coarse (60-100" GSD)	Medium (10-60" GSD)	Fine (1-10" GSD)
Low Target Motion	30, 15, 1, 0 fps	30, 15, 1, 0 fps	30, 15, 1, 0 fps
High Target Motion	30, 15, 1, 0 fps	30, 15, 1, 0 fps	30, 15, 1, 0 fps
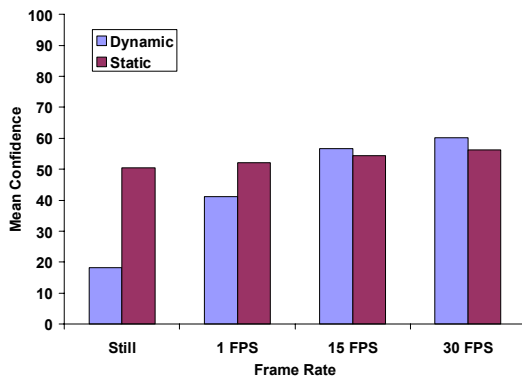
Figure 2. Mean Confidence Ratings By Frame Rate and Task Type

In a second evaluations of frame rates, imagery analysts performed target detection and target identification tasks on motion imagery clips presented at multiple frame rates. The evaluations used both synthetic and measured imagery. The findings were similar, with a decline in performance occurring between 5 and 15 frames per second, depending on the task (Figure 3).
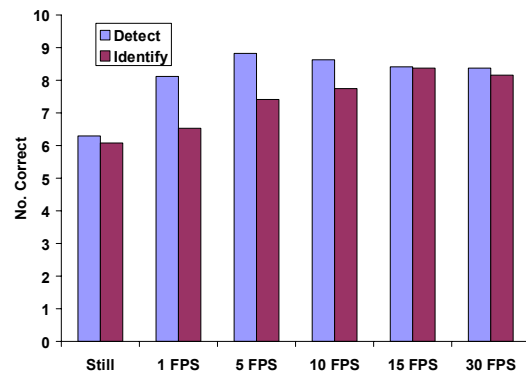
Figure 3. Correct Responses By Frame Rate and Task Type

3.3 *Target Motion, Camera Motion, and Scene Complexity*

Evaluations have indicated a small, but significant perceptual effect due to target motion. The clips with high target motion have been rated slightly higher in interpretability, compared to clips with little or no target motion. We investigated this issue with a new data set and found small, but significant effects. The levels of target motion, camera motion, and scene complexity were rated (high, medium, and low) for each clip. These ratings were provided by the research team, not imagery analysts. These ratings were incorporated into a stepwise regression analysis. The dependent variable was the mean interpretability rating for the clip and the candidate independent variables were $Log_{10}(GSD)$, the rating of target motion, the rating of camera motion, and the rating for scene complexity. While $Log_{10}(GSD)$ is the dominant explanatory variable, the other 3 factors were statistically significant (Table 3). This analysis suggests that target motion, camera motion, and scene complexity have small but perceptible effects on the interpretability of a motion imagery clip. Note that the difference in R^2 due to these factors is small – 0.84 vs. 0.80. The effect of ignoring these factors in a model such as an image quality equation would be to increase the error term. Clearly, further investigation is warranted.

4 Image Compression

The perception studies provide the foundation for our evaluation of image compression. We conducted a study of the effects of image compression on the interpretability of motion imagery. The dataset for the study consisted of the original (uncompressed) motion imagery clips and clips compressed by three compression methods at various compression rates[1]. The three compression methods were:

- Motion JPEG 2000 – intraframe
- MPEG-2 - intraframe: main profile at main level
- H.264/AVC – intraframe, main profile

All three were exercised in intraframe mode. Each of the parent clips was compressed to three megabits per second, representing a modest level of compression. In addition, each parent clip was severely compressed to examine the limits of the codecs. Actual bitrates for these severe cases depend on the individual clip and codec (Table 4).

Table 3. Regression Analysis of Factors Affecting Perceived Interpretability

	Coefficient	t-statistic	P- value
(Constant)	94.4	20.88	> 0.00001
$Log_{10}(GSD)$	-26.28	-23.96	> 0.00001
Target Motion	-6.27	-5.12	> 0.00001
Camera Motion	4.046	4.11	> 0.00001
Complexity	-3.40	3.13	0.002

$R^2 = 0.84$

The study used the Kakadu implementation of JPEG2000, the Vanguard Software Solutions, Inc. implementation of H.264/AVC, and the Adobe Premiere MPEG-2 codec.

The choice of compression methods and levels supports two goals: comparison across codecs and comparisons of the same compression method at varying bitrates. Table 4 shows the combinations represented in the study. Note that the lowest bit rate achievable not only varies across the codecs but across clips. We recorded the actual bit rate for each product and use this as a covariate in the analysis.

The study consists of two parts. For both parts, a set of compression products were generated using each of the codecs at the various bit rate. The first part of the study implemented a set of image metrics and examined their behavior with respect to bitrate and codec. The second part was an evaluation in which trained imagery analysts viewed the compressed products and the original parent clip to assess the effects of compression on interpretability.

4.1 Analysis of Image Metrics

The first metric for image quality is the **Structural SImilarity Metric (SSIM)**. SSIM quantifies differences between two images, I_1 and I_2, by taking three variables into consideration, luminance, contrast, and spatial similarity. For grey level images, those variables are measured in the images as the mean, standard deviation, and Pearson's correlation coefficient between the two images respectively. Let

$$\mu_k = mean(I_1)$$

$$\sigma_1 = standard\ deviation(I_1)$$

and

$$\sigma_{12} = \frac{1}{N-1}\sum_{i,j}(I_1(i,j)-\mu_1)(I_2(i,j)-\mu_2)$$

then

(1) $$SSIM(I1, I2) = \frac{2\mu_1\mu_2}{\mu_1^2+\mu_2^2} * \frac{2\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2} * \frac{\sigma_{12}}{\sigma_1^2\sigma_2^2}$$

The formulation in (1) was modified to avoid singularities, e.g., when both means are 0. SSIM is computed locally on each corresponding MxM sub-image of I_1 and I_2. In practice, the sub-image window size is 11x11. An entire image is created by implementing Equation 1 as a convolution window filter. The SSIM value is the average across the entire image.

Table 4. Codecs and Compression Rates

Bitrate	Uncompr.	H.264 (VSS) Intra-frame	JPEG 2000 (KDU) Intra-frame	MPEG 2 (Premiere) Intra-frame
Native	X			
3 MB /sec		X	X	X
Severe			X	

Where the severe bitrate represents the limit of the specific codec on a given clip.

Two edge metrics were examined. The first is denoted by CE for Common Edges and the second is denoted SE for strength of edges. Heuristically, CE measures the ratio of the number edges in a compressed image to the number of edges in the original; whereas SE measures a ratio of the strength of the edges in a compressed version to strength of the edges in the original.

In detail, given two images I_1 and I_2, $CE(I_1, I_2)$ and $SE(I_1, I_2)$ are computed as follows. Color images are HSI transformed to generate gray level intensity images. From the grey level images, edge images are constructed using the Canny edge operator. The edge images are designated as E_1 and E_2. Assume that the values in E_1 and E_2 are 1 for an edge pixel and 0 otherwise. Let "*" denote the pixel wise product. This is the intersection of the two edge images. Let G_1 and G_2 denote the gradient images of I_1 and I_2 respectively. G(m,n) was approximated as the maximum of absolute value of the set { I(m,n) - I(m+t1,n + t2) | -6 < t1 < 6 and -6< t2 < 6 }, i.e. the maximum difference between the center value and all values in a 5x5 neighborhood around it. With that notation,

(2) $$CE(I_1, I_2) = \frac{2*\sum(E_1*E_2)}{\sum E_1 + \sum E_2}$$ where the sum

is taken over all the pixels.

(3) $\quad SE(I_1, I_2) = \dfrac{\sum (E_2 * G_2)}{\sum E_1 * G_1}$ where the sum is

taken over all the pixels.

An additional set of edge operators were also applied. These operators are called **edge strength (ES)** metrics. Let I_1 be the luminance component of an original frame from an MI clip and let I_2 be the corresponding frame after compression processing, also in luminance. We apply a Sobel filter, S, to both I_1 and I_2, where for a grayscale frame F:

(4) $\quad S(F) = \sqrt{(H * F)^2 + (V * F)^2}$

The filters H and V used in the Sobel edge detector are

(4.1) $\quad H = \begin{array}{ccc} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{array}$

(4.2) $\quad V = H^T$

We define two metrics, one for **local loss of edge energy (EL)** (thus finding blurred edges from I_1 in I_2) and the other for the addition of edge energy (thus finding edges added to I_2 that are weaker in I_1). Each metric examines the strongest edges in one image (either I_1 or I_2) and compares them to the edges at the corresponding pixels in the other (I_2 or I_1).

For the grayscale image F, let I(F,f) be the set of image pixels, p, where F (p) is at least as large as f * max(F). That is:

(5) $\quad I(F,f) = \{P pixels : F(p) \geq f * \max(F)\}$

Using the definition of I(F,f), the two edge metrics are:

(6) $\quad BlurIndex = \dfrac{mean(S(I_2))}{mean(S(I_1))}$

where the means are taken over the set I(S(I_1), 0.99)

(7) $\quad AddedEdgeEnergy = \dfrac{mean(S(I_1))}{mean(S(I_2))}$

where the means are taken over the set I(S(I_2), 0.99).

Finally, we examined the **peak signal to noise ratio (PSNR)**. The PSNR is defined for a pair of m×n monochrome images, I_1 and I_2. Let MSE be defined by,

$$MSE = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} \| I_1(i,j) - I_2(i,j) \|^2$$

The PSNR is defined as:

$$PSNR = 10 \log_{10} \frac{MAX_I^2}{MSE} = 20 \log_{10} \frac{MAX_I}{\sqrt{MSE}}$$

where MAX_I is maximum pixel value of the image. In our case, MAX_I is taken to be 255. For color images with three RGB values per pixel, the definition of PSNR is the same except the MSE is the sum over all squared value differences divided by image size and by three spectral bands. However for the color videos we used luminance signal, Y, which we derived from RGB by the formula

$$Y = 0.299R + 0.587G + 0.114B .$$

The analysis shows that each of these metrics is monotonic in bitrate (Figure 4). In addition, both H.264 and JPEG 2000 outperform MPEG-2 for intraframe compression.

4.2 Evaluation of Image Compression

In addition to the analysis based on image metrics, we conducted an evaluation with imagery analysts. Using the uncompressed clips and a range of compression products, the analysts rated their confidence in meeting specific criteria, i.e., performing specific image exploitation tasks, on the various products. For each parent clip, three criteria (image exploitation tasks) were assigned. These were selected from the criteria set being used for scale development [9]. The considerations for selecting the criteria were:

- The criteria should "bound" the interpretability of the parent clip, i.e. at least one of the three should be difficult to do and one should be easy.
- The criteria (or at least some of the criteria) should reference objects and activity that are comparable to the content of the clip
- The criteria should have exhibited low rater variance in the previous evaluations

Image analysts rated their confidence in performing each criteria or image exploitation task with respect to each compression product, including the original (uncompressed) clip. We calculated an overall interpretability rating from each analyst for each clip. The method for calculating these ratings was as follows: Each of the three criteria used to rate each clip was calibrated (on a 0-100 scale) in terms of interpretability, where this calibration was derived from an earlier evaluation [9]. Multiplying the calibrated interpretability level by the IA's confidence rating produces a score for each criterion. The final interpretability score was the maximum of the three scores for a given clip. Let:

Interpretability Score(j, k) = max $\{C_{i,j,k} I_{i,k} : i=1,2,3\}$

where $C_{i,j,k}$ is the confidence rating by the jth IA on the kth clip for the ith criterion and $I_{i,k}$ is the calibrated interpretability level for that criterion. All subsequent analysis presented below is based on this final interpretability score.

All three codec yielded products for the evaluation, although MPEG-2 would not support extreme compression rates. Bitrate was the dominant factor, but pronounced differences among the codecs emerged too (Figures 5 and 6). At modest compression rates, MPEG-

2 exhibited a substantial loss in interpretability compared to either H.264 or JPEG-2000. Only JPEG-2000 supported more extreme intraframe compression and highly compressed renditions were produced from all of the parent clips. There were systematic differences across the clips, as expected, but the effects of the codecs and bitrates were consistent. The analysis of covariance confirms these statistical effects (Table 5). When modeled as a covariate, the effects of bitrate dominate. The effect due to codec is modest, but still significant. As expected, there is a significant main effect due to scene, but no scene-by-codec interaction.
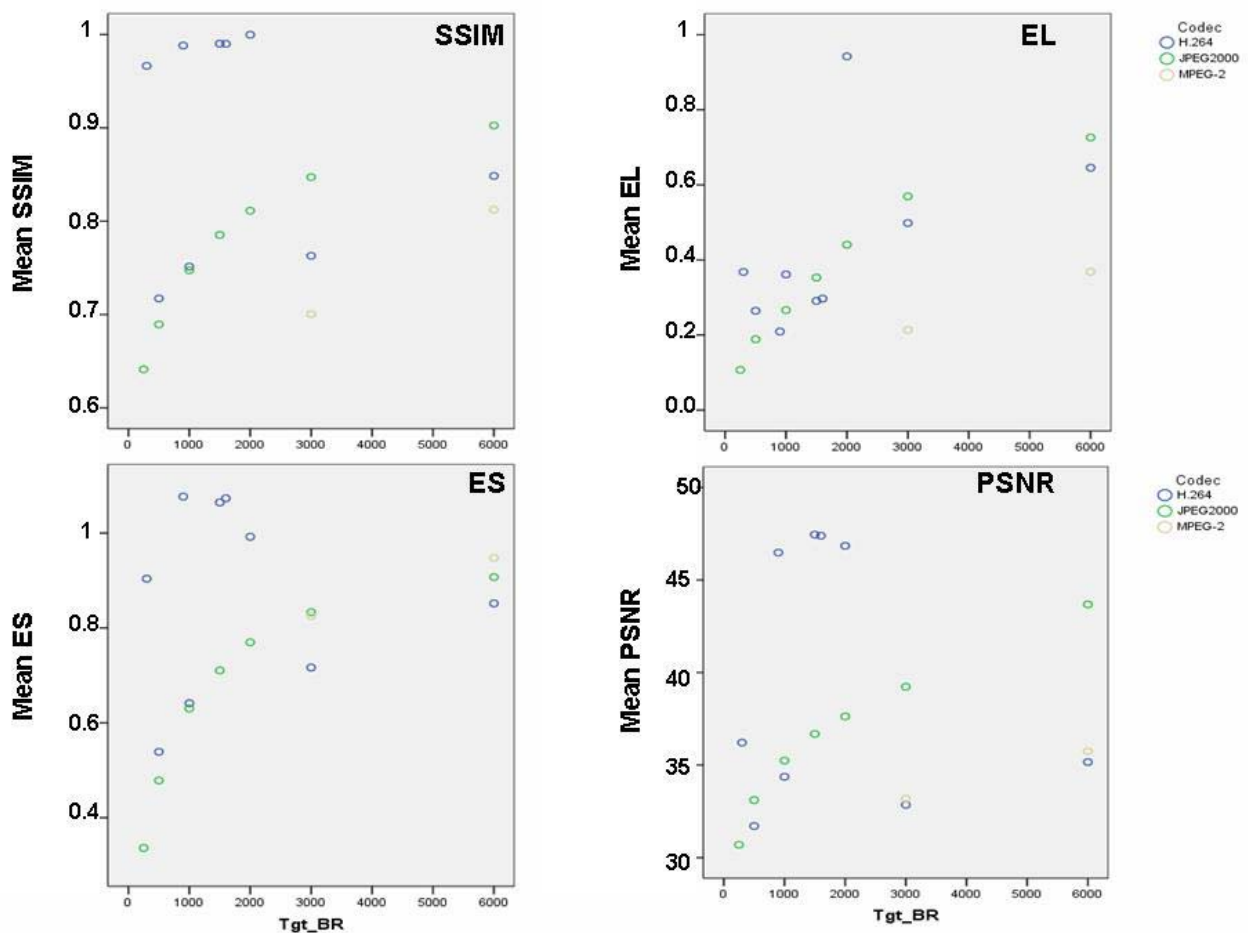

**Figure 4. Image Metrics Applied to the Compressed Products**

Table 5. Analysis of Variance for Interframe Comparisons

Source	Sum of Squares	Deg. of Freedom	Mean Square	F-statistic	Tail Probability
Intercept	1348896.07	1	1348896.07	0.850	0.38373
BitRate	21851680.16	1	21851680.16	24.428	0.00780
Codec	9436528.77	3	3145509.59	5.571	0.01008
Scene	48995713.72	4	12248928.43	26.225	> 0.00001
Codec * Scene	5350083.24	12	445840.27	0.498	0.84237

5   Conclusion

The evaluations and analyses presented in this paper characterize the primary factors affecting the perceived interpretability of motion imagery. The dominant factors are GSD (spatial resolution) and frame rate (temporal resolution). Secondary, but significant, effects have been observed for target motion, camera motion, and scene complexity. The team has made substantial progress in characterizing and quantifying these effects, laying the foundation for development of an interpretability scale for motion imagery.

Evaluation of image compression for motion imagery illustrates how interpretability-based methods can be applied to the analysis of the image chain. We present both objective image metrics and analysts' assessments of various compressed products. The results show good agreement between the two approaches.
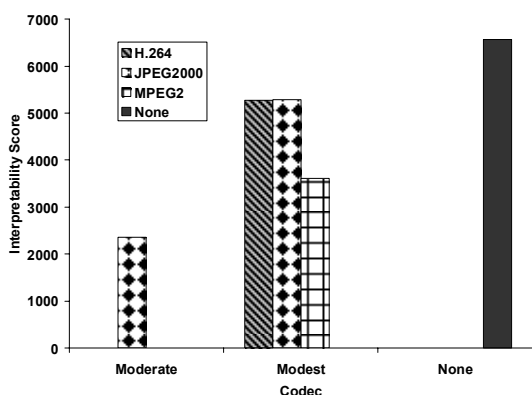
Figure 5. Summary Comparison Across Codec and Bitrate

References

[1]   J.M. Irvine,. "National Imagery Interpretability Rating Scales (NIIRS): Overview and Methodology" Proceedings of the International Society for Optical Engineering (SPIE), 29-30 July, 1997, Volume 3128, pp.93-103.

[2]   J.M. Irvine 2003. "National Imagery Intelligence Rating Scale (NIIRS)" in Driggers, R.G. (ed) The Encyclopedia of Optical Engineering, Marcel Dekker., October 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, April 1996.

[4]   J. C. Leachtenauer, W. Malila, J. M. Irvine, L. Colburn, and N. Salvaggio, "General Image-Quality Equation: GIQE," Applied Optics, vol. 36, pp. 8322-8328, 1997.

[5]   L.A. Maver, C.D. Erdman, K. Riehl, "Imagery Interpretability Rating Scales", Digest of Technical Papers, International Symposium Society for Information Display, Society for Information Display, Vol. XXVI, pp: 117-120, 23 May. 1995.

[6]   John M. Irvine, Charles Fenimore, David Cannon, John Roberts, Steven A. Israel, Larry Simon, Charles Watts, James D. Miller, Ana Ivelisse Avilés, Paul F. Tighe, Richard J. Behrens, Michelle Brennan, Donna S. Haverkamp, (2005) "Factors Affecting Development of a Motion Imagery Quality Metric", SPIE Defense and Security Symposium, Orlando, FL, 28 Mar-1 Apr 2005.

[7]   John M. Irvine, Charles Fenimore, David Cannon, John Roberts, Steven A. Israel, Larry Simon, Charles Watts, James D. Miller, Ana Ivelisse Avilés, Paul F. Tighe, Richard J. Behrens (2004) "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, 13-15 October 2004.

[8]   Charles Fenimore, John Irvine, David Cannon, John Roberts, Ivelisse Aviles, Steven Israel, Michelle, Brennan, Larry Simon, James Miller, Donna Haverkamp, Paul F. Tighe, Michael Gross (2006) "Perceptual Study of the Impact of Varying Frame Rate on Motion Imagery Interpretability", SPIE Conference on Human Vision and Electronic Imaging XI, SPIE 6057-17, January 2006.

[9]   John M Irvine, David Cannon, James Miller, Jeffrey Bartolucci, Laurie Gibson, Charles Fenimore, John Roberts, Ivelisse Aviles, Michelle Brennan, Aloise Bozell, Larry Simon, Steven A. Israel (2006)

"Methodology Study for Development of a Motion Imagery Quality Metric" SPIE Defense and Security Symposium, 17-21 April 2006, Orlando, Florida.

[10] John M. Irvine; Charles Fenimore; David Cannon; Donna Haverkamp; John Roberts; Steven A. Israel; Larry Simon; James Miller; Ana Ivelisse Avilés;

Michelle Brennan, "Development of a Motion Imagery Quality Metric," Proceedings of the American Society for Photogrammetry and Remote Sensing (ASPRS) Annual Meeting, 1-5 May 2006, Reno, Nevada, USA, 2006.
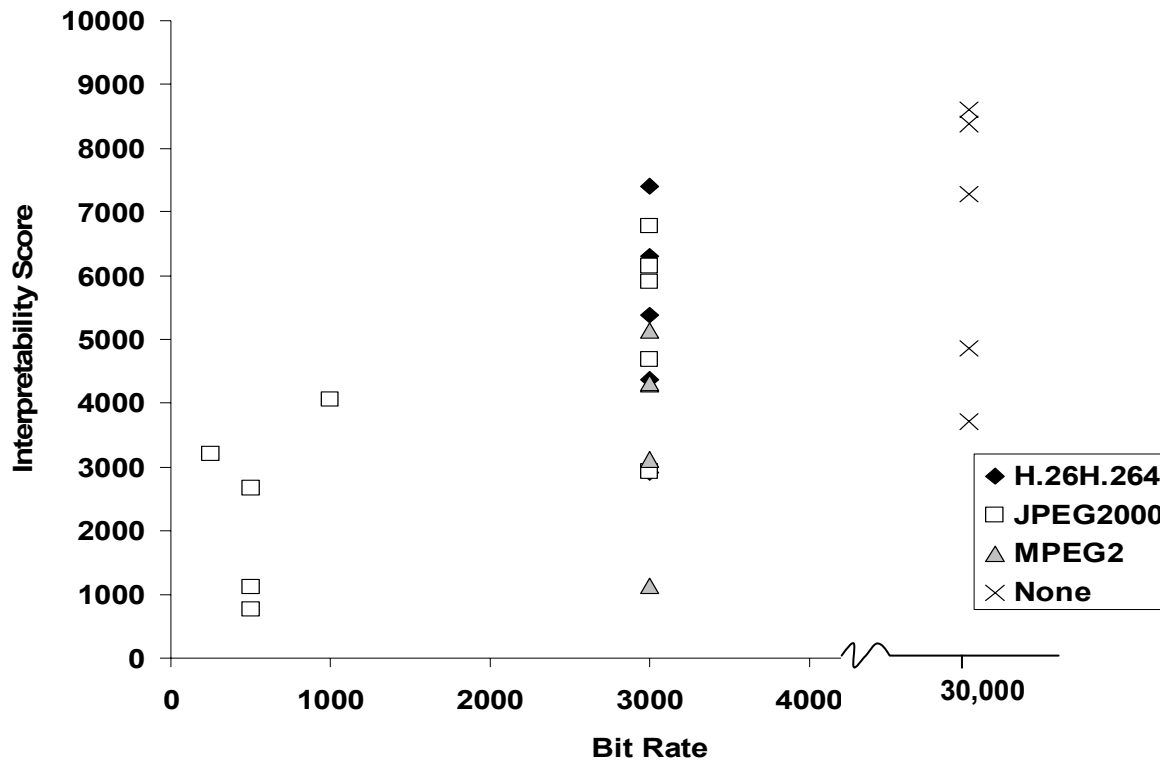
Figure 6. Interpretability Scores for Compression Products, by Codec and Bitrate

---

[1] The mention of commercial products in this paper is not intended as an endorsement by NIST, nor are such products necessarily the best available for the purposes of the described research.