A Comparison of Visual Statistics for the Image Enhancement of FORESITE Aerial Images with Those of Major Image Classes

Daniel J. Jobson[†] Zia-ur Rahman[‡], Glenn A. Woodell[†], Glenn D. Hines[†],

[†]NASA Langley Research Center, Hampton, Virginia 23681
[‡]College of William & Mary, Department of Applied Science, Williamsburg, VA 23187

ABSTRACT

Aerial images from the Follow-On Radar, Enhanced and Synthetic Vision Systems Integration Technology Evaluation (FORESITE) flight tests with the NASA Langley Research Center's research Boeing 757 were acquired during severe haze and haze/mixed clouds visibility conditions. These images were enhanced using the Visual Servo (VS) process that makes use of the Multiscale Retinex. The images were then quantified with visual quality metrics used internally within the VS. One of these metrics, the Visual Contrast Measure, has been computed for hundreds of FORESITE images, and for major classes of imaging—terrestrial (consumer), orbital Earth observations, orbital Mars surface imaging, NOAA aerial photographs, and underwater imaging. The metric quantifies both the degree of visual impairment of the original, un-enhanced images as well as the degree of visibility improvement achieved by the enhancement process. The large aggregate data exhibits trends relating to degree of atmospheric visibility attenuation, and its impact on the limits of enhancement performance for the various image classes. Overall results support the idea that in most cases that do not involve extreme reduction in visibility, large gains in visual contrast are routinely achieved by VS processing. Additionally, for very poor visibility imaging, lesser, but still substantial, gains in visual contrast are also routinely achieved. Further, the data suggest that these visual quality metrics can be used as external standalone metrics for establishing performance parameters.

1. INTRODUCTION

During August and early September, 2005, the research Boeing 757 of NASA Langley Research Center was used to conduct a series of test flights under the Follow-On Radar, Enhanced and Synthetic Vision Systems Integration Technology Evaluation (FORESITE) program to test and demonstrate a number of new aviation safety technologies. One of our goals in participating in FORESITE was to acquire a high volume of aerial imagery. In particular, we were interested in acquiring imagery under visibility conditions that mimicked situations under which the pilot relies on instrumentation to fly the plane rather than the visibility of the terrain. The imagery that was acquired in these poor visibility conditions was needed to test a new metric for visual quality, both before and after non-linear image enhancement using the Visual Servo (VS)[1] and Multiscale Retinex (MSR) methods.[2–5] The VS adds an active visual quality measurement and feedback control element to the passive MSR processing, and extends performance from wide dynamic range imaging to encompass all imaging conditions including the very narrow dynamic range associated with poor visibility conditions such as fog, haze, rain, snow, dust, and dim light. The MSR has been implemented in hardware to provide real-time, video enhancement of flight data.[5–7] This hardware was flight-tested during FORESITE, and the experimental results and a description of the hardware are provided in a companion paper.[8]

The FORESITE experiments were conducted during the height of late summer stagnant air conditions. As a result, visibility along the flight paths between the Langley Air Force Base, Hampton, VA and the NASA Wallops Flight Facility, Wallops Island, VA was seriously impaired by conditions that ranged from severe haze to combinations of severe haze interspersed with patchy clouds. Human observers on the test flights routinely

Contact: DJJ: daniel.j.jobson@nasa.gov; ZR: zrahman@as.wm.edu; GAW: glenn.a.woodell@nasa.gov; GDH: glenn.d.hines@nasa.gov;

reported that the ground terrain was only very faintly visible, and on several flights there was no visibility of the ground terrain. For the FORESITE flights in which we participated, such conditions were prevalent at altitudes greater than 2000 feet, but did not exist for lower altitudes. Therefore we were able to acquire considerable image data at or exceeding the extremes of visibility. Data were collected from a total of eight flights in the form of 100–200 images per flight together with limited anecdotal human assessments of visibility conditions. The images were taken with a Nikon D1 digital color (still) camera shooting through the side window of the aircraft. From each flight, 50 images representative of very poor visibility were selected and enhanced with VS processing. This formed the basis for computing large aggregate visual statistics for pre- and post-enhancement imagery for one (visual) metric, the Visual Contrast Measure (VCM). The same measure was then applied to large aggregates of image data for other major image classes: consumer images, NOAA aerial photos of hurricane damage, underwater images, and Earth and Mars orbital images. The VCM was computed for both pre- and post-enhancement imagery in all cases. This provides the most general context possible for analyzing and interpreting the performance of the enhancement process and the suitability of the VCM as an overall metric of that performance.

The VCM is an internal, primary metric that the VS uses for "smart" control of MSR enhancement.[1] In this paper, however, we examine the VCM as a stand-alone *external* metric that can be used to determine visual quality. Computationally, the VCM is given by

$$\mathrm{VCM} = 100 * R_v / R_t \tag{1}$$

where R_v is the number of regions in an arbitrary image that exceed a specific threshold for regional signal standard deviation, and R_t is the total number of regions into which the image has been divided. Therefore VCM scores should be interpreted as answering the question: "How much of the image frame has good contrast?" A VCM score of zero simply means there were no regions of good contrast in that specific image. As we will see, quite a number of FORESITE images did possess zero or near-zero scores. So, in this crude sense, we can conclude that we were able to capture imagery largely near, or at, the limits of visibility. We will also see that the FORESITE VCM scores are much lower than the scores for other (obviously) poor visibility classes—underwater imaging and NOAA aerial surveys of hurricane damage taken during unavoidable haze conditions.

For all of the large aggregate statistics to be presented, attempts were made to avoid any bias due to a specific sensor or mission factor. Therefore, imagery for the various classes was selected to be as inclusive as possible. Earth orbital imagery came from Quickbird, IKONOS—the highest quality data perhaps—, astronaut photography of earth, and Landsat, and included wide ranging atmospheric conditions. The best imagery that we were able to obtain was acquired during very clear atmospheric conditions, while the worst was hazy but not totally obscured. Extremely obscured images are likely to simply be left out of public archives, so it should be understood that extremes of visibility are not likely to be encountered in these archives. With this understanding, we can still frame a context for understanding the FORESITE data where we were intentionally trying to get the "worst" case possible. The Mars image data were compiled from all existing Mars orbital image sensors—Mars Global Surveyor, Mars Odyssey, and Mars Express. For all classes diverse subject matter, terrain, etc., were included to attempt to avoid any statistical biases due to subject matter. Subject matter or scenes which are feature-impoverished were not included as these represent the "moot" case where there isn't anything to enhance or measure.*

2. TYPICAL EXAMPLES AND OVERALL VISUAL ASSESSMENT

In a companion paper,[9] we show the impact of VS processing on a more extensive collection of FORESITE and NOAA hurricane damage aerial photos. So in this paper we will give just a few typical examples from these and other image classes to convey the connection between the VCM scores and visual judgment. In Figure 1 we show examples that cover the full range of VCM scores and allow the reader to judge that the scores track degrees of visual quality in a useful quantitative way.

The examples show the usefulness of the VCM score, but also expose a shortcoming that arises because the VCM was originally designed as a control function for the VS. This shortcoming is that the VCM score tends

*See the section on "Data Sources" for the websites that we used.

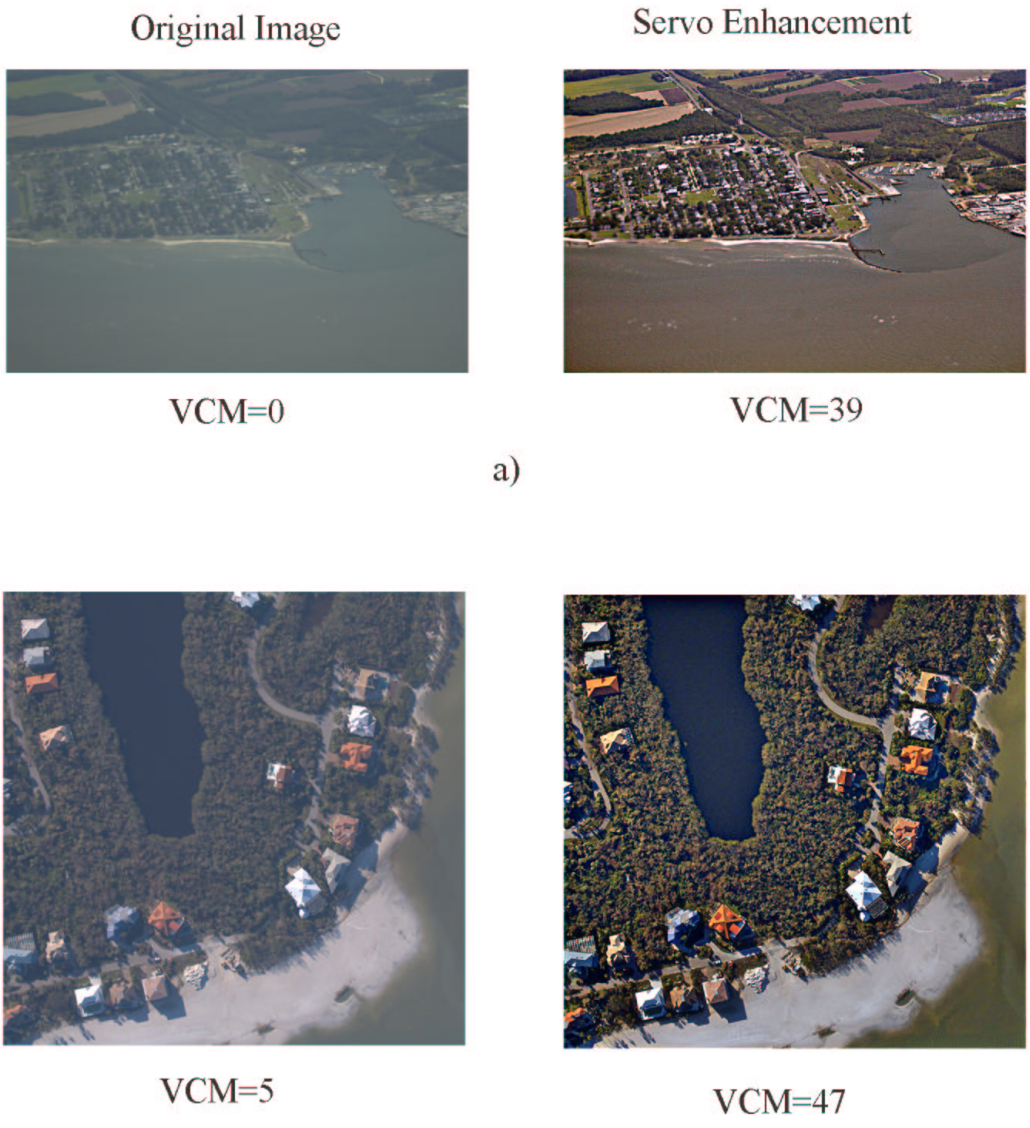

Figure 1. Typical examples of images and their VS enhancements: (a) FORESITE data with VCM score of zero; (b) NOAA Hurricane Wilma damage aerial photo with single-digit VCM score. The VCM scores for the VS processed image are much higher than the VCM scores for the unprocessed data. The correlation between the VCM scores and the visual quality is clear from the visual impact on scene visibility.

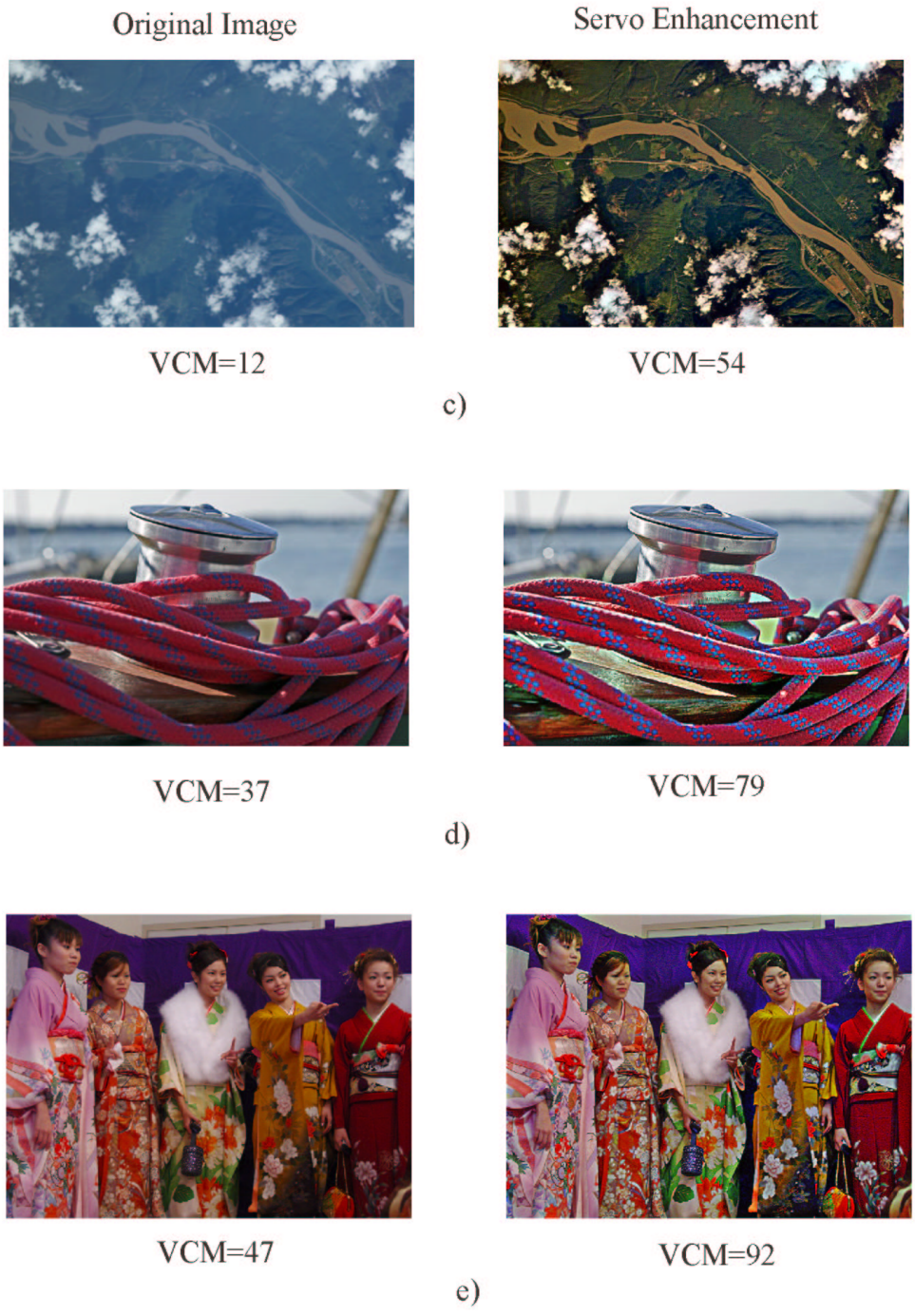

Figure 1. (continued) More examples of images and their VS enhancements (c) International Space Station Earth observations photo; (d) lower quality consumer digital photo; (e) higher quality consumer digital photo. The un-enhanced image in (e) is typical of the best contrast that was encountered in all of the classes of imagery that we sampled. The VCM scores for the VS processed image are much higher than the VCM scores for the unprocessed data. The correlation between the VCM scores and the visual quality is clear from the visual impact on scene visibility.

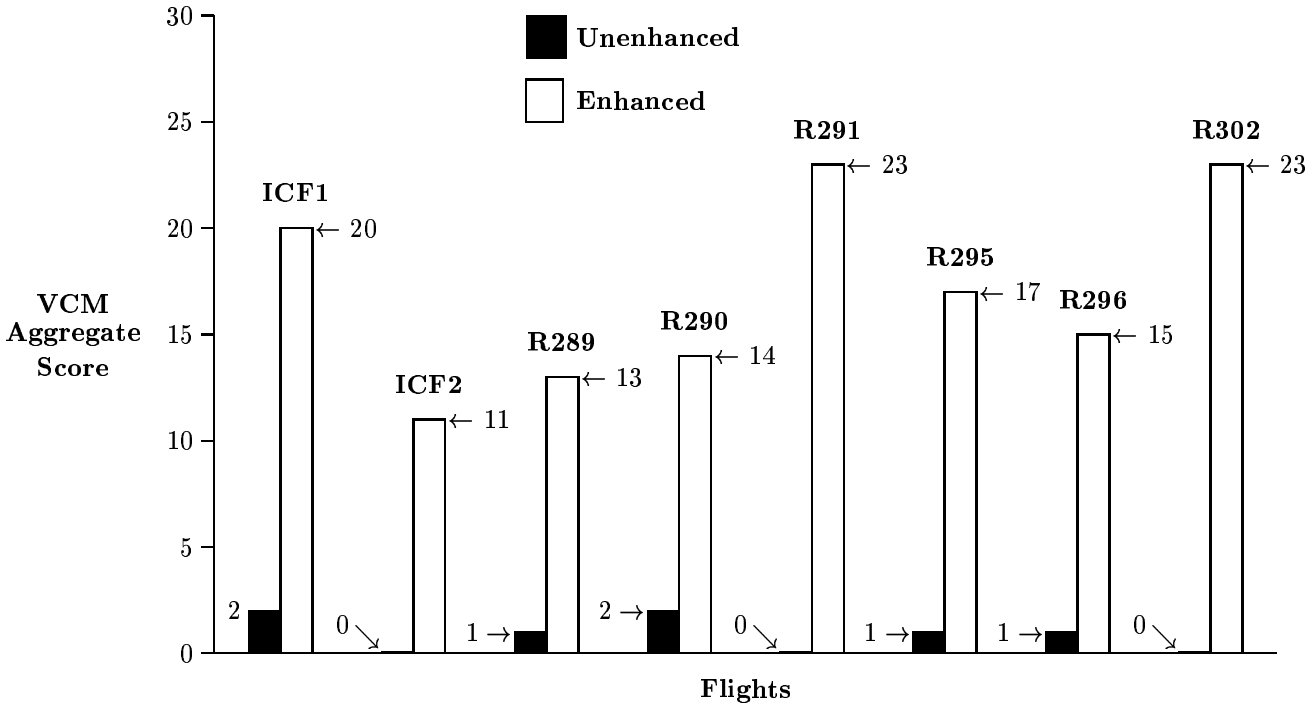

Figure 2. Summary of FORESITE flight data statistics. The individual flights are labeled by their official designations assigned by the FORESITE program.

to zero even when there is still some weak scene visibility, and that the perceived visual quality rises somewhat faster than the VCM scores at the low score range. In other words, a slow increase in the VCM scores from zero into single digits, and then low teens, is accompanied by a dramatic increase in (perceived) visual quality. Beyond this initial sluggishness, as the score continues to rise, visual quality tracks the VCM in a visually logical manner. So VCM scores > 10 do track visual judgment reasonably well. A score between 0–9 in the following large aggregate data sets should be interpreted with this behavior in mind.

3. LARGE AGGREGATE STATISTICS FOR FORESITE DATA

The eight data acquisition flights in the FORESITE experiments exhibit overall consistency in the VCM statistics. The variations in the data reflect mostly a range from poor visibility to extremely poor visibility. As we will show later, the VCM scores of *both* original and VS enhanced imagery for FORESITE are lower than we encountered in any other image class we analyzed. The summary of the eight flights is shown in Figure 2.

Each bar in Figure 2 represents 50 images selected from the larger set of about 100–200 images taken during flight. Other imagery that was acquired either on ground or during takeoff and landing phases of the flight exhibits significantly better visibility because of the much lower altitude. Such imagery was not used in our analysis since we were interested in analyzing conditions with the poorest visibility. VCM scores for the original images range between 0–2, but most were between 0–1 indicating very poor visibility. VS enhanced scores range between 11–23 with an average of 17. Since the original scores were at 0–2 with an average of 1, the average increase in score due to the servo enhancement process is 16. Note that we have previously discussed that visual quality rises rapidly relative to the VCM score advancing from single digits to teens, so this amount of increase goes with a relatively large perceived increase in visual quality for the VS enhanced FORESITE imagery. However, we will see that this falls short of the increase in VCM scores achieved by VS enhancements of the other major images classes. This smaller score increase for FORESITE imagery is consistent with our previous work on post-enhancement visibility limits.[1] The extremely poor visibility conditions during FORESITE flights result in sensor noise being a relatively larger component of image signal variation, and this noise sets the ultimate limit on post-enhancement visibility improvement.

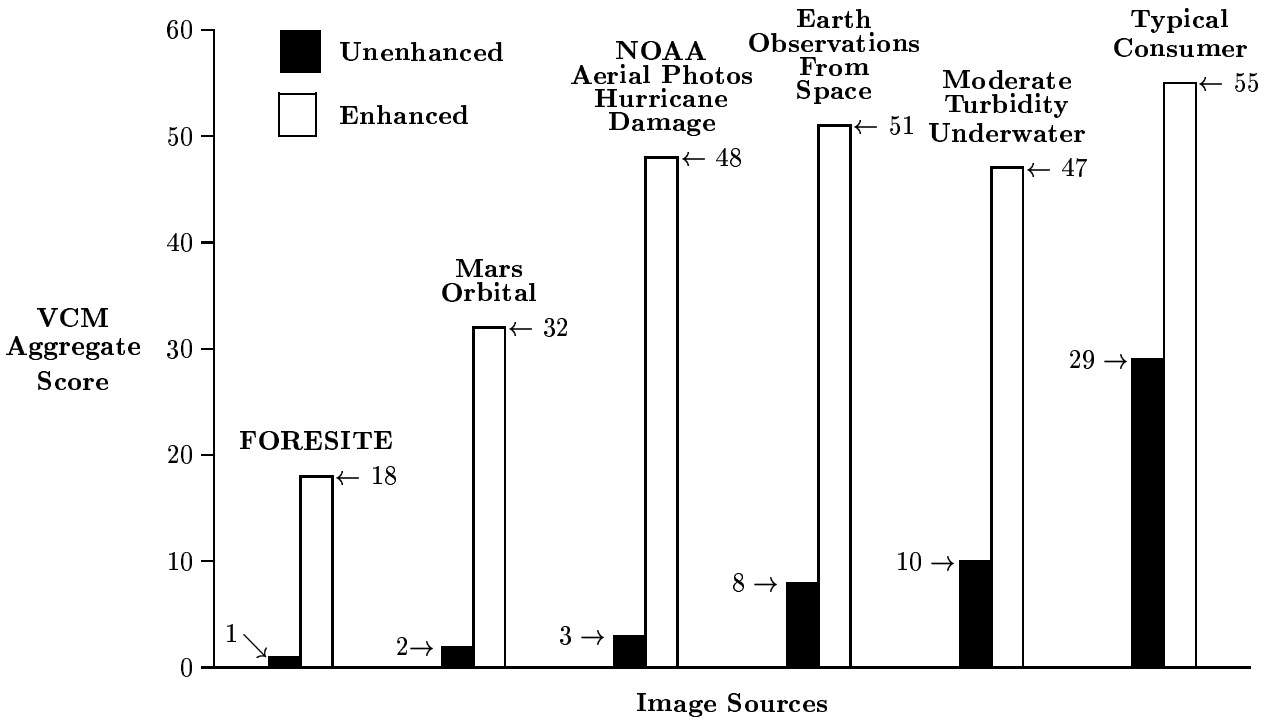

Figure 3. Comparison of FORESITE data with major classes of imagery. The data was obtained with 100–200 images from each image class.

4. COMPARISON OF LARGE AGGREGATE STATISTICS FOR MAJOR IMAGE CLASSES

We now summarize the FORESITE data in the context of similar data for other image classes. For the other image classes, the aggregate data was computed from 100–200 images within each class, with very diverse scene content. The image classes were selected in order to elicit interesting scientific comparisons, and (hopefully) find a plausible explanation for any trends—especially unexpected ones—that emerge. Because the image classes represent all major classes that we thought would exhibit possible statistical differences, we are confident that the results are general in nature. The VCM aggregate scores for both original and VS enhanced data are shown in Figure 3.

Of all the image classes, the FORESITE data, both pre- and post-enhancement, has the lowest aggregate VCM scores. This is not at all surprising since one of our objectives for FORESITE was to acquire poor visibility imagery. Further, while images with this poor a visibility may often be acquired, they do not often make it into image archives since viewers of galleries are not interested in looking at basically blank frames. So the FORESITE flights did fulfill our need for some very extreme visibility image data not available from other sources.

The big surprise to us was that the next lowest scoring class of both original and enhanced data was the Mars orbital imagery from three different on-going planetary missions—Global Surveyor, Odyssey, and the ESA's Mars Express. We do not think it plausible that any Mars atmospheric phenomena is responsible for the poor imagery: only rare images show any signs of dust storms and none of these were included in this data sample. Rather, we favor the explanation that Mars imagery lacks a dimension intrinsic to most terrestrial imaging, namely, reflectance variation. While some reflectance changes occur on the Martian surface, their contribution to image signals is relatively minor compared to the contribution due to topographical variations. This impoverishment of reflectance variation on Mars, and the rather static nature of the planet (images of a Mars location taken years apart are usually rather similar), are our best explanation for the data trend we show in Figure 3.

The underwater image class provided a surprise of a different sort. The VCM scores for underwater imagery are much higher in both pre-and post-enhancement cases than was expected. We certainly would have expected underwater turbidity to be even more severe and limiting than atmospheric turbidity, yet the underwater image data was on par with the Earth orbital imaging class statistics. We suspect that this may be explainable as being due to bias beyond our control: underwater photographers probably take images only in relatively low turbidity conditions, and, either do not take shots in worse conditions or else omit the "blank" frames from their archives. Thus, the underwater data that we were able to access is only representative of the visibility range over which images are likely to be taken and posted in galleries.

As we did expect, however, the consumer image class is the highest scoring, again for pre- and post-enhancements. This simply reflects the fact that most everyday photos are not taken in extreme turbidity conditions or very dim lighting.

The NOAA hurricane damage aerial imagery deserves some discussion. These images should *not* be considered representative of aerial photography as a whole. The urgency of the need for damage assessment in the wake of the major hurricanes requires that photography flights be made as soon as possible after the hurricane. During the peak of the hurricane season this means flights are scheduled as soon as cloud cover is past. This is also, though, a time of very high humidity and significant amounts of haze. The hurricane data is an aggregate of aerial photos of damage by hurricanes Katrina, Wilma, Ivan, Ophelia, Dennis, and Rita, so the data can be considered representative of the visibility in the aftermaths of hurricanes during the peak season and high humidity conditions.

Finally, an interesting overall trend is obvious from the data shown in Figure 3. Most image classes do approach, or achieve, a post-enhancement score of 45–55. The only exceptions are Mars orbital imagery, and the FORESITE data. Therefore, for most terrestrial imaging, even from Earth orbit, we can conclude that the VS enhancement performs well and approaches a maximum achievable contrast score. The FORESITE data falls short of this only because it is an extreme poor visibility case, and sensor noise has begun to limit the magnitude of contrast enhancement achievable with the processing. Even so, the VS enhancements of the FORESITE data do represent major visibility improvements, especially when we recall that visual quality rises very rapidly as the VCM scores increase from zero through single digit values on into values in the teens which is the case for the FORESITE data.

5. CONCLUSIONS

A study of the Visual Contrast Measure intrinsic to the Visual Servo (active) image enhancement processing suggests that it can also serve as a stand-alone metric of image quality. For the VCM scores to be interpreted properly as a visual metric, we found that visual quality rises rapidly as the VCM scores rise from zero into the teens, but thereafter seems to track visual assessments in a more linear manner. Further, we see room for improving the VCM as a stand-alone visual metric by extending the current cut-off of zero to a lower range. This can be accomplished by softening the sharp cutoff in the original thresholding and classification process.

The analysis of the large aggregate statistics for FORESITE flight data and other major images classes supports some additional conclusions. The FORESITE data was the poorest visibility image data that we have encountered and satisfied our image acquisition goal for the flight experiments. The VS enhancements of the FORESITE data did substantially improve visibility as quantified by significant increases in the VCM scores. The overall trends in the large aggregate statistics support the idea that the Visual Servo for most imaging approaches or achieves a maximum visual contrast possible and that this coincides with VCM scores of 45–55. The exceptions to this were Mars orbital imaging and the FORESITE data. The former may be limited because the imagery is largely topographic, while the latter is limited because sensor noise is a larger component of image data in extreme poor visibility imaging.

Image Data Sources

1. NOAA aerial photos of hurricane damage: `http://www.ngs.noaa.gov/index.shtml`
2. Mars orbital images:

(a) Mars Global Surveyor MOCS Narrow Angle Images: `http://www.msss.com/moc_gallery/`

(b) Mars Express: `http://www.esa.int/esa-mmg`

(c) Mars Odyssey: `http://mars.jpl.nasa.gov/odyssey/gallery/images.html`

3. Earth Orbital Images:

   (a) Gateway to Astronaut Photography `http://eol.jsc.nasa.gov/`

   (b) Human Spaceflight(Shuttle and International Space Station) `http://www.spaceflight.nasa.gov/gallery/index.html`

   (c) Quickbird `https://www.digitalglobe.com/sample_imagery.shtml`

   (d) IKONOS: `http://www.spaceimaging.com/gallery/default.htm`

4. Consumer Images:

   `http://www.pbase.com`

5. Underwater Images:

   (a) `http://www.dancingfish.com/main.php`

   (b) `http://diver.net/kathy/`

   (c) `http://www.ramblincameras.com/Rcamindex.htm`

Acknowledgments

The authors wish to thank the Synthetic Vision Sensors element of the NASA Aviation Safety Program for the funding which made this work possible. In particular, Dr. Rahman's work was supported under NASA cooperative agreement NNL04AA02A.

REFERENCES

1. D. J. Jobson, Z. Rahman, and G. A. Woodell, "Feature visibility limit in the nonlinear enhancement of turbid images," in *Visual Information Processing XII*, Z. Rahman, R. A. Schowengerdt, and S. E. Reichenbach, eds., Proc. SPIE 5108, 2003.
2. D. J. Jobson, Z. Rahman, and G. A. Woodell, "Properties and performance of a center/surround retinex," *IEEE Trans. on Image Processing* **6**, pp. 451–462, March 1997.
3. D. J. Jobson, Z. Rahman, and G. A. Woodell, "A multi-scale Retinex for bridging the gap between color images and the human observation of scenes," *IEEE Transactions on Image Processing: Special Issue on Color Processing* **6**, pp. 965–976, July 1997.
4. Z. Rahman, D. J. Jobson, and G. A. Woodell, "Retinex processing for automatic image enhancement," *Journal of Electronic Imaging* **13**(1), pp. 100–110, 2004.
5. G. D. Hines, Z. Rahman, D. J. Jobson, and G. A. Woodell, "DSP implementation of the multiscale retinex image enhancement algorithm," in *Visual Information Processing XIII*, Z. Rahman, R. A. Schowengerdt, and S. E. Reichenbach, eds., pp. 13–24, Proc. SPIE 5438, 2004.
6. G. D. Hines, Z. Rahman, D. J. Jobson, and G. A. Woodell, "Real-time enhanced vision system," in *Enhanced and Synthetic Vision 2005*, J. Verly, ed., pp. 13–24, Proc. SPIE 5802, 2005.
7. G. D. Hines, Z. Rahman, D. J. Jobson, and G. A. Woodell, "Single-scale retinex using digital signal processors," in *Proceedings of the GSPx*, 2004.
8. G. D. Hines, Z. Rahman, D. J. Jobson, and G. A. Woodell, "Real-time enhancement, registration, and fusion for an enhanced vision system," in *Enhanced and Synthetic Vision 2006*, J. Verly and J. J. Guell, eds., Proc. SPIE 6226, 2006.
9. G. A. Woodell, D. J. Jobson, Z. Rahman, and G. D. Hines, "Advanced image processing of aerial imagery," in *Visual Information Processing XV*, Z. Rahman, S. E. Reichenbach, and M. A. Neifeld, eds., Proc. SPIE 6246, 2006.