Assessing Geologic Image Interpretations Errors Occurring in Extraterrestrial Robotic Exploration*

Jacob L. Wagner
Mechanical and Industrial Engineering
University of Iowa
Iowa City, IA, U.S.A.
jacob-wagner@uiowa.edu

Geb W. Thomas
Mechanical and Industrial Engineering
University of Iowa
Iowa City, IA, U.S.A
Geb-thomas@uiowa.edu

Justin Glasgow
Mechanical and Industrial Engineering
University of Iowa
Iowa City, IA, U.S.A
Justin-glasgow@uiowa.edu

Abstract-*Measuring the error introduced by the machine and human in a man-machine system is an important step in understanding and improving the system. Two experiments assessed the error from the total system in length measurements and shape classification. A separate analysis calculated the error introduce by the camera. The average error for length measurements was 2.33 pixels with a standard deviation of 2.399. The error in edge location introduce by the camera was 5-10 pixels. This indicates that humans generally are good at measuring lengths, but have difficulty near the threshold of resolution. Results from the classification task indicate that humans have lower then expected intra and inter geologist consistency. The study indicates that increasing the resolution of the camera system increases length measurement accuracy and that training may be the best way to improve classification consistency.*

Keywords: man-machine system, error, extraterrestrial geology

1 Introduction

Robotic exploration of the Martian surface requires interpretations of imaged data that is analogous to observations made in field geology, which is a discipline traditionally based on direct observation. This work explores the hypothesis that the differences in robotic and direct geology may lead to potentially important differences in the accuracy of the interpretation in scientific data returned from Mars. In man-machine systems, errors can be subdivided into those caused by the machine processes and those caused by human interpretation. In this work, errors are defined as loss of interpretation accuracy or misinterpretation caused by the differences in direct and robotic geology caused by the imaging system (lack of resolution, blurring, noise, digitization) or human interpretation (skill based, rule based, or knowledge based errors)[2]. In past field tests, both types of errors presented themselves. Thomas [4] found that geologists misclassified rocks because the cameras could not resolve the speckling on them. Anderson noted that geologists overlooked important science targets that were clearly visible in spectral images [1]. Tracing the information flow through the system and identifying locations where errors occur provides useful information to mission geologists and gives system designers powerful information for improving future systems, either by refining the design of the robot or improving the training techniques for the geologists. This paper assesses how information about specific sediment characteristics is transformed by a camera system and how this transformation affects interpretation of the images.

During data acquisition, the robot gathers information from the environment, processes the information, and sends the images to Earth for interpretation. Using the rover to collect the information adds an extra layer of information filtering for a field geologist. The difference between observing a physical specimen and observing an image of the specimen has never been studied; however, there is a loss of sensory cues that geologists may use in their interpretation. To analyze the loss of information that occurs during the image processing, a series of experiments were conducted to measure the errors and noises that are introduced into the system by its various components.

Two geologists from the 2004 MER science team were interviewed to determine specific, mission relevant interpretation features for the study. The focus of the 2004 mission is determining whether water (liquid or solid) flowed on the Martian surface. Sediment analysis is an important technique that will be used by the geologists to distinguish between wind and water processes; determine the extent, volume, magnitude, and nature of various flows; and other aspects of the geological history of the studied areas. Sediment information was used for this study because of its importance in the final conclusions reached by the mission team and the fact that it relies heavily on interpretation of imaged data, which would normally be done using physical specimens.

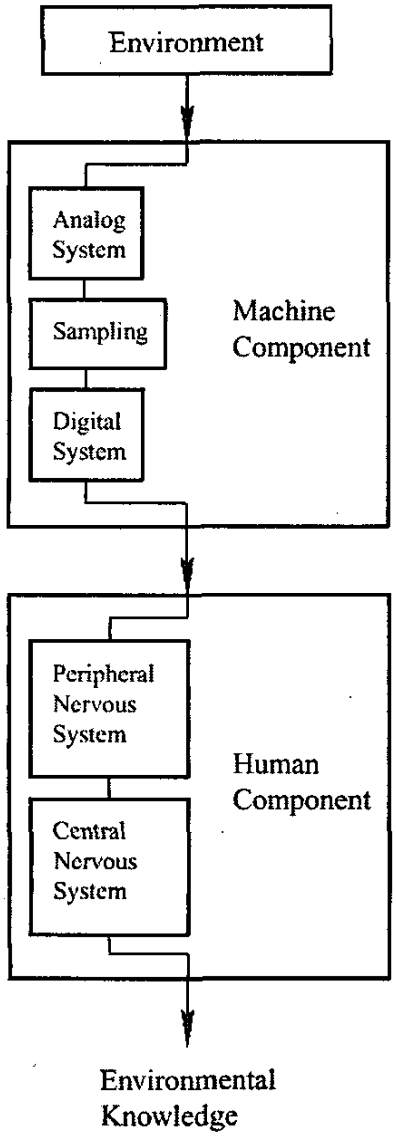

Figure 1: Model of man-machine system showing flow of information through both the machine and human components of the system.

2 System Model

The entire system model consisted of three components (Figure 1): the environment, representing the true state of the world; the machine, consisting of the rover; and the human. Information originating in the environment flows through the system and is eventually transformed into knowledge about the environment. The machine can be broken down into three subsystems, an analog subsystem (the camera lens), a sampling subsystem (the arrangement of imaging elements on the CCD chip), and a digital subsystem (the output of the CCD chip). All three of these subsystems can introduce error into the information as it passes through. The human was modeled using two subsystems, which are roughly split along the lines of the central and peripheral nervous systems. The Peripheral Nervous System (PNS) is the sampling and filtering system for the human and the Central Nervous System (CNS) is the cognitive or decision system. This division is somewhat arbitrary because some decision-making does occur in the PNS and some areas of the CNS perform filtering and processing tasks. The error caused by the cameras and the errors in the total system were measured separately. Using these two error measurements, the error caused by the two systems can be determined.

The model highlights all of the different subsystems that can cause problems and it provides information on where and why errors occur. For this experiment the machine component of the system was a Canon EOS D30 digital camera. The EOS D30 possesses a 3.11 Megapixel (2160 X 1440 pixel) CMOS imaging chip fitted with either a Sigma 300mm telemacro lens or a Canon 52mm zoom lens. The subjects were either professional geologists (used for the shape experiments and some of the size experiments) or university students.

3 Camera Error

Images of a calibration target were collected and analyzed to determine the error introduced by the machine. The test picture consisted of 2.54 mm wide black strips spaced 2.54 mm apart on a white background. The error contributed to the total system by the machine results mainly in the blurring of edges. Two factors contribute to the blurring of the edges: charge overflow and lens aberrations. These two factors combine to create a per edge pixel error of 5 to 10 pixels.

In the three images used to determine the error, 31 alternating black and white strips covered an average of 2123.67 pixels. With 27 pixels per mm, each bar should measure 68.51 pixels. Measurements from a calibration picture concurred with the calculation of 27 pixels per mm. All calculations were made using a value of 68 pixels, rounding downward because it leads to a conservative estimate of camera error.

Analysis began with the center of the black strip located in the center of the image. The amount of aberration varies with the deviation from the optical axis, so in the center there should be minimum blur. Measurement of the number of black pixels between two edges produced a bar size of 62 pixels. Measuring from the edge where the white pixels begin across the bar to the next white edge gives a bar size of 72 pixels. These numbers suggest that, due to charge overflow and aberration, the white strip blurred into the black 3 pixels on each edge. The last two pixels of the white strip have a slight coloring to them, which may be caused by chromatic lens aberration. Measurements at 8 different points, the top and bottom of the center black strip plus the top, center and bottom of the far right and far left black strip agreed with the value of a 3 pixel per edge blurring due to monochromatic lens aberration and charge overflow. For this set of images the maximum amount of chromatic lens aberration was 4 pixels, which occurs at the bottom of the far right bar.

The chromatic aberration causes coloring along each edge. From left to right of the image the colors vary from red to blue, with the right edge of an individual bar a different color from the left. The right bar edges are red on the left side of the image and are green on the right side of the image. The left bar edges vary from blue-green to blue.

4 Total System Error

Error introduced by the human can be more difficult to quantify than the physical imaging parameters. In robotic geology, the geologist uses images and sensor readings to understand the environment. In field geology, the geologists go into the environment and use all of their senses to collect information. There are many differences between these experiences, including the availability of multiple viewpoints, richer context information, and the ability to interact with and manipulate the environment. To measure the magnitude of the changes in interpretation caused by some of these differences, two experiments were conducted comparing the geologists' interpretation of imaged data with their interpretation of physical rock specimens. We define this difference to be the total error in the system (camera error plus human error). Interpretation differences that cannot be accounted for by camera error are defined to be human error. A measurement of the information conveyed by an imaged specimen versus a physical specimen is another way to quantify the total system information loss. Further, comparing interpretations across different geologists provides a measurement of observer variability.

The two experiments focused on sediment grains. Two characteristics were measured: grain size and grain shape. Grain size was just the measurement of a grain along a predetermined axis. Grain shape is a classification of the smoothness and flatness of a grain. These two attributes measure different things, one is a length measurement, and the other is a classification. The principle hypothesis in both experiments is that there is human error caused by the difference in direct and indirect observation; any information loss between the interpretation of the physical specimens and the image specimens that is greater then the loss that could be explained by the camera alone is attributed to the human component.

Three different camera settings were used in each experiment. The settings varied in their resolution and the type of lens used. Two settings were designed to mimic the panoramic camera on the MER robot with working distances of 1 and 2 meters. This results in resolutions of approximately 100 and 50 pixels per inch. The third setting was designed to mimic the microscopic imager used to gather magnified images. This camera setting had a resolution of approximately 590 pixels per inch. These were the target resolutions; each image was calibrated using a calibration target consisting of geometric shapes with known edge lengths to determine the actual resolution.

4.1 Grain Length Experiment

4.1.1 Methods

The length of a predetermined axis on a grain is the simplest measurement the geologists can make. The lengths of 280 rocks were measured along an arbitrary but specific axis using a micrometer with an accuracy of 0.001 inches. These 280 rocks were then photographed using the three camera settings. The rocks were arranged so that the measured axis was parallel to the cameras image plane and the axis was identified in the image using paper arrows. After all the specimens were photographed, the lengths were converted form units of inches to pixels using coefficients derived from the calibration targets. These distances were considered the true value to compute the accuracy of the subjects.

Seven subjects participated in the length experiment (5 geologists and 2 students). In each trial, the subject was required to measure the length of the grain, in pixels, along the specified axis. To make this measurement, the subjects had to determine both edges of the grain. Subjects used either the Photoshop or NSF Image software to view the images and make the measurements. Both pieces of software have a measuring tool that outputs the Euclidian distance in pixels for any line segment identified in the image. This reduced the error that could occur if the subjects were required to count the number of pixels by hand. Subjects were also allowed to zoom into any region of the image, if they desired.

4.1.2 Results and Discussion

The measurements gathered from the images were compared with the true measurements to determine the total system error. Two different error measurements were used: the actual difference between the physical distance and the image measurement, and the absolute value of this difference. The average error for the actual difference was -1.255 ± 0.90 (SEM) pixels ($p < 0.0001$) with a standard deviation of 3.077 pixels. This indicates that the subjects tended to measure the images slightly longer then the actual measurements. A normal probability plot of the error indicates that the error is normally distributed and the mean is shifted slightly. The measurement of the absolute error was done to determine the measurement error. On average the measurements made by the subjects was off by $2.333 \pm .070$ (SEM) pixels ($p < 0.0001$) with a standard deviation of 2.399. This measurement is similar to the error introduced by the camera.

4.2 Grain Shape Experiment

4.2.1 Methods

The loss of information between imaged specimens and physical specimens for grain shape interpretations required each subject to interpret an image of a specimen and the physical specimen. 195 different grains were selected for this experiment. The grains were selected to include a wide range of sizes and shapes. The grains were imaged, numbered and store in individual containers.

Six subjects (all geologists) interpreted both the grain roundness and the grain sphericity using an image of each individual grain. The geologists used a scale containing 6 degrees of roundness varying from well round to very angular and 6 degrees of sphericity varying from spherical to flat. Each subject was familiar with the classification system before the experiment started. Each category was assigned a number and the subjects recorded the number for each grain on the provided data sheet. Once the subject had interpreted all of the imaged grains they were given the physical grains to interpret. The order of the grains was randomized between the image analysis and the physical analysis to reduce the likelihood that the geologists would remember their previous classifications. The two interpretations of each grain (imaged and physical) made by each subject were then compared.

4.2.2 Results

The results from the grain shape experiment indicate that there is a difference in the classifications of the physical and imaged specimens. The average difference in sphericity was -0.151 ± 0.046 (SEM) categories with a standard deviation of 1.407 categories ($p = 0.0010$), which indicates that the imaged specimens were classified to be slightly flatter then the physical specimens. The average difference in roundness was –0.114 ± 0.043 (SEM) categories with a standard deviation of 1.325 ($p = 0.0082$), which indicates that the geologists found the imaged specimens rounder then the physical specimens. For both classifications, the errors were normally distributed.

By comparing the classifications made of the imaged and physical samples, we can estimate the total system error introduced by robotic geology, using information theory [3]. Information theory measures the transfer of information through a communication system in the presence of system (in this case both human and machine) noise. The theory measures both input and output signals in terms of the number of bits it would take to minimally encode the relevant information. The number of input bits minus the number of output bits determines the information lost. The calculation is based on a 6X6 confusion matrix. To generate a confusion matrix, we selected a pair of subjects and placed the categorizations made by the first subject along each row and the categorizations made by the second subject along each column. If the first subject coded 5 rocks with the value 1, and the second subject also categorized 4 of the same rocks with the value one, but one of rocks with the value 2, then the first row of the confusion matrix would be [4 1 0 0 0 0]. The confusion matrix allows the calculation of both the input and output signals, according to the following formulae:

$$H(S) = -\sum_{s=1}^{n} p(S)\log_2(p(S)) \qquad (1)$$

$$H(R) = -\sum_{s=1}^{n} p(R \mid S)\log(p(R \mid S) \qquad (2)$$

where H(S) is the information in the signal (in bits), H(R) is the information in the response (in bits), p(S) is the probability of a particular signal, and p(R|S) is the probability of a particular response given a particular signal.

Table 1: Average Information Communicated Among Subject Ratings of Roundness and Sphericity of Imaged and Physical Rock Specimens.

Comparison	Average Info in Stimulus	Average Info Communicated	N	Standard Deviation
Physical Roundness	2.3	1.1	30	0.4
Physical Sphericity	1.9	0.8	30	0.5
Imaged Roundness	2.1	0.7	30	0.4
Imaged Sphericity	2.0	0.6	30	0.5
Physical v Imaged Roundness	2.3	0.8	36	0.5
Physical v Images Sphericity	1.9	0.5	36	0.5

We constructed a total of 218 confusion matrices for the categorizations of all combinations of each pair of participants and the following categorizations: physical roundness versus physical roundness, physical sphericity versus physical sphericity, imaged roundness versus imaged roundness, imaged sphericity versus imaged sphericity, physical roundness versus imaged roundness and physical sphericity versus imaged sphericity. We then averaged the amount of information in the original classifications for each comparison and the average amount of information in the response. For the response classification averages involving the same classification material, we did not include contributions of participants with themselves, because these contributions would represent the perfect transmission of information. Table 1 summarizes the results.

4.2.3 Discussion

The value of 2.3 bits of information in the physical roundness of rocks and 1.9 bits of information in the physical sphericity of rocks indicates that geologists generally believed that there was a wide distribution in the rocks provided for study, since if the rocks were evenly distributed across the 6 classification levels, the information in the signal would be at a maximum of $\log_2(6) = 2.58$ bits. The third column, however, suggests that the geologists were not very consistent in their rock categorization. The value of 1.1 and .8 bits of communicated information suggests that if one geologist classifies a physical specimen of a rock, only approximately 1 bit of information would be transmitted to another geologists. Since one bit of information was used, the information could be most efficiently coded into just two categories. This result suggests that instead of six levels of categorization for roundness and sphericity, only two may be required to accurately convey the geologists' interpretations.

The fourth and fifth rows indicate that although the information containing the image inputs signals is similar, much less information is conveyed among the geologists by their categorizations, particularly for estimates of roundness, which drops from 1.1 bits of information communicated for physical specimens to .7 bits of information for imaged specimens. The table indicates a smaller information drop for judgments of sphericity: from .8 bits (physical) to .6 bits (imaged). This indicates that the geologists disagreed more about roundness and sphericity categorizations made from physical specimens than they did for imaged specimens.

The best measure of the difference between physical and imaged judgments may be made from the last two rows of the table, which compare average information transmitted between geologists when comparing physical and imaged specimens. This measure includes measurements of each subject when viewing the same physical and imaged specimens, which would presumably allow more information to be transferred, compared to the other measures. However, the information transferred is just .8 or .5 bits for the roundness and sphericity judgments, respectively. This indicates that there is very little agreement in the categorizations of the rocks made by the geologists when comparing categorizations of the physical versus imaged specimens. Comparing the between subject information loss with the between presentation style loss of information, suggests that most of the information lost is caused by categorization disagreements among the geologists, but that a substantial information loss is caused by the imaging process.

5 Conclusion

The two experiments produced different results for people's abilities to measure sediment grain characteristics. For the length measurement experiments, the error measured for the total system was similar to the error caused by the camera system. This suggests that people are good at finding edges and determining lengths for well lit, near field images. The best way to increase the accuracy of the measurements is to increase the resolution of the camera since the error in pixels did not change for different camera resolutions. The experiment also suggests that people have difficulty measuring objects that are only a few pixels long. The blurring and diffraction added by the camera along with the error from the humans would indicate that objects that are less then 5-6 pixels long would be difficult to accurately measure. This means that if the geologists expect to see something that is near this lower threshold of observability they may need to switch to a higher resolution camera.

The shape experiment showed that there was agreement between shape classification using the images and classification using physical specimens, but the information that a classification conveyed between geologists was rather low. Although it is difficult to precisely determine how the properties of the imaging system affect these judgments, the results suggest that the imaging system decreases the agreement of rock classifications. Considering the broader problem of rock classification, however, the results suggest that the primary issue is learning how to improve the classification agreement between geologists for the same physical specimen, and then addressing how the imaging system affects these agreements. A critical component of such a study would include the identification of within-subject classification for the same specimen. If an individual geologist is not consistent in classification then there is little hope that inter-geologist classification would be consistent. We are currently conducting a follow-up experiment to determine the size of this effect.

This work attempts to bridge the gap between the scientific interpretations made by geologists and the designers of the information collection systems. By experimentally assessing the impact of system parameters on specific scientific judgments, this work helps designers understand what system parameters affect scientific results, and shows scientists how their judgments may be affected by the system they are using.

7 Acknowledgements

The material is based upon work supported by NASA under award No: NAG5-11981. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Aeronautics and Space Administration.

References

[1] R. Anderson, Personal communication, San Jose, CA. 2003

[2] J. Rasmussen, "Skills, rules, and knowledge; signals, signs and symbols, and other distinctions in human performance models." *IEEE Transactions on Systems, Man, and Cybernetics*, Vol 13, No. 3, pp. 257-266, 1983

[3] C. E. Shannon, "A mathematical theory of communication", *Bell Systems Technical Journal*, Vol 27, pp. 379-423, 623-656, July and October 1948.

[4] G. W. Thomas, M. Reagan, et al., "Analysis of science team activities during the 1999 Marsokhod rover field experiment: implications for automated planetary surface exploration", *Journal of Geophysical Research-Planets*, Vol 106, No. E4, pp. 7775-7783, 2001