QUASI-THEMATIC FEATURE DETECTION AND TRACKING FOR FUTURE ROVER LONG-DISTANCE AUTONOMOUS NAVIGATION

Affan Shaukat [(1)], Conrad Spiteri [(1)], Yang Gao [(1)], Said Al-Milli [(1)], and Abhinav Bajpai [(1)]

[(1)] Surrey Space Centre, University of Surrey, Guildford, Surrey, GU2 7XH, (United Kingdom),
Email: {a.shaukat, c.spiteri, yang.gao, s.al-milli, a.bajpai}@surrey.ac.uk

ABSTRACT

This paper investigates state-of-the-art approaches for object detection and tracking employing models that can efficiently detect objects (specifically '*rock*' on planet surfaces) in the visual scene in terms of semantic descriptions. Two models (i.e., *"visual saliency"* and *"blob (shape-based) detection"*) are presented here specifically focused towards future planetary exploration rovers. We believe that these two object detection techniques will abate some of the algorithmic limitations of existing methods with no training requirements, lower computational complexity and greater robustness towards visual tracking applications over long-distance planetary terrains. Comprehensive (quantitative) experimental analysis of the proposed techniques performed using three challenging benchmark datasets (i.e., from PANGU, RAL Space SEEKER and SSC lab-based test-bed) will be presented in this paper.

Key words: **Planetary rovers, autonomous visual navigation, object detection and tracking, thematic features**

1. INTRODUCTION

The use of autonomous robotic platforms has seen a tremendous growth over the past five decades of planetary exploration missions. There is a diversity of space exploration technologies used in these missions (e.g., orbiting spacecraft [1], space telescopes [2], stationary landers [3, 4] etc.), however there is an increasing recognition that planetary rovers form one of the most important sources of exploratory information in that they provide greater mobility, ability of physical experimentation, autonomous navigation and microscopic level of observations. Over the past few decades, research in planetary rovers has evolved from the most primordial technology to more complex autonomous and intelligent systems.

Autonomous visual navigation for planetary rovers is one of the most popular research topics for space roboticists, e.g., those involved in MER, ExoMars and MSL missions [5]. Such systems are foreordained to operate in remote and sometimes hostile environments where they are subjected to a number of key constraints that deems the imperativeness of long hours of autonomous operation with resilience to subtle environmental, technical and physical perturbation. Furthermore, remote-controlled operation is implausible due to limitations within the propagation velocity of electromagnetic waves. The uncertainty within the complex operating environment of an autonomous rover quite firmly underpins the conjecture of employing vision-based algorithms that quantify perceptual inputs in computationally efficient and simplistic manner.

Most existing visual modelling techniques utilised by rovers use saturated texture features to describe and track objects in the visual scene [11], which tend to have high computational load and will lose reliability over long distance in remote and uncertain environments such as planetary surfaces. This paper investigates object detection and tracking techniques based on quasi-thematic features (in contrary to the more conventionally used point-based descriptors, e.g., '*SIFT features*') [6] in order to reliably detect and track objects (e.g., '*rocks*') from a sequence of planetary 2D images. Two proposed solutions based on the state-of-the-art algorithms (i.e., saliency and blob (shape-based) detection) are presented. These techniques will abate some algorithmic limitations of existing methods with no training requirements, lower computational complexity and greater robustness towards applications over long-distance planetary terrains.

The first proposed approach employs a ***'visual saliency'*** [7] model that uses multi-modal stimuli (i.e., colour, orientation, and intensity, etc.,) for the segregation (i.e., detection) of conspicuous regions (such as rocks) in the visual scene. These visually salient regions are then tracked and labelled throughout subsequent frames via an instance-based search algorithm using a distance metric. The second proposed approach uses ***'shape-based (blob-like) descriptors'*** for modelling objects, where image patches that define *'region of interest'* (ROI) are delineated via spatial filtering based on binary thresholding and edge detection. An optimal set of image patches are then selected that comprise higher level of thematic descriptions for objects (i.e., *'rocks'* in the current application) based on image moments. This is followed by tracking and labelling analogous blobs between two subsequent frames using Hu set of invariant moments.

This paper will set out to quantify whether the two proposed approaches can appropriately detect and track planetary surficial rocks potentially facilitating future long-distance autonomous rover navigation.

2. OBJECT DETECTION AND TRACKING IN SPACECRAFT NAVIGATION

The success of future planetary missions greatly depend upon the ability of rovers to autonomously traverse long distances preferably at higher speeds in order to achieve its mission objectives. This requires the rovers to efficiently detect possibly dangerous path manoeuvres, highly accurate path planning, and intelligent mediums of interactions with scientists and ground stations. Since a great deal of these technologies is based on visual sensory inputs, computer vision will continue to play an important role in majority of all space exploration missions.

Although the applications of automated computer vision algorithm vary from precision descent and landing to efficient orbits including contour mapping and terrain visualization, real-time autonomous path planning and localisation to surface exploration including collision detection and obstacle avoidance [8, 9], however, our current discussion will specifically focus towards image processing techniques useful for autonomous navigation, localisation and path planning in autonomous rovers (others fall beyond the scope of this paper). We begin with a succinct literature survey of the more commonly used object detection and tracking methodologies found in literature.

Visual object detection and tracking has been carried out on planetary exploration missions using either stereoscopic or monocular cues. There has been some success in using stereo vision [8] for obstacle avoidance by employing various techniques such as, 'feature matching' [10, 11] and 'relational matching' [12, 13]. Tracking in stereo vision is usually applied for stereoscopic visual odometry and is more formally approached in stochastic sense, such as, using *'maximum likelihood estimation'* [14], *'particle filtering'* [15], *pyramidal 'Pseudo Normalized Cross Correlation' (MER-VO algorithm)* [16], displacement prediction via *'dense optic flow'* [17] or *'probabilistic graphical models'* [18].

Rock detection and tracking remains an important capability within autonomous planetary rovers, although primarily used for obstacle avoidance, visual odometry and autonomous long distance visual navigation, it also helps facilitate the rover and scientists to investigate and preform image analysis on specific objects of interest. Various modelling techniques have been used in such applications, for instance, '*Multiple Viola-Jones (MVJ)'* originally based on a supervised learning algorithm (utilising *AdaBoost*) [19], *'Support vector machines (SVMs)'* that use point-based (pixel-level) descriptions for classification, (i.e., windows around pixel intensity values are either classified as rock or non-rock [20]), and also using *'edge-based descriptors'* [21]. Some of the recent advancements have enabled carrying out highly detailed autonomous mineral content analysis on Martian surficial rocks [22].

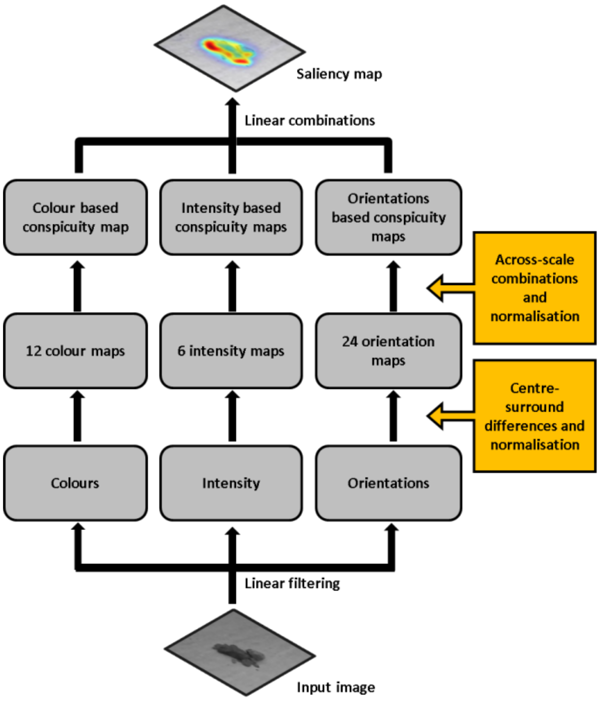

Saliency map
Linear combinations
Colour based conspicuity map
Intensity based conspicuity maps
Orientations based conspicuity maps
Across-scale combinations and normalisation
12 colour maps
6 intensity maps
24 orientation maps
Centre-surround differences and normalisation
Colours
Intensity
Orientations
Linear filtering
Input image

Figure 1. Itti-Koch-Niebur visual saliency model [7]

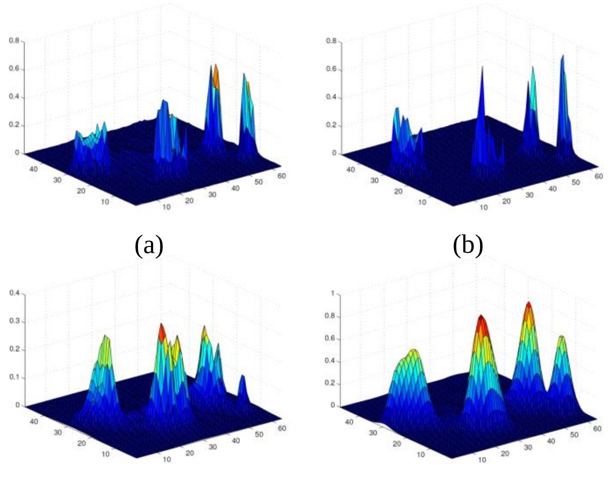
3. PROPOSED OBJECT DETECTION AND TRACKING TECHNIQUES

Our problem in this paper is thus to detect and track (recognition is beyond the scope of the current work) rocks on the surface of planets using appropriate *unsupervised* modelling techniques that can identify and describe *'regions of interest'* (ROI) in terms of quasi-thematic features (labelled as detected rocks). This can be achieved either on the basis of their visual saliency within the input scene, or using image moments (i.e., Hu's set of invariant moments). We implement these two distinct strategies as follows.

3.1. Visual Saliency Based Detection and Tracking

Specific surface characteristics of objects on planetary surfaces (such as rocks) may provide sufficient information for it to be distinguished in the visual scene. As such computer vision paradigms that use descriptions of objects in terms of their visual saliency to segregate them from their periphery seems to be a feasible approach in the aforementioned situation. Thus we propose to perform rock detection via visual saliency based semantic description of objects using the ***'Itti-Koch-Niebur'*** saliency model [7], followed by tracking via the k-nearest neighbour instance-based search algorithm (eliminating explicit model training).

The Itti-Koch-Niebur saliency model builds on the architecture by [23, 24] and relates to human visual search strategies [25]. It is a purely bottom-up model of visual attention for selection of conspicuous regions in the visual scene without top-down control.



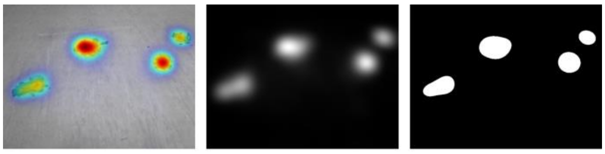
(a) (b)

(c) (d)

Figure 2. Conspicuity maps for a sample image generated for colour (2a), intensity (2b), orientation (2c), combined into the final saliency map (2d)

The model is a cascade of multimodal feature processing steps, commencing with linear filtering of the input image into three topographical feature maps (i.e., feature extraction), followed by saliency-based processing for identifying conspicuous regions (creating conspicuity maps) within each of the three feature maps (i.e., feature processing), and finally, combining the three conspicuity maps into a single (final) saliency map for local conspicuity over the entire visual scene [7] (refer to Fig. 1).

Details of the Itti-Koch-Niebur saliency model are explained in [7]; however we highlight some of the main steps involved. The process begins with the extraction of the early visual features via linear filtering (feature extraction). Original image is subsampled and low-pass filtered using Gaussian pyramids at nine spatial scales. Using the RGB colour channels of the input image an intensity image is obtained. The first set of feature maps are based on ***'intensity contrast'***. A total of 6 intensity feature maps are computed using centre-surround differences (via across-scale differentiation). Similarly a second set of feature maps is generated for the ***'colour channels'***. Taking into account the *'red/green and green/red'*, *'blue/yellow and yellow/blue'* double opponency, a total of 12 colour feature maps are generated in this process. A third set of feature maps is generated from the intensity image using oriented Gabor pyramids ($\theta \in \{0°, 45°, 90°, 135°\}$). The ***'orientation feature maps'***, define the local orientation contrast between the centre and the surround scales. This process generates a total of 24 orientation feature maps. This is followed by combining individual types of feature maps into "conspicuity maps" in order to signify visually salient regions. In order to over come the complexities of non-comparable modalities, dynamic ranges, noise obscuration within the direct combination of the whole set of 42 feature maps, [7] proposes a map normalisation operator, $\mathcal{N}(.)$, which promotes maps with few strong peaks, whereas suppressing those with many comparable peaks:



(a) (b) (c)

Figure 3. Original input image with saliency heat maps (3a), followed by saliency-based illustration of the ROIs (3b), and the final binarised ROIs obtained via the Otsu's method (3c)

- Normalisation of the map to range *[0, …, M]*
- Finding the global maximum *M*
- Computation of the average $\overline{m}$ of all local maximum *M*
- Multiply the map by $(M - \overline{m})^2$

Combination of feature maps into 3 *"conspicuity maps"* (i.e., $\bar{\boldsymbol{I}}$ for intensity, $\bar{\boldsymbol{C}}$ for colour and $\bar{\boldsymbol{O}}$ for orientation, refer to Fig. 2), is carried out via across-scale addition. The final saliency map is generated by normalisation followed by summation of the three conspicuity maps [7],

$$S = \frac{1}{3}\left(\mathcal{N}(\bar{\boldsymbol{I}}) + \mathcal{N}(\bar{\boldsymbol{C}}) + \mathcal{N}(\bar{\boldsymbol{O}})\right) \qquad (1)$$

The next step is to track the visually salient regions over the subsequent frames (images). Saliency maps generated specify the regions of interest (ROI) in the visual scene. They characterise salient objects (which on a homogeneous planetary surface will point towards *rocks*). In order to convert them to ROI patches with hard boundaries segregating them from the background, we perform histogram shape-based image thresholding. We use Otsu's method [26] to reduce our saliency map to a binary image with the assumption that the saliency maps have a bimodal distribution, that is, two classes of pixels; the salient image pixels (ROIs) and the background (pixels describing non-salient regions).

This method essentially follows an exhaustive search strategy (discriminant form of pattern recognition technique) to compute the optimum threshold that minimises the intra-class variance or maximising the inter-class variance [26]:

$$\sigma_\beta^2 = \omega_1(\tau)\omega_2(\tau)\big(\mu_1(\tau) - \mu_2(\tau)\big)^2 \qquad (2)$$

Where, $\omega_1(\tau)$ and $\omega_2(\tau)$ are the probabilities of the two classes (C_1 and C_2) respectively ($\tau \in \{1,2, \dots ,256\}$ represents any level within the full range of gray level histogram values). In most cases, the Otsu's [26] method iteratively computes the optimum by

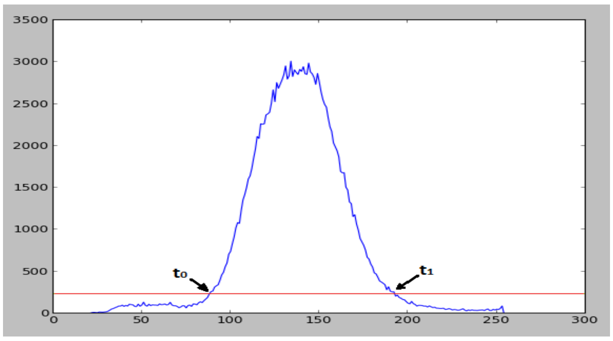


Figure 4. Image histogram showing two threshold points: t_0 and t_1

maximising Eq. 2, more formally,

$$\sigma_\beta^2(\eta^*) = \underset{1\leq\eta\leq K}{\operatorname{argmax}}\, \sigma_\beta^2(\eta) \qquad (3)$$

A flavour of the results obtained using this method is illustrated in Fig. 3.

Tracking these binary patches (ROI blobs) forms an important part of our experimental process. The use of binary blobs (as ROI descriptors) as compared to complex contextual features simplifies our tracking problem. This involves using an instance-based search algorithm that does not require any explicit model training or *a priori* information regarding the dataset. We use the k-NN search algorithm [27] that collates salient regions throughout subsequent frames by applying Euclidean norm (*ℓ^2 norm*) as the distance metric. The concept is based on the standard k-NN classification paradigm where the pairwise distances are measured between the centroids of salient ROI patches in the current frame and the centroids of the patches in the previous frame. Similar objects are labelled as a single object over subsequent frames. Thus the current method performs quasi-clustering of similar objects using a spatiotemporal based similarity check and thus achieves tracking.

The centroid of the detected salient object (ROI blob) in the current frame 't' is set as the *reference point* (R_l^t where $l = \{0, 1, \dots, n\}$ is the *object id*), followed by exhaustively searching for the centroid of the detected salient object in the previous frame '$t-1$' (the *query point*) (Q_γ^{t-1} where $\gamma = \{0, 1, \dots, n\}$ is the *object id*), such that the Euclidean distance between the two centroids is a minimum. Pairs of detected objects identified by this process are associated with a common label ($\mathcal{L} = \{0, 1, \dots, n\}$) throughout subsequent frames to achieve tracking (i.e., *object id* of the ROI in the previous frame is associated with the ROI in the current frame), more formally,

$$\forall l \forall \gamma, \mathrm{kNN}(R_l^t) \rightarrow \mathcal{L}: \mathcal{L} = \underset{\gamma}{\operatorname{argmin}}\{\|R_l^t - Q_\gamma^{t-1}\|\} \qquad (4)$$

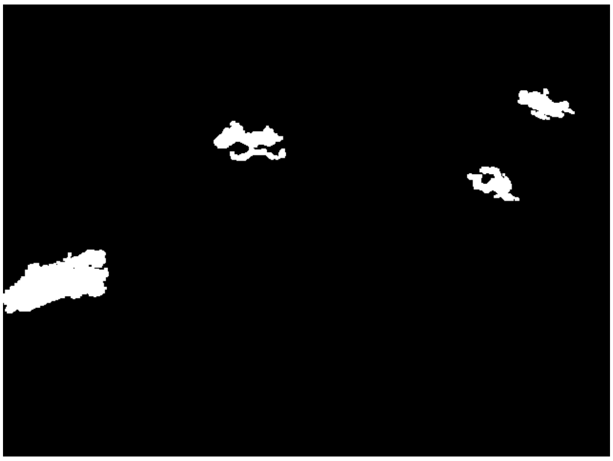


Figure 5. Binary image containing blobs

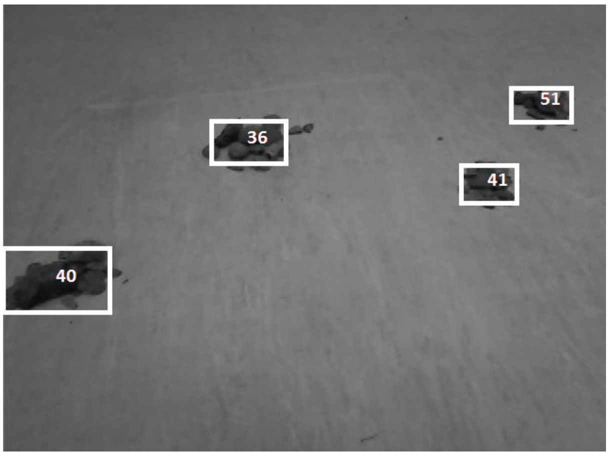


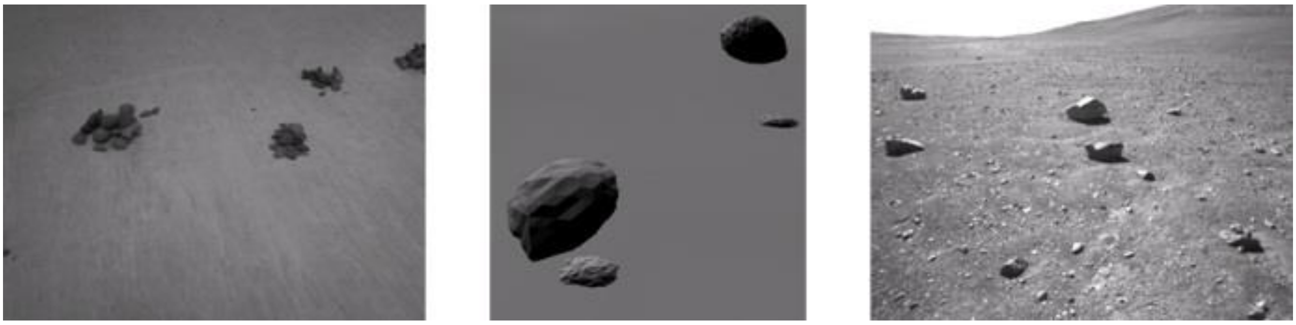
Figure 6. Sample image illustrating detected rocks along with tracked labels

3.2. Blob (Shape-based) Detection and Tracking

The '*blob detection and tracking'* algorithm is a cascade of a number of distinct processes. The original image is segmented via binerization using a threshold selection criterion. Contours of individual patches (i.e., blobs) are extracted using a border following method. Hu set of invariant moments are computed for each contour. These Hu moments (of each individual blob) within two subsequent frames are collated in order to achieve tracking.

For the image segmentation stage, we use the MAT algorithm [28]. The algorithm utilises a methodology that takes the advantage of local image statistics of *mean* and *variance* within a cluster and two thresholds obtained from the intensity distribution histogram. The algorithm uses a simple percentile (of the brightness) measurement procedure as follows [28]:

$$\int_{-\infty}^{a} p_r(t)dt = \mathcal{H}\% \qquad (5)$$



(a) (b) (c)

Figure 7. Example images from the Lab-based (7a), PANGU simulated (7b) and SEEKER (7c) datasets

Where $p_r(t)$ represents the intensity distribution of an input image and $\mathcal{H}\%$ is an arbitrary chosen percentage value. The threshold values 't_0' and 't_1' (refer to Fig. 4) are computed as follows,

$$\int_{-\infty}^{t_0} p_r(t)dt = \mathcal{H}\%, \qquad \int_{t_1}^{\infty} p_r(t)dt = \mathcal{H}\% \quad (6)$$

The values 't_0' and 't_1' are used in a multi-thresholding sequence to produce a binary image containing blobs that represent any rocks detected within the image (refer to Fig. 5), i.e., more formally,

$$dst_{(r,c)} = \begin{cases} 0 & if\ src_{(r,c)} > t_0 \\ 0 & if\ src_{(r,c)} < t_1 \\ 1 & otherwise \end{cases} \quad (7)$$

Where $dst_{(r,c)}$ is the output binary image, $src_{(r,c)}$ is the source image captured from the camera, r and c are the pixel row and column within the image, 't_0' and 't_1' are the calculated threshold values.

The blobs within the thresholded binary image are then indexed in order to allow operations to be performed on a single blob rather than the contents of the whole image. The binary image contains noise in the form of relatively smaller blobs. As such we eliminate the ones that are smaller than an *a priori* defined threshold (*outlier rejection*). This is achieved via the analysis of the blobs zeroth moment (i.e., the *area*). Smaller blob pixels are isolated and categorised as part of background. The general equation of a moment for a binary image with j rows and k columns is as follows,

$$\mathcal{M}_{vw} = \sum_{r=1}^{p}\sum_{c=1}^{q} r^v c^w \mathcal{I}(r,c) \quad (8)$$

Where $\mathcal{I}(r,c)$ in case of a binary image is (with '$\mathcal{B}$' defining a ROI within the image),

$$\mathcal{I}(r,c) = \begin{cases} 1, & if(r,c) \in \mathcal{B} \\ 0, & otherwise \end{cases}$$

The zeroth order moment (area) "*Ar*" of a binary image is computed as follows,

$$Ar = \sum_{r=1}^{p}\sum_{c=1}^{q} \mathcal{I}(r,c) \quad (9)$$

A border following method in the sense of [29] is performed on the individual blobs in order to extract their contours. The Hu set of (translation and rotation) invariant moments [30] are computed for these contours, yielding a vector of 7 values that describe the shape of each blob. An exhaustive search strategy is applied in order to carry out a comparison of the Hu moments belonging to each individual blob between two subsequent frames resulting in matched pairs. These matched pairs are then uniquely labelled throughout subsequent frames to achieve tracking (refer to Fig. 6). The matching is formalised as follows,

$$I(A,B) = \sum_{i=1}^{7} \left| \frac{1}{\mathcal{F}_j^A} - \frac{1}{\mathcal{F}_j^B} \right| \quad (10)$$

Where, $\mathcal{F}_j^A$ and $\mathcal{F}_j^B$ are defined as,

$$\mathcal{F}_j^A = sgn(h_j^A) \cdot \log|h_j^A| \text{ and } \mathcal{F}_j^B = sgn(h_j^B) \cdot \log|h_j^B|$$

The terms h_j^A and h_j^B are the Hu moments of objects *A* and *B* respectively.

4. EXPERIMENTAL OBJECTIVES

This paper aims to detect objects (specifically rocks) on planetary surfaces using two distinct modelling techniques in terms of semantic descriptions rather than point-based features. Detected objects are defined in terms of blobs that are tracked over subsequent frames using heuristic tracking techniques. The paper will examine the two techniques by comparing them against human annotated ground-truth datasets in terms of their detection and tracking accuracies using standard quantitative evaluation measures. ***Note:*** *For proof-of-concept evaluations we*

Table 1. Performance evaluation for the Itti-Koch-Niebur Saliency Detection/Tracking Method

Itti-Koch-Niebur Saliency Model								
Dataset	**N-MODA**	**MOTA**	**Avg Tp/Img**	**Avg Fp/Img**	**Avg Fn/Img**	**Avg Objs/Img**	**Miss rate**	**TPR**
LAB-based	0.80	0.80	3.83	0.22	0.74	4.79	0.15	0.85
PANGU	0.84	0.83	2.37	0.05	0.41	2.83	0.11	0.88
SEEKER	0.61	0.61	2.19	0.26	1.11	3.56	0.29	0.71

Table 2. Performance evaluation for the Blob (Shape-based) Detection/Tracking Method

Shape-based Detection Model								
Dataset	**N-MODA**	**MOTA**	**Avg Tp/Img**	**Avg Fp/Img**	**Avg Fn/Img**	**Avg Objs/Img**	**Miss rate**	**TPR**
LAB-based	0.75	0.71	4.36	1.36	0.08	5.80	0.01	0.99
PANGU	0.63	0.61	2.14	1.07	0.18	3.39	0.07	0.93
SEEKER	0.65	0.62	3.11	1.54	0.13	4.78	0.03	0.97

subsampled the original datasets into smaller subsets.

5. EXPERIMENTAL DATA COLLECTION

5.1. SSC Lab-based Test-bed

The dataset was recorded in our research lab at the Surrey Space Centre, University of Surrey. The collected dataset (from a sensor equipped 4-wheeled rover, traversing a flat surface littered with scattering of rock clusters) consists of low-level features including odometry data. External video scene is captured using a single on-board camera mounted with a known reference frame relative to the rover's pose. The dataset comprises a total of 230 frames (per frame monocular image size of 640×480 pixels, at 15-fps sampling). Refer to Fig. 7a.

5.2. PANGU

The dataset was generated using a combination of the Planet and Asteroid Natural scene Generation Utility (PANGU), (developed at the University of Dundee) and image capture software from the *PMSLAM* project at the SSC. PANGU simulates planetary environments using parameters such as the levelness of the terrain and the number, size and distribution of craters and boulders. The PMSLAM software has the ability to place a virtual camera at a given location, field of view and record images. The dataset comprises a total of 111 frames (per frame monocular image size of 512×512 pixels). Refer to Fig. 7b.

5.3. RAL Space SEEKER

The dataset is a small subset of the original data generated by the *'SEEKER'* consortium. It consists of rectified monocular images from the left camera *(a 'Bumblebee XB3'*) of a sensor equipped rover (providing additional low-level sensor data, such as, DGPS, IMU Data, DTM, VO and Path data). The experimental dataset comprises a total of 54 frames at 5-fps sampling (per frame monocular image size of 512×384 pixels). Refer to Fig. 7c.

5.4. Ground-Truth Annotation of Image Data

For each of the three experimental datasets, per image visual scene annotation (as well as labelling using numeric nomenclature) is carried out by two individuals in terms of observed objects (i.e., rocks). Both individuals use a common (planetary rock) annotation tool purposely built at the Surrey Space Centre. This results in laboriously annotated ground-truth datasets with bounding boxes encapsulating the regions of interest (i.e., rocks in the current experimental scenario) with negligible observed variation among the annotators (voiding the requirement of an interrater reliability study). These will serve as our test datasets.

6. EXPERIMENTAL EVALUATION AND RESULTS

We examine the detection *accuracy* of the proposed systems using the evaluation protocols similar to [31]. For any given frame 't', the number of *'false positives'* (fp_t), *'misses'* (ms_t) and *'true positives'* (tp_t) is calculated by measuring the spatial overlap between the ground-truth and the system output objects. If for a given frame 't', G_i^t is the *i*th ground-truth object and D_i^t is the i^{th} detected object then the spatial overlap ratio (OR_i^t) is calculated as,

$$OR_i^t = \frac{|G_i^t \cap D_i^t|}{|G_i^t \cup D_i^t|} \quad (11)$$

The detected object is considered a *true positive* for $OR_i^t \geq 0.2$ and *false positive* for $OR_i^t < 0.2$. Whereas any unmatched objects in the ground-truth set are considered *misses*. The Normalised Multiple Object Detection Accuracy (N-MODA) is computed for the entire image sequence of each dataset as follows,

$$\text{N-MODA} = 1 - \frac{\sum_{t=1}^{N_{frames}} \left(c_{ms}(ms_t) + c_f(fp_t)\right)}{\sum_{t=1}^{N_{frames}} N^t} \quad (12)$$

Where,

$$N^t = \begin{cases} N_G^t, & if\ N_G^t \geq N_D^t \\ N_D^t, & if\ N_G^t < N_D^t \end{cases}$$

For $\sum_{t=1}^{N_{frames}} N^t = 0$ we force N-MODA = 0. The parameters, c_{ms} and c_f are weighting parameters that can be varied according to the specified application (in the current paper, $c_{ms} = c_f = 1$), N_G^t and N_D^t are the number of ground-truth and system detected objects respectively.

In order to analyse the system's tracking accuracy the Multiple Object Tracking Accuracy (MOTA) is computed (according to [31]) as follows,

$$\text{MOTA}= 1 - \frac{\sum_{t=1}^{N_{frames}} \left(c_{ms}(ms_t) + c_f(fp_t) + c_s\left(ID_{SW_t}\right)\right)}{\sum_{t=1}^{N_{frames}} N^t} \quad (13)$$

Where, ID_SW_t is the number of object labels mismatches in the current frame 't' relative to the previous frame$(t-1)$. The parameter c_s is given as, $c_s = \log_{10}$, hence ID_SW_t counts always start from 1. We compute these accuracy measures along with other important ROC values (i.e., average *true positives*, *false positives*, *false negatives*, *total number of objects* per image and the *true positive rate*) for both the proposed techniques.

Performance evaluations for *visual saliency-based* algorithm are shown in Tab. 1. Results indicate a very good performance overall (especially in the case of LAB-based and PANGU datasets) whereas it is within an acceptable range for the challenging SEEKER dataset (specifically observing the ROC ('*receiver operating characteristic'*) values, such as miss rate and false positives per image). It is worth reiterating that this performance is achieved in purely unsupervised manner, without any top-down feedback as well as *a priori* knowledge of the test datasets used. Results for the Shape-based detection and tracking model are shown in Tab. 2. We observe good performance overall, whereas it slightly outperforms the saliency-based system in the case of the ROC values (i.e., miss rate and true positive rate).

7. CONCLUSION AND FUTURE WORK

We proposed two distinct approaches towards object detection and tracking (planetary rocks) with specific focus of application within the domain of long-distance autonomous navigation of planetary rovers. The algorithms proposed in this paper used quasi-thematic features for describing rocks on the basis of their visual saliency and Hu set of invariant moments such that the ROIs were represented by blobs rather than complex features. This greatly reduced the complexity within the feature space, and as a result enabled the use of simplistic heuristic-based tracking techniques. Performance of the two proposed techniques was thoroughly examined using standard quantitative evaluation protocols in terms of their detection and tracking accuracies on human annotated ground-truth datasets. Results thus achieved for both techniques generally showed good performance especially in the case of the saliency-based system. We believe such paradigms that can model objects in terms of semantic descriptions could potentially form a very effective basis for object detection and tracking problems specifically for applications in future long-distance autonomous rover navigation.

With the proof-of-concept evaluations showing promising results for both these distinct methodologies, we anticipate exploring many other novel dimensions of semantic features-based visual object detection and tracking techniques in future. Furthermore we also anticipate to experiment with more challenging noisy datasets in order to achieve a solid foundation for the proposed concept.

8. ACKNOWLEDGEMENT

The work presented here was partially supported by the U.K. Engineering and Physical Sciences Research Council under Grant EP/J011916/1 (Reconfigurable Autonomy). The authors are very grateful to the Autonomous Systems Group, RAL Space, for providing the SEEKER dataset.

9. REFERENCES

1. Howell, L. W., & Ruf, J. H. (1986). *Graphical techniques to assist in pointing and control studies of orbiting spacecraft*, NASA, Scientific and Technical Information Branch.

2. Koekemoer, A. M., et al. (2007). *The cosmos survey: Hubble space telescope advanced camera for surveys observations and data processing*, The Astrophysical Journal Supplement Series 172, 196.

3. Mutch, T., et al. (1976). *Fine particles on mars: observations with the Viking 1 lander cameras,* Science 194 (4260) 87–91.

4. Mutch, T., et al. (1976). *The surface of mars: the view from the Viking 2 lander,* Science 194 (4271) 1277–83.

5. Cabane, C., et al. (2009). *Mars Science Laboratory (MSL) and the future missions to Mars,* Highlights of Astronomy, Volume 15, XXVIIth IAU General Assembly, Ian F. Corbett, ed.

6. Chen, G., Barnes, D., LiLan, P. (2012). *An Approach for Matching Desired Non-feature Points on Mars Rock Targets Based on SIFT*, TAROS, Springer, LNCS, (7429), pages (418-419).

7. Itti, L., Koch, C., Niebur, E., (1998). *A model of saliency-based visual attention for rapid scene analysis*, Pattern Analysis and Machine Intelligence, IEEE Transactions on , 1254 –1259.

8. Matthies, L., et al. (2007). *Computer vision on mars*, International Journal of Computer Vision 75 (1), 67–92.

9. Makhlouta, M., Gao, Y., Shala, K., (2008). *A Vision and Behaviour Based Approach for Short-Range Autonomous Navigation of Planetary Rovers*, In proc: ESA Workshop on ASTRA, Noordwijk, Netherlands.

10. Olsen, C. F., & Abi-Rached, H. (2010). *Wide-baseline stereo vision for terrain mapping*, Mach. Vision Appl. 21 (5), 713–725.

11. Cao, F., & Wang, R., (2010). *Study on stereo matching algorithm for lunar rover based on multi-feature*, In proc: 2010 ICICC, IEEE Computer Society, USA, pp. 209–212.

12. Shapiro, L. G., (1994). *Handbook of pattern recognition and image processing* (vol. 2), Academic Press, Inc., FL, USA, Ch. Relational matching, pp. 475–496.

13. Brown, M. Z., Burschka, D., Hager, G. D., (2003). *Advances in computational stereo*, IEEE Trans. Pattern Anal. Mach. Intell. 25 (8), 993–1008.

14. Snderhauf, N., Konolige, K., Lemaire, T., Lacroix, S., (2005). *Comparison of stereovision odometry approaches*, in: ICRA05 Barcelona, Planetary Rover Workshop.

15. Zhu, J., Yuan, L., Zheng, Y., Ewing, R., (2012). *Stereo visual tracking within structured environments for measuring vehicle speed*, Cir. & Sys. for Vid. Tech., IEEE Trans. on (1471-1484).

16. Johnson, A., et al. (2008). *Robust and efficient stereo feature tracking for visual odometry*, in proc: ICRA, pp. 39–46.

17. Cumani, A., Guiducci, A., (2006). *Visual odometry for robust rover navigation by binocular stereo*, in proc: SSIP'06, pp. 74–79.

18. Saxena, A., Schulte, J., Ng, A. Y., (2007). *Depth estimation using monocular and stereo cues*, in proc: IJCAI'07.

19. Viola, P., Jones, M., (2001). *Rapid object detection using a boosted cascade of simple features*, in proc: CVPR, Vol. 1, 2001, pp. I–511 – I–518.

20. Thompson, D. R., Casta, R., (2007). *A performance comparison of rock detection algorithms for autonomous planetary geology*, in proc: IEEE Aerospace Conference.

21. Castano, R., et al. (2005). *Current results from a rover science data analysis system*, in proc: IEEE Aerospace Conference, 2005.

22. Grotzinger, J., et al. (2012). *Mars science laboratory mission and science investigation*, Space Science Reviews 170, 5–56.

23. Milanese, R., Gil, S., Pun, T., (1995). *Attentive mechanisms for dynamic and static scene analysis*, Optical Engineering 34 (8), 2428–2434.

24. Baluja, S., Pomerleau, D. A., (1997). *Expectation-based selective attention for visual monitoring and control of a robot vehicle*, Robotics and Autonomous Systems, 329–344.

25. Treisman, A. M., Gelade, G., (1980). *A feature-integration theory of attention*, Cognitive Psychology 12 (1) (1980) 97-136.

26. Otsu, N., (1979). *A threshold selection method from gray-level histograms*, IEEE Transactions on Systems, Man and Cybernetics 9 (1) 62–66.

27. Samet, H., (2008). *K-Nearest Neighbor Finding Using MaxNearestDist*, Pattern Analysis and Machine Intelligence, IEEE Transactions on , vol.30, no.2, pp.243,252.

28. Yan, F., Zhang, H., Kube, C. R., (2005). *A multistage adaptive thresholding method*, Pattern Recognition Letters 26 (2005) 1183–1191.

29. Suzuki, S., (1985). *Topological Structural Analysis of Digitized Binary Images by Border Following*, Computer Vision, Graphics and Image Processing, N 32-46.

30. Hu, M. K., (1962). *Visual Pattern Recognition by Moment Invariants*, PROC. IRE vol. 49, p. 1428.

31. Kasturi, R. et al. (2009). *Framework for Performance Evaluation of Face, Text, and Vehicle Detection and Tracking in Video: Data, Metrics, and Protocol*, PAMI, IEEE Trans. on, vol.31, no.2, pp.319,336.