

MASTER: A MOBILE AUTONOMOUS SCIENTIST FOR TERRESTRIAL AND EXTRA-TERRESTRIAL RESEARCH

Iain Wallace⁽¹⁾, Mark Woods⁽²⁾

⁽¹⁾ SCISYS, 23 Clothier Road, Bristol, BS4 5SS, UK, Email: iain.wallace@scisys.co.uk

⁽²⁾ SCISYS, 23 Clothier Road, Bristol, BS4 5SS, UK, Email: mark.woods@scisys.co.uk

ABSTRACT

We present results from the ESA MASTER project to develop and study a domain-independent software architecture to identify scientific phenomena in images. For example an autonomous detection component deployed on a planetary rover could be used to automatically detect known targets of interest, survey the geological characteristics of an area while in transit or take full advantage of serendipitous events by detecting novel phenomena. MASTER builds upon existing machine learning, classification and machine vision techniques to advance the state-of-the-art in space autonomy. The paper includes outline assessments of the power and memory requirements for the proposed algorithms obtained from a dedicated field trial.

1. INTRODUCTION

Bandwidth or communication limitations may make real-time control of instruments for scientific discovery difficult or impossible. For planetary rovers there is a trade-off between detailed observation to ensure targets are not missed, which requires slow traverses to downlink all data, and maintaining sufficient progress to visit many science targets. In the Earth Observation (EO) domain interesting transient events may be missed if they appear in low resolution images but these are not processed for higher resolution targeting in time. Similarly, bandwidth to process the sheer amount of EO data may be limited, meaning novel data are overlooked. MASTER focussed on the problems of detecting (and defining) novelty, and the need for expert scientist input to be used to train the detector.

In the most general case, regions of an image may be novel for several reasons:

- A known type of phenomenon is detected, but it is in some way different to others seen before.
- Unknown phenomena are observed, unlike any seen before.
- Known phenomena are observed but in a context unseen before.

To address this detection problem we present the MASTER architecture, a novel combination of visual

saliency, feature extraction and machine learning.

Although the presented system is generic with respect to input for evaluation we focussed on imager data from EO or a planetary rover. We present results based on MERIS Reduced Resolution Earth Observation data and planetary rover imagery from MER/MSL PANCAM and the ESA SAFER and SEEKER field trial campaigns.

1.1. Background

The challenge of improving the science return of space missions by science identification and thus capitalising on opportunistic science has received much attention in recent years. Researchers have been interested in identifying the occurrence of certain, pre-defined indicators of scientific interest, in identifying unexpected or anomalous features and in using this information to prioritise images for transmission or to provide autonomous sensor placement (see [1], [2],[3],[4]).

Science-directed autonomous operation is largely made possible by advances in the machine vision and learning communities and as such still represents a nascent field in space exploration. Typically, systems in this domain have used different vision and data analysis techniques to detect and prioritise known classes of interest. An example is the variety of approaches developed to detect, segment and characterise rocks and other geological features from imagery obtained by planetary rovers (see, for example, [5]–[8]). We summarise here the most prominent published approaches to autonomous science discovery.

ASE: The Autonomous Sciencecraft Experiment (ASE) flying on the Earth Observing One (EO-1) spacecraft was designed with the remit to autonomously detect and respond to dynamic, scientifically interesting events observed from low earth orbit such as thermal anomaly detection, cloud detection and flood scene classification [9].

OASIS: The Onboard Autonomous Science Investigation System (OASIS)[10] was developed to evaluate and autonomously act on

data gathered by platforms such as planetary landers and rovers.

AEGIS: The AEGIS system provides automated targeting for (narrow field-of-view) remote sensing instruments on the MER mission.

CREST: Work carried out as part of the U.K. STFC CREST initiative set out to develop an autonomous science system in the context of ExoMars [11]. Compared to related works this aimed to tackle a broader geological remit by characterising a target’s structure, texture and composition as well as other parameters.

Science on the Fly: One of the most advanced (in terms of the vision/learning aspects) approaches[12] and evaluated during field trials in the Amboy Crater Lava Field.

Compared to these approaches the key innovation of MASTER is the goal of truly autonomous novelty detection, rather than identification of a priori known targets and conditions. The MASTER project was conceived with a desire to mitigate against this limitation by considering for science deployment potential targets of unknown type as well as unconventional targets of known type.

2. THE MASTER ARCHITECTURE

Figure 1 shows an overview of the MASTER image processing pipeline which serves to explain the processing steps. First a saliency map is calculated, to identify locally outstanding regions of the image. This is used as the input to a segmentation algorithm to extract areas of interest for classification. A collection of classifiers trained on an expert-labelled dataset are then applied to each segment. Finally the confidence measures from each classifier are used by a novelty detection algorithm to assign a novelty class (or known

label) to each segment. Thus the final output from the system is the location and extent of any novelty detected, and additionally the class of novelty present. In the example here, three rocks stand out in the saliency map, which is used to produce a segmentation highlighting the two nearest. These are then classified, with “rock” being the most likely outcome, but below some defined level of certainty. As such, the final output describes each rock as “classified novelty, rock”. This can loosely be interpreted as “a rock, but unusual compared to those in training”.

The MASTER software architecture was designed to implement this processing pipeline. It allows for each component to be modified and evaluated in isolation. In such a complex system to investigate and understand the performance of the whole it is necessary to decompose and analyse the parts. Figure 4 shows the main components of the architecture. Broadly, the structure is to have a main controller component to manage the flow of data through the subcomponents, running tests and generating evaluation data. This allows for testing of components in isolation or together, with different discrete implementations.

2.1. Training Component

The training component encapsulates the store of training datasets – expert annotated images – and various means of partitioning these datasets into test and training sets. Meaningful evaluation of the system requires tests based on expertly labelled imagery to evaluate system performance and so the training component must support this and various partitioning schemes. For example, it can k-fold the datasets to provide many test/training sets for statistical analysis of classifiers and then also provide test sets with guaranteed novel phenomena not present in training.

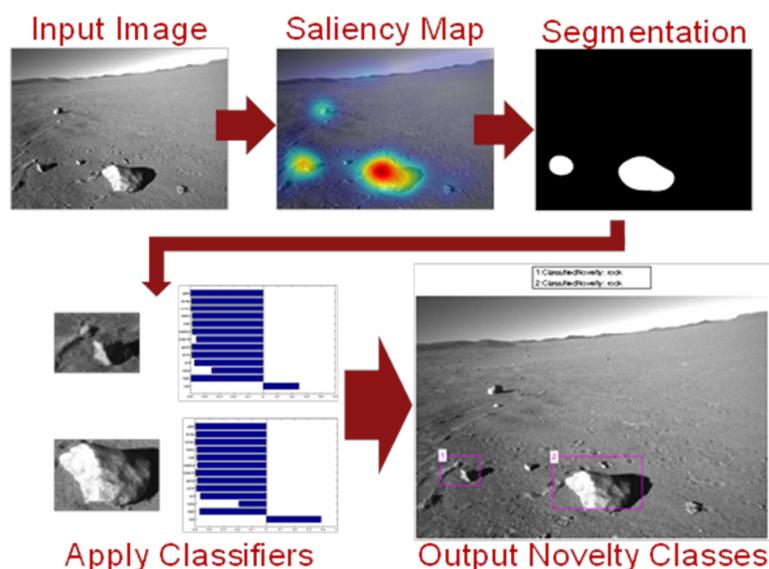


Figure 1 - Image Processing Pipeline Example

2.2. Saliency Map Generation

We wished to evaluate several algorithms, as the saliency map generation effectively gates input to the rest of the system and provides one limit on performance (a non-salient region will never be identified as novel). For the final system, we implemented several algorithms and allowed the system to automatically choose the best performing algorithm and operating point for segmentation based on the training data. Figure 2 illustrates some of the differences between two such algorithms applied to the same image.

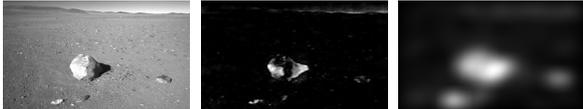


Figure 2 - Example Saliency Maps

2.3. Classifiers

Although we refer to this component as the classifier component, we also include the feature extraction and generation of appropriate feature vector. We primarily tested with one-versus-rest SVM classification based on libSVM [13] and using a Histogram of Oriented Gradients (HOG) [14] descriptor with a bag-of-visual-words based spatial histogram used to create the feature vector. As with the saliency component, the design is modular, and allowed experimentation with other feature vectors and classifiers, including Aggregate Channel Features (ACF) with AdaBoost decision-tree classifiers [15] and Integrate Channel Features (ICF) [16] in conjunction with SVMs.

2.4. Novelty Detection

The dedicated novelty detection part of the system is relatively simple, as it relies primarily on the classifier output – indeed the entire architecture could be considered as novelty detection. We primarily

considered two concepts of novelty. In both cases a strong confidence from a single classifier was considered not-novel, with the class of (non) novelty being assigned based on the frequency of occurrence in the training data. In the simple case, novelty is represented by there being no single confident classifier. In the more complex group novelty classification scheme we also consider a set of confident classifiers to represent a novel class (confusion) and a different type of novelty where there is a class with a large margin over the rest but not certain.

3. SYSTEM EVALUATION

Evaluating a system such as MASTER is a complex topic due to the breadth of problems tackled. To attempt to define performance, one must first know what the system's goal is. For MASTER as defined, this prompts some key questions, which may have different answers depending on the application:

- What is novelty, in relation to the training data?
- How important is the detection of novelty versus its location in an image?
- How desirable is it to classify regions of interest that are not novel?
- Where is the trade-off between detecting novelty and the cost of false positives?
- Is the problem one of selection – “is this

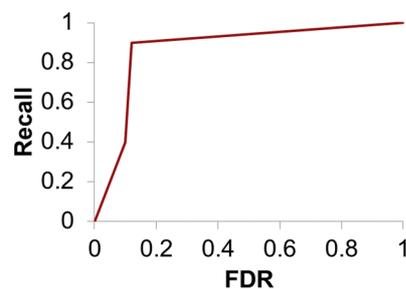


Figure 3 - An Example Recall-FDR Graph. Best performance would be top left.

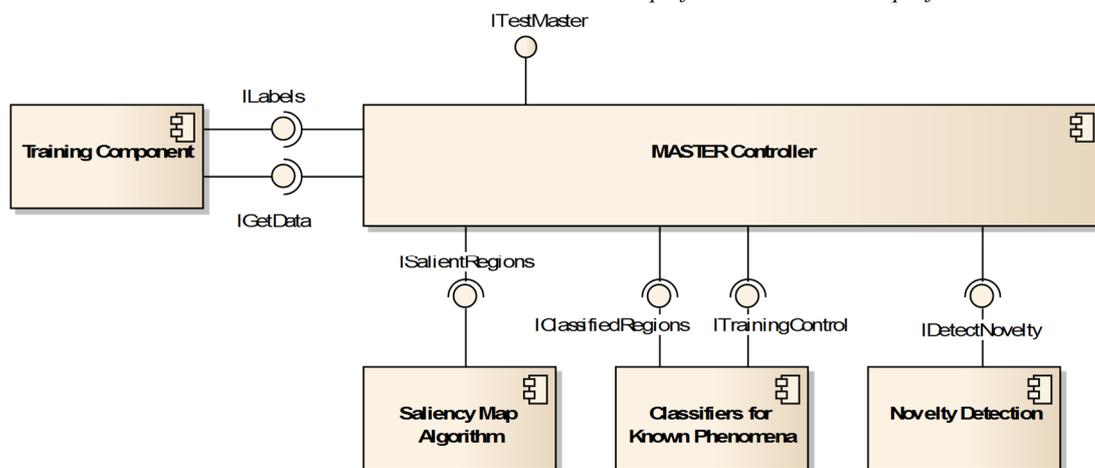


Figure 4 - The MASTER Architecture

image novel?”

- Is the problem one of ranking – “which is the most novel image?”

This list implies trade-offs, emphasising one aspect of performance often impacts on another.

An initial thought might be to conflate the last two points, one could simply select the most novel (by some threshold) images as novel - but they speak to two different system goals. Consider the difference between a rover which may only downlink 1 in 100 images, so one must be selected and the others discarded, and a system to process EO data which should rank images to direct human attention to the most critical. In the first case the penalty for false positives is high, whereas in the second if all images are to be inspected eventually, we would admit more false alarms to identify more images of interest sooner.

To cover all these criteria would require different metrics, and they have different measures of success against which they could be measured. Regardless of the metric there are desirable properties of any measure of performance:

- It should allow comparison between different MASTER systems. The MASTER project did not just aim to create one system, but investigate and research different configurations and their effect on performance.
- It should allow for configuration of the system to trade off detecting novelty against the chance of false positives (or other similar concession).
- Performance should depend on the input data and the system approach, and not be affected by implementation details. E.g. a definition of a negative result based on the number of areas of an image evaluated would result in a penalty for a system that chose a smaller window size.
- It is desirable to have an easily understood transparent measure..

These points and looking for a good general measure of system performance led us to plots of Recall against False Discovery Rate (FDR, [17], [18], though in both papers they describe FDR as 1-Precision). To briefly summarise this measure, it is based on a confusion matrix which considers positive (novelty) detections, negative results and if they are correct (true) or not (false). Based on the numbers of true and false positive and negatives for a test dataset we can calculate various performance measures. As we vary parameters of a system, we can plot a performance curve for comparison

3.1. Evaluation datasets

For evaluation we focussed on two test domains – Earth Observation (EO) and Planetary Science (PS) rover imagery, with several datasets annotated by expert scientists. The EO imagery used was from ESA MERIS Reduced Resolution datasets, processed into both pseudo true-colour images and false colour chlorophyll images. These were annotated by a domain expert with a focus on phytoplankton blooms as the phenomena of interest. The PS imagery consists of several datasets from the NASA/JPL Photojournal Planetary Image Archive and rover imagery from the ESA SEEKER [19] and SAFER [20] Mars-analogue field trials. Images are annotated by an expert geologist with outcrops and layered rock formations as the phenomena of interest. The datasets are completely labelled, as this allows using a portion for training the system and the remainder as ground truth for system evaluation. In the results below we use three PS datasets “SK” from SEEKER, “PSsafer” from SAFER and “PS” from the NASA/JPL data. There is one large “EO” dataset.

3.2. Saliency Component Evaluation

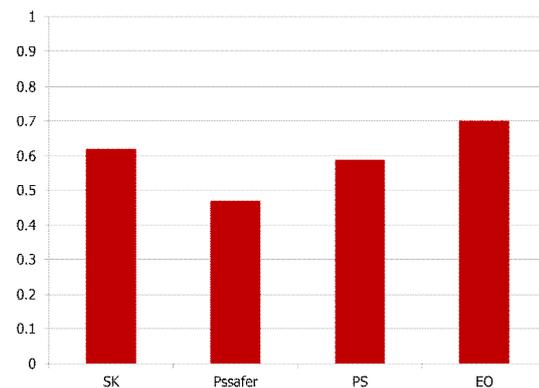


Figure 5 - Saliency AUC Scores. Comparing mean AUC scores for the test data, calculated from ROC curves. Values represent the chance a given pixel is correctly marked salient.

We only present here the aggregate results of tests of ten saliency algorithms across all datasets, this leads to the apparently poor AUC scores in Figure 5. It is important to realise maximum performance is higher than this, and that the AUC score represents an average over possible operating points and this can be set for higher precision at the expense of recall or higher recall at the expense of false positives depending on application. It is useful to consider performance in terms of this trade-off for MASTER, as the saliency step represents a gate on performance for the rest of the system – it is used to pre-segment for classification. The absolute performance scores are not important, only the effectiveness of a given segmentation method using saliency as input. These numbers suggest we are likely to admit many false positive pixels as salient, and so segmentation will be important.

Qualitatively, we can see that the saliency algorithms are capable of providing useful output for novelty detection in practice. Figure 7 shows navigation images from the SAFER trial of a feature later named “Carnot”. The feature clearly shows up in the saliency map, and thus could be selected for classification and identified as of interest. In the actual trial, Carnot was not detected till two Sols of operation later when by chance a PANCAM image captured the edge of the trench feature. The above images were processed after the fact on navigation images that are normally discarded, highlighting the potential for a MASTER system to identified features of interest that would otherwise be missed.

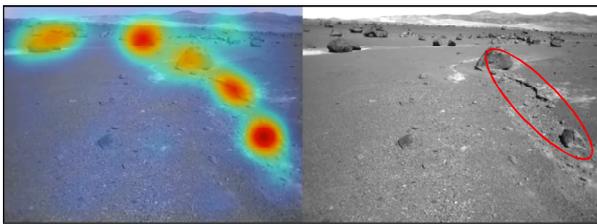


Figure 7 - "Carnot" NAVCAM Image with Saliency Map Applied

3.3. Classifier Evaluation

Evaluating classifiers in isolation is a straightforward task. ROC curves and AUC scores are widely accepted methods for the purpose of evaluating classifiers. The AUC score gives a particularly intuitive measure, as it can be interpreted as representing the chance a random segment will be correctly classified.

The classifiers work well in at a high level, which is an

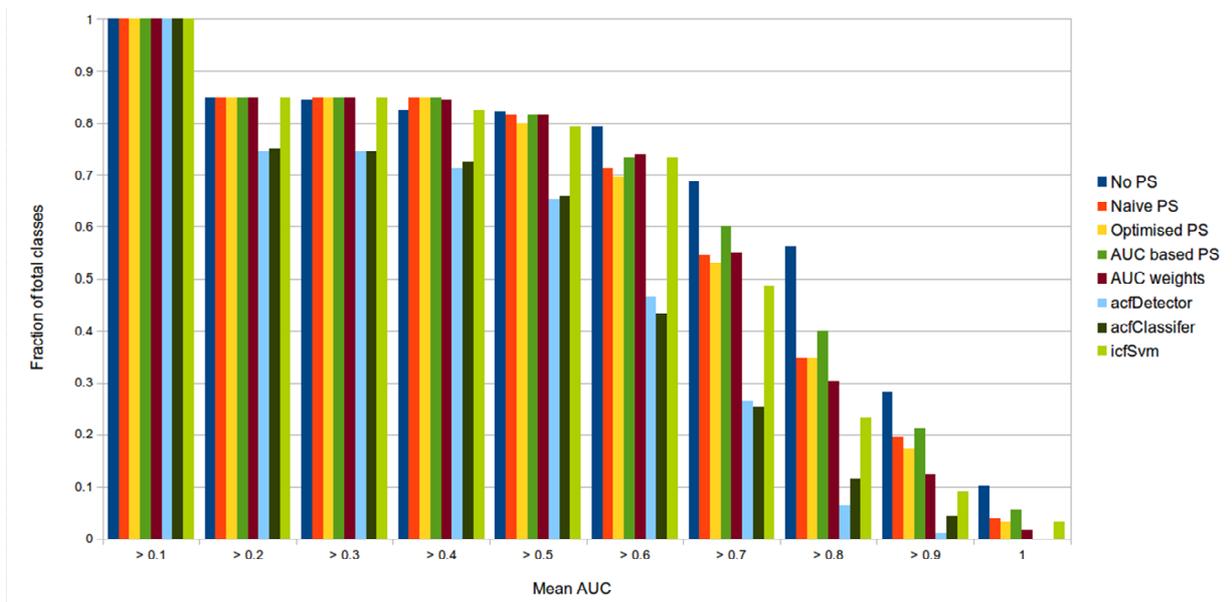


Figure 6 - Summary of classifier performance across all classes and domains.

unexpected result of applying such general purpose classifiers to the space domain. Figure 6 shows aggregate results from several experiments to test different parameter selection algorithms and classifiers across all datasets using k-fold cross validation. The class mappings and datasets are not the final ones, but they do include substantial PS, EO and Seeker data, only the SAFER data is not included. To interpret the graph read the mean AUC threshold along the bottom, the bar indicates the fraction of all classes across all datasets that achieved that AUC score or better. To evaluate these results pessimistically, they show that for all classifiers or feature vectors tried, in all datasets, most classes (>0.5) are correctly classified most of the time (AUC >0.5). Performance is much better than this in the main, several implementations achieve mean AUC scores over 0.7 in most classes.

Performance was best where there is homogeneity between test and training data – ideally field of view and camera used should be the same seen in the PS SAFER and SK results). Classes that were better defined and more homogenous also performed better as would be expected, and in these cases as few as 25 training examples was sufficient for good performance.

3.4. System Evaluation

Figure 8 shows graphs illustrating the performance of complete systems under varying novelty and saliency thresholds, representing levels of “trust” for the classifiers and saliency respectively. High FDR with varying novelty suggests that the saliency gating is defining performance as more discriminating classifiers (higher threshold) do not reduce FDR. This is backed up by varying saliency threshold showing movement on the

FDR axis.

The top left point with high recall and low FDR appears to be good performance, but is an outlier showing the danger of generalised metrics. Whilst performance is good by this measure, more detailed analysis showed it represents classifying the image as novel (or not) correctly but the position accuracy was poor.

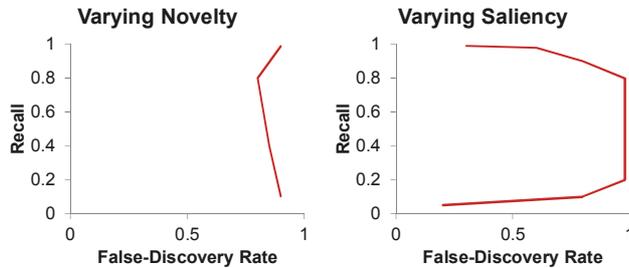


Figure 8 - Summary of Complete System Performance

4. APPLICATIONS AND FUTURE WORK

For future development of the MASTER system as-is there are a several potential avenues of investigation and application:

1. Class definitions and labelling.
2. Saliency based segmentation.
3. Classifier choice and feature vectors.
4. Novelty detection algorithms.

The first point would be to address capturing of expert knowledge, perhaps introducing a hierarchical classification structure. Secondly, the experimental results suggested that improving segmentation based on the saliency might decrease the error rate of the complete system. Finally there are many more potential classifiers and feature vectors that could be investigated, similarly novelty detection algorithms.

One of the main outcomes from MASTER was that to improve performance it is necessary to narrow the scope of the problem with an application focus. This would

have some implications for the many performance trade-offs required in the system. We see three key potential avenues for development where results are promising:

- Assisted Labelling
- Directing Attention
- Development on planetary exploration flight hardware.

The first point comes out of our experiences working with the ExoMars instrument science teams on the SAFER field trial, where there was a need to allow collaborative organisation and labelling of the returned data. This was at an early stage in the MASTER project, yet we were able to use an early prototype of the labelling tool which was well received. The good performance of the classifiers suggests a MASTER system to aid a human user in suggesting likely labels for identified regions of interest.

Especially in the EO case, directing attention for a higher resolution imager or further acquisition is an interesting case – the focus would be on location of novel events rather than just their detection – a different emphasis to refine development.

Finally, there is potential to focus on a subset of the planetary exploration case and develop a system capable of running on flight hardware. This comes directly from some exciting results out of the UKSA Chameleon field trials, expanded in the next section.

4.1. Power and Execution Time

The recent UK Space Agency (UKSA) Chameleon project involved rover field trials in the Atacama Desert Martian analogue site, returning to the same area as the earlier ESA Seeker trials [21]. As part of these trials we were able to apply some of the technology developed during MASTER for image classification and analysis.

Table 1 - Summary of Power and Memory Measures from Chameleon Trial. *The highlighted row shows the cost of applying four feature extraction and classification operations as described in this document to a NavCam image. Energy cost was measured by execution time and recorded CPU power consumption reported by internal registers.*

Data Product or test	Product Size (Mb)	Component Memory Usage (Mb)	Energy cost over idle (Joules)
Inertial Measurement Unit Pose Reading (IMU)	0.00008	43	0.001
Rectified Image pair (lower res. 1Mp)	0.3	166	0.1
3D Map to Occupancy Grid for nav planning	0.05	87	0.2
Rectified Image pair (High res. 2.3Mp)	1.12	190	0.4
Visual Odometry Estimate (low res, 1Mp)	0.003	93	1.5
Machine Learning Based Terrain Analysis (1 frame)	0.0001	85	1.7
Visual Odometry Estimate (High res, 2.3Mp)	0.003	142	3.8
Low Resolution Point Cloud, (7m range)	7	122	4.7
generating map low res	0.175	Not recorded	13.4
generating map high res	0.14	Not recorded	39.2
High Resolution Point Cloud, (30m range)	26	228	92.1

We captured CPU power consumption data for various algorithms and sensor processing and so we are able to compare the runtime performance and power costs of the classifiers presented here to other common (and less common) rover sensors and GNC algorithms. **Error! Reference source not found.** shows a summary, sorted by energy cost. Note that these are preliminary numbers on research software and not a rigorous test, but they are novel results and suggest promising future work. In particular memory and energy costs can only be realistically compared with each other, and not external systems. The key result is that the terrain analysis process – which involved running four classifiers and feature extractors – has a power cost on the same order as a Visual Odometry (VO) estimate. VO is already in use on flight rovers, and these results show that whilst the training part of machine learning algorithms is computationally expensive, execution is well within the realms of feasibility for flight hardware.

5. CONCLUSIONS

The use of field trial data from UKSA and ESA trials greatly enhanced our results, supplying a large amount of relevant test and labelling data. To summarise our

results:

- We varied the novelty threshold, recall changed but FDR was similar.
- The implication is noisy inputs, as we showed the classifiers perform well in isolation.
- We varied the saliency threshold, showing the expected shift in the FDR axis.
- This combined with saliency results suggest that better segmentation may improve performance in the general case.
- Qualitatively however saliency shows great performance, for example in the Carnot discovery example.

In conclusion, the MASTER project shows a promising future for computer vision and machine learning approaches in the space domain. The system prototyped here could be developed into a broad range of potential applications ranging from flight systems on planetary rovers to labelling and annotation tools to support terrestrial scientists.

6. REFERENCES

- [1] R. Castaño, M. Judd, R. C. Anderson, and T. Estlin, "Machine learning challenges in Mars rover traverse science," in *2003 ICML Workshop on Machine Learning Technologies for Autonomous Space*, 2003.
- [2] D. R. Thompson, T. Smith, and D. Wettergreen, "Information-optimal selective data return for autonomous rover traverse science and survey.," in *ICRA*, 2008, pp. 968–973.
- [3] D. Barnes, A. Shaw, P. Summers, R. Ward, M. Woods, M. Evans, G. Paar, and M. Sims, "AUTONOMOUS IMAGE BASED LOCALISATION FOR A MARTIAN AEROBOT," presented at the ISPRS Symposium, Paris., 2006.
- [4] D. R. Thompson, W. Abbey, A. Allwood, D. Bekker, B. Bornstein, N. Cabrol, R. Castaño, T. Estlin, T. Fuchs, and K. Wagstaff, *Smart cameras for remote science survey*. Citeseer, 2012.
- [5] T. L. Roush, "Essential autonomous science inference on rovers (EASIR)," in *Aerospace Conference, 2004. Proceedings. 2004 IEEE*, 2004, vol. 2, pp. 790–800.
- [6] D. Thompson, S. Niekum, T. Smith, and D. Wettergreen, "Automatic detection and classification of features of geologic interest," in *Aerospace Conference, 2005 IEEE*, 2005, pp. 366–377.
- [7] H. Dunlop, "Automatic rock detection and classification in natural scenes," Carnegie Mellon University, 2006.
- [8] D. R. Thompson and R. Castaño, "Performance comparison of rock detection algorithms for autonomous planetary geology," in *Aerospace Conference, 2007 IEEE*, 2007, pp. 1–9.
- [9] S. Chien, R. Sherwood, D. Tran, B. Cichy, G. Rabideau, R. Castaño, A. Davis, D. Mandl, B. Trout, S. Shulman, and others, "Using autonomy flight software to improve science return on Earth Observing One," *J. Aerosp. Comput. Inf. Commun.*, vol. 2, no. 4, pp. 196–216, 2005.
- [10] R. Castaño, T. Estlin, R. C. Anderson, D. M. Gaines, A. Castaño, B. Bornstein, C. Chouinard, and M. Judd, "OASIS: Onboard autonomous science investigation system for opportunistic rover science," *J. Field Robot.*, vol. 24, no. 5, pp. 379–397, 2007.
- [11] M. Woods, A. Shaw, D. Barnes, D. E. Price, D. Long, and D. Pullan, "Autonomous science for an ExoMars Rover-like mission," *J. Field Robot.*, vol. 26, no. 4, Apr. 2009.
- [12] D. R. Thompson, D. S. Wettergreen, and F. J. C. Peralta, "Autonomous science during large-scale robotic survey," *J. Field Robot.*, vol. 28, no. 4, pp. 542–564, 2011.
- [13] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, 2011.
- [14] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, 2005, vol. 1, pp. 886–893.
- [15] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast Feature Pyramids for Object Detection."
- [16] P. Dollár, Z. Tu, P. Perona, and S. Belongie, "Integral Channel Features.," in *BMVC*, 2009, vol. 2, p. 5.
- [17] S. Agarwal, A. Awan, and D. Roth, "Learning to detect objects in images via a sparse, part-based representation," *Pattern Anal. Mach. Intell. IEEE Trans. On*, vol. 26, no. 11, pp. 1475–1490, 2004.
- [18] S. Agarwal and D. Roth, "Learning a sparse representation for object detection," in *Computer Vision—ECCV 2002*, Springer, 2002, pp. 113–127.
- [19] M. Woods, A. Shaw, E. Tidey, B. Van Pham, U. Artan, B. Maddison, and G. Cross, "SEEKER-AUTONOMOUS LONG RANGE ROVER NAVIGATION FOR REMOTE EXPLORATION," in *I-SAIRAS 2012*, Torino, Italy, 2012.
- [20] M. Woods, A. Shaw, I. Wallace, Malinowski, Mateusz, and Rendell, Phillip, "Simulating Remote Mars Rover Operations in the Atacama Desert for Future ESA Missions," presented at the 13th International Conference on Space Operations, 2014.
- [21] M. Woods, A. Shaw, and I. Wallace, "THE CHAMELEON FIELD TRIAL: TOWARD EFFICIENT, TERRAIN SENSITIVE NAVIGATION," presented at the ASTRA 2015, 2015.