Validating Rover Image Prioritizations

R. Castaño,[1] K. Wagstaff,[1] L. Song,[2] and R. C. Anderson[3]

Planetary exploration missions often have the capability of collecting more data volume than the mission is allocated in downlink bandwidth. Currently, instruments often collect and transmit data at a rate that equals their bit limit, remaining idle much of the remaining time. One scenario to take advantage of instrument capability is to permit continuous data collection and to autonomously select onboard the most scientifically valuable images for downlink. Given the same bandwidth constraint, this approach should increase the diversity and value of the transmitted data. A critical issue with this approach is how to prioritize the data. In this article, we present a quantitative technique for evaluating prioritizations. We apply the technique to assess the agreement of human experts on image rankings. These results can provide an upper limit to the kind of agreement we may expect between a set of rankings generated by humans and those produced by an automated algorithm. In addition, our analysis of the results leads to insights for maximizing the performance of automated algorithms that could be used onboard a spacecraft.

I. Introduction

As the number of planetary exploration missions increases, contention over the limited bandwidth available through the Deep Space Network (DSN)[4] will lead to critical decisions regarding what data to download. Currently, each mission and each instrument are allocated a specific amount of bandwidth. Instruments are generally tasked to collect exactly the quantity of data that will fill their bandwidth allocation. As a consequence, many instruments are idle much of the time. For example, the Thermal Emission Imaging System (THEMIS) instrument [1] on Mars Odyssey operates at a duty cycle of about 15 percent, although it is capable of much higher performance. Currently, the instrument collects only a small fraction of its possible observations and then transmits them all back to Earth. THEMIS could, however, collect at least three times this amount of data. An onboard mechanism would then be required to determine which data should be sent to Earth. Any onboard analysis that would select which data will be downloaded must be reliable and have a demonstrated ability to make good decisions, where the decision quality is evaluated by domain scientists. Evaluation of such methods for THEMIS, or any other instrument, to confirm that they do prioritize data as desired is critical.

---

[1] Modeling and Data Management Systems Section.

[2] Flight Software and Data Systems Section.

[3] Planetary and Life Detection Section.

[4] http://deepspace.jpl.nasa.gov/dsn/

The research described in this publication was carried out by the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

In this article, we present a quantitative assessment method for evaluating prioritization algorithms. The operational scenario we focus on is prioritizing images for rover traverse science. This assessment method assumes that there exists a "best" or "optimal" ranking of the data that represents how a human scientist would order the data. To test this hypothesis, we performed a study of the agreement between human experts. The level of agreement between experts defines the upper bound of possible performance for an automated prioritization algorithm. We find that the experts generally agree about which data is of the highest "science value." We also compared the agreement between non-experts on the same task. As we might expect, the non-experts showed a lower degree of agreement due to differences in how they interpreted "science value." These results indicate that automated methods, like humans, will require that the selection criteria—how the "best" ranking should be determined—be very precisely specified.

Section II of this article describes a method for quantitative comparison of two prioritizations of a set of data. In Section III, we present the results of our study of human experts. Section IV summarizes our findings and concludes with recommendations for improving the performance of automated methods.

II. Quantitative Assessment of Prioritization Methods

There is no generally accepted approach for evaluating prioritization methods. Therefore, we applied a statistical method for comparing two rankings of the same data set, where each ranking represents a prioritization.

Our method uses the Spearman rank-order coefficient [3] to calculate the agreement between two rankings, R and S, of the same data set. Let the set of n items (here, images), where $n > 1$, be given by $X = \{x_1 \cdots x_n\}$. Let R_i and S_i be the rank of item x_i according to R and S, respectively. We calculate the agreement between the two rankings using the Spearman rank-order correlation coefficient, r, which is expressed as

$$r = 1 - \frac{6D}{n^3 - n}$$

where D is the sum of the squared difference of the ranks:

$$D = \sum_{i=1}^{n} (R_i - S_i)^2$$

The Spearman coefficient r ranges in value from 1 to -1. When R and S agree perfectly, it is clear that $D = 0$ and, thus, $r = 1$. In contrast, when R and S completely disagree, we have

$$D = \sum_{i=0}^{n} (R_i - S_i)^2$$

$$= \sum_{i=0}^{n} \left(R_i^2 - 2R_iS_i + S_i^2\right)$$

$$= 2\sum_{i=0}^{n} i^2 - 2\sum_{i=0}^{n} i(n+1-i)$$

$$= \frac{n^3 - n}{3}$$

from which it follows that $r = -1$. Once r is known, the significance of r, which depends on the number of items being ranked, can also be calculated. To compare the rankings from multiple individuals, the mean correlation coefficient is computed. We use Student's t-distribution to calculate the significance of the mean agreement.

We can use this method to compare the output of an autonomous prioritization algorithm with the desired ranking of the same items. To do so, we require knowledge of the "correct" ordering of the items. In this article, we present the results of a study we conducted to determine the level of agreement between different humans. The results of this analysis, presented in the next section, provide insights on how to assess automated methods as well as recommendations for the design of those automated methods.

III. Experimental Results and Analysis

Automated prioritization of data for downlink is particularly useful for planetary rovers, which can typically take many more images than they can transmit back to Earth. For example, as a rover travels from one designated study site to another, it steadily records images for navigation (navcam) and hazard avoidance (hazcam) purposes. Some of these images may also be of scientific interest. Although the rover cannot send back every image at full resolution, it could use an automated prioritization method to rank the images for possible transmission. We refer to this scenario as rover traverse science.

For these experiments, we collected 25 images taken by the navigation cameras on the Field Integrated Design and Operations (FIDO) rover [6] during a field test near Flagstaff, Arizona, in August 2002. FIDO is an experimental test rover for the 2003 Mars Exploration Rovers. Each image taken by FIDO consists of 640×480 pixels; the average image size is 900 kilobytes. The images range from panoramic views of the landscape to local close-up views of rocks (see Fig. 1).

The goal of this study was to assess the level of agreement between humans when asked to prioritize a set of images. We collected ranking information from each person independently. First, we presented all 25 images and asked which image was the most important from a scientific standpoint—that is, if the rover had taken the 25 images and it could only send back one, which would be the most important to transmit to the science team? After removing that image, we asked which of the remaining images was now the most important. We proceeded in this way until we obtained a full ranking of all 25 images from each person. A total of sixteen people, of diverse backgrounds, participated in the study. Seven of these individuals were geologists with rover experience (this includes one geology student) and were considered experts. The remaining nine individuals had technical backgrounds but no particular training in geology or rover traverse science and were considered non-experts. A vulcanologist was included in the non-experts for this study, because while an experienced field geologist, he did not have experience with rover science.

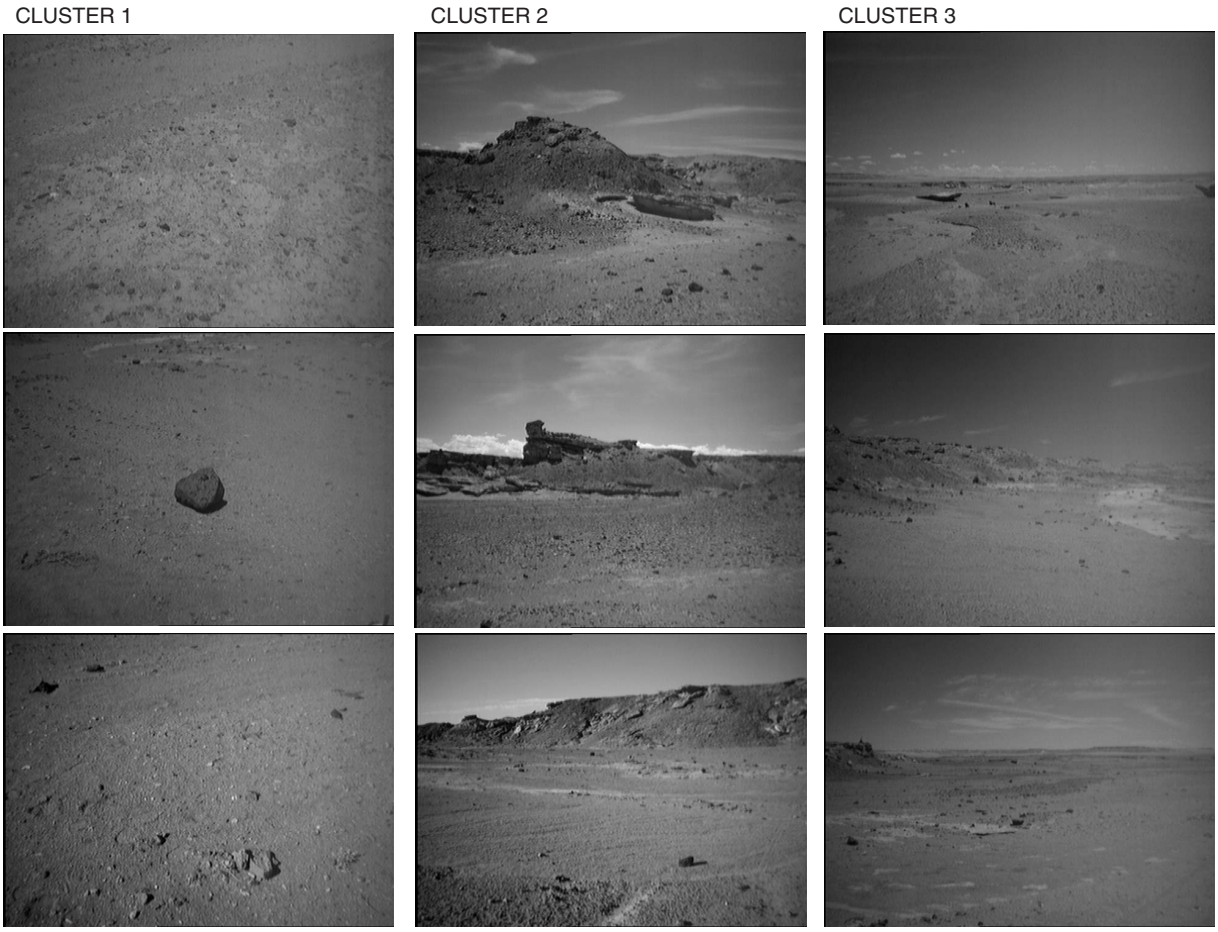

Fig. 1. Sample FIDO images used in the experiment. Referring to the portion of the experiment in which the images were divided into three clusters, the first column contains examples that were placed in cluster 1, the second column shows examples from cluster 2, and the third column shows examples from cluster 3.

A. Complete Ranking

The first test that we performed was to directly compare the complete image rankings, calculating the Spearman rank-order coefficient between each pair of individuals. The Spearman correlation between members of the study varied from a low of -0.23 to a high of 0.93. For the overall correlation, the mean of the Spearman correlations between each pair of individuals (16 people result in 120 pairs) was used. The mean correlation was 0.34. More importantly, the mean correlation among experts with previous rover experience was 0.41, which is significant at the 95 percent level. These seven individuals had pairwise correlations that ranged from 0.43 to 0.93. All but two of these pairwise correlations are significant at the 99 percent level. This indicates that, in general, the experts were positively correlated on the ranking of the data. The positive correlations, while not always strong, indicate that there is agreement that some orderings of the data are preferable to others. Thus, if not all data could be transmitted, there would be a distinct benefit to prioritizing the data for downlink even if the exact best ordering is not agreed upon.

The relationships of correlations between individuals are shown graphically in Fig. 2, which is a multidimensional scaling of the pairwise correlation values, projected into two dimensions [2]. Here, distance corresponds to (lack of) correlation. Six of the individuals with previous rover image experience are indicated with the letters A through F, where E is a geology student. The seventh expert is I. H is a vulcanologist and had a very different set of priorities. The rest of the participants, non-experts, are more widely scattered.

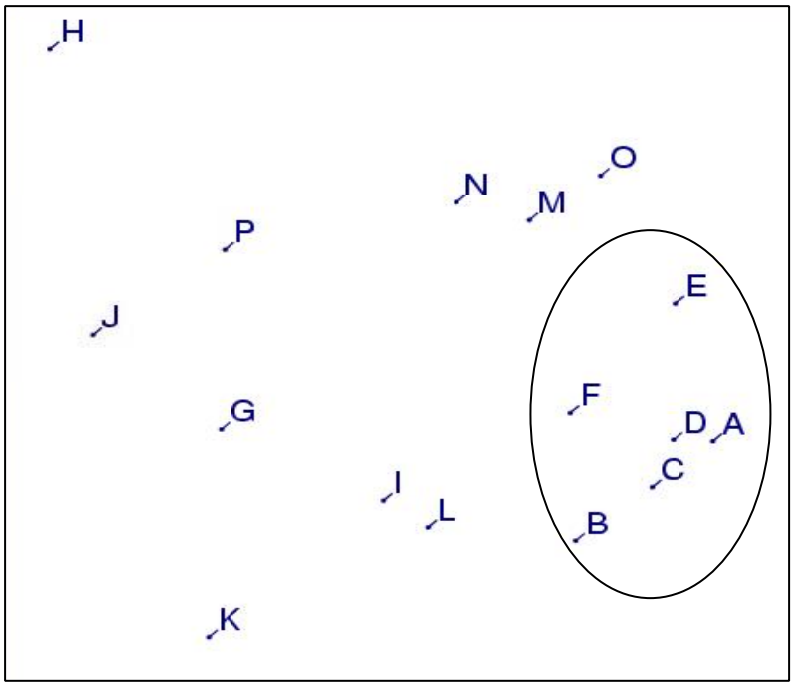

Fig. 2. Visualization of similarities between human rankings. The ellipse is drawn around the points representing six of the seven experts, emphasizing that the experts are relatively tightly grouped. The last expert is I, slightly beyond the group.

B. Correlations for Image Clusters

After analyzing agreement on the complete ordering, we evaluated the data and prioritizations to determine if there was a distinction between general classes of images present in the data set. The data were sorted into three clusters with the k-means clustering algorithm [4] using features automatically extracted for each image, as described in [5]. The properties of the images in each resulting cluster can be described as follows (see also Table 1). Cluster 1 consists of images of the near field. They are filled with a view of the ground and do not have a horizon in the image. Clusters 2 and 3 contain images with a horizon, where the images in cluster 2 are distinguished as having features on the horizon (hill, cliff, etc.) and images in cluster 3 generally have a flat horizon. Examples for each of the clusters are shown in Fig. 1. After clustering and assigning the cluster identification (id) to each image, we then evaluated each individual's ranking of the images to determine if they had a preference for a certain cluster. We found that, in general, individuals preferred horizon images to foreground-only images.

For this statistical test, each image was labeled as either belonging to cluster 1, 2, or 3. Our goal was to determine if there was any difference between the rankings of clusters. The null hypothesis was that "there is no significant difference between the clusters." We then sought to disprove this hypothesis, i.e., to show that the clusters are meaningful with respect to rankings. We found that 12 of the 16 people (75 percent) show significant differences in their clustered ranks at the 90 percent significance level; 9 of 16 (56 percent) show differences at 95 percent, and 7 of 16 (44 percent) at the 99 percent level. The six geologists with rover experience were at the 99 percent level. One expert was at the 95 percent level. None of the individuals at the 90 percent or lower level of significance were experts. All of the experts preferred cluster 2 over cluster 3 and cluster 3 over cluster 1. All but two individuals in the experiment felt that cluster 2 was the most important. The implications of this result are that if onboard analysis can determine a semantic label (cluster) for an image, a priority can be assigned based on the label.

Table 1. Description of the subsets into which the collection of images was divided.

Cluster	Number of members	Image properties
1	8	Near field, no horizon
2	10	Horizon with feature
3	7	Flat horizon

Interestingly, some individuals actually physically divided the images into three classes while conducting the experiment and explicitly took an image from one class and then another, rotating through them. Such a method emphasizes diversity, while assigning all the images from one class a high priority focuses on depth of information on that class. In addition, several experts stated that an image with only foreground was of little use without an image to provide context, in this case an image with a horizon. Thus, the scientific value of the images is not independent of each other.

C. Correlations Restricted to Single Clusters

Having determined that the experts and many of the non-experts did consider some sets of images to be of higher priority on the average than others, we then wished to determine if there was agreement on the rankings of images within each cluster. The results are shown in Fig. 3. As the clusters have fewer samples (image members), the significance for the same level of agreement decreases. Even taking this into consideration, cluster 2 has a very low agreement. It is interesting that, while there was almost complete agreement that cluster 2 was the most important, there is almost no agreement on which images within cluster 2 are the most important. The effect is the most pronounced with the experts, for whom there was over 95 percent confidence in the ranking order of cluster 2 for every expert and close to zero correlation on the ordering of images within cluster 2. This indicates that the most important information in the prioritization was captured by the cluster membership. The non-experts agreed very well on the images in cluster 1, while the experts did not. The result indicates that there were distinguishing features in the images of this class that the non-experts all thought were important but that the experts did not feel were important.

IV. Conclusions and Recommendations

One method for maximizing science return when downlink bandwidth is limited is to prioritize the data onboard before transmission. Automated prioritization methods must be evaluated for their ability to construct data rankings that reflect the priorities human scientists would assign to them. In this article, we presented a qualitative method for comparing two rankings of the same data set. We used that method to assess the level of agreement between humans and found that the broad goal of "ranking by science value" led to agreement that was statistically significant, but did not lead to a "best" ordering of the data among experts. Among non-experts, however, there was less agreement. When the images were divided into three distinct groups, it was determined that most of the expert agreement on the overall ranking is explained by membership in a class and very little on the order within the class. This is a very promising point for onboard prioritization since the images were divided into three clusters using an automated method.

We found that one of the biggest obstacles to higher agreement was the inherent vagueness in the term "scientific value." Each person interpreted this criterion differently. Those with rover image experience interpreted it in that context and, since most of them had worked with images for navigation purposes, they tended to prioritize images that would be most useful for planning a rover's future traverse. In contrast, the vulcanologist (H) ranked the images in a very different way (H is actually negatively correlated

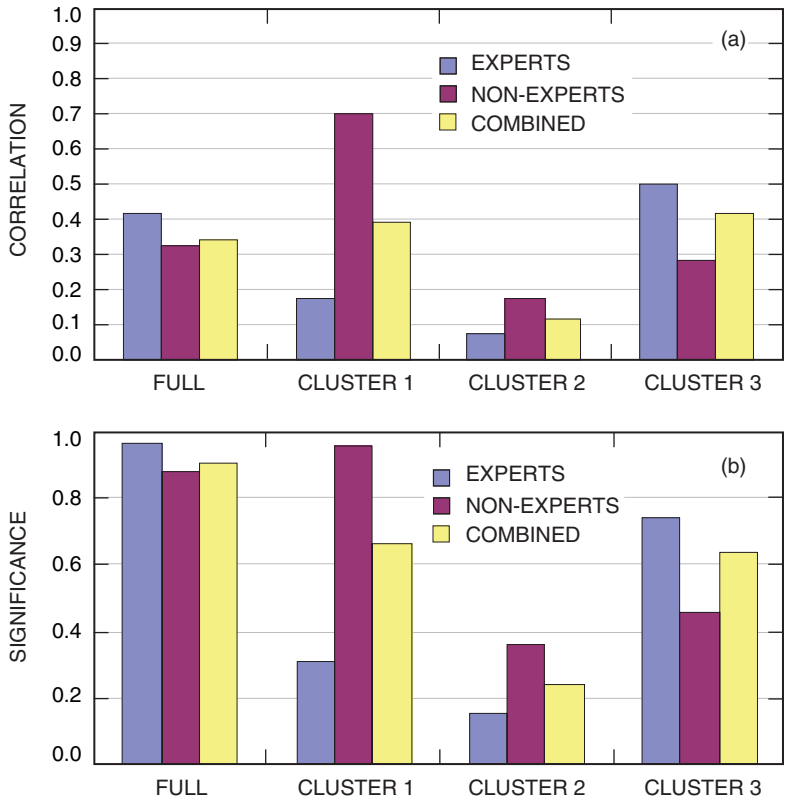

Fig. 3. Agreement on the rankings of images within each cluster: (a) mean correlation between rankings from pairs of individuals and (b) significance of the ranking correlation.

with two-thirds of the other participants). The remaining participants had other interpretations, such as prioritizing close-up images of rocks or looking for images with interesting structure on the horizon.

Several of the experts commented that scientists rarely consider the broad question of which data are scientifically most interesting. Usually, they operate in a scenario where they are studying a certain phenomenon or evaluating a specific hypothesis. In addition, each scientist had a bias towards what was interesting based on his or her area of expertise. For example, a sedimentologist and petrologist might disagree on the relative value of a soil image versus a rock image. The diverse interests of a science team are moderated by the mission scientist or instrument principal investigator. Decisions are made based on anticipated information gain. These observations led us to conclude that, to achieve the highest level of agreement and reliability, the ranking goal must be very precisely stated by an expert.

We recommend for the future development of automated prioritization methods that, rather than ranking items by their scientific importance, we are likely to see more success when ranking items based on their relevance to a specific scientific issue, such as evidence for past water. This approach would also lead to higher agreement with human rankings produced with the same goal in mind, and it is likely that any disagreements could be more readily interpreted and resolved.

Acknowledgments

We would like to thank Ray Arvidson, Ben Bornstein, Natalie Cabrol, Ashley Davies, Tara Estlin, Forrest Fisher, Edmund Grin, Ed Guinness, Michele Judd, Sherri Klug, Scott McLennan, Dominic Mazzoni, and Erik Pounders for their participation in the image prioritization experiment. We also thank Michele Judd for her helpful comments. This work was funded by the Interplanetary Network Directorate.

References

[1] P. R. Christensen, B. M. Jakosky, H. H. Kieffer, M. C. Malin, H. Y. McSween, Jr., K. Nealson, G. L. Mehall, S. H. Silverman, S. Ferry, M. Caplinger, and M. Ravine, "The Thermal Emission Imaging System (THEMIS) for the Mars 2001 Odyssey Mission," *Space Science Reviews*, vol. 110, pp. 85–130, 2004.

[2] J. B. Kruskal and M. Wish, *Multidimensional Scaling*, Beverly Hills, California: Sage Books, 1978.

[3] E. L. Lehmann and H. J. M. D'Abrera *Nonparametrics: Statistical Methods Based on Ranks*, Englewood Cliffs, New Jersey: Prentice Hall, 1998.

[4] J. B. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," *Proceedings of the Fifth Symposium on Math, Statistics, and Probability*, vol. 1, pp. 281–297, 1967.

[5] D. Mazzoni, K. Wagstaff, and R. Castaño, "Using Trained Pixel Classifiers to Select Images of Interest," *The Interplanetary Network Progress Report*, vol. 42-158, Jet Propulsion Laboratory, Pasadena, California, pp. 1–8, August 15, 2004. http://ipnpr/progress_report/42-158/158G.pdf

[6] P. S. Schenker, E. T. Baumgartner, P. G. Backes, H. Aghazarian, L. I. Dorsky, J. S. Norris, T. L. Huntsberger, Y. Cheng, A. Trebi-Ollennu, M. S. Garrett, B. A. Kennedy, A. J. Ganino, R. E. Arvidson, and S. W. Squyres, "FIDO: A Field Integrated Design & Operations Rover for Mars Surface Exploration," i-SAIRAS, Montreal, Canada, June 2001.