Vision-Based Hazard Detection with End-to-end Spatiotemporal Networks for Planetary Landing

Cheng Huang¹,² and Hongmei Wang¹,²
¹National Key Laboratory of Aerospace Flight Dynamics
²School of Astronautics, Northwestern Polytechnical University, Xi’an 710072, Shaanxi, Province, China

Abstract. Hazard detection is the key technique for autonomous landing during a planetary exploration mission. This paper proposes an end-to-end spatiotemporal network to detect all hazards in the image sequences captured by an optical camera. The spatial stream processes the colour image sequences and the temporal one learns features from optical flow image sequences. The spatial features and temporal features are fused by the proposed metric fusion method, and then the spatiotemporal features become more discriminative through triplet loss. A hazard map can ultimately be obtained in the testing phase after processing on the full-size image sequences. The evaluation results prove the effectiveness of the proposed network, and the testing results show the feasibility of practical application.

1. Introduction
In recent years, with the development of the planetary and comet exploration technology, autonomous and intelligent probing technique is more and more widely used. And it is truly a great challenge to autonomously land on an unknown celestial body. Therefore, avoiding all hazards and selecting a safety landing area during the descent process is the key to the success of the space exploration mission. Fig.1 shows that there are mainly four hazards that may influence the landing safety including crater, discontinuities, rock and slope[1]. For traditional researches, different methods are adopted for different hazards.

For automatic detection of craters, the geometric information is usually utilized. Since the edge of the crater is simply like a circle, circles are used to fit craters edge[2][3] to detect the existence of craters. And texture features [4] are also distinguished information that can be combined with supervised learning classifiers to discriminate craters and non-craters[5]. There is one more good method to model the craters according to the photos taken from different angles[6]. These methods are simple and easy to be implemented, but the detection accuracies are not enough to meet expectations. Nowadays, with the rapid development of machine learning methods, there are many more intelligent architectures for craters detection, such as simple convolutional neural networks for
craters feature learning [[7]-[8]]. Artificial neural networks can also interpret the hazard features and generate a hazard map at different scales[[9]].

There are mainly two methods for the rock detection, including edge contour detection and shadow detection. Edge fragments can be concatenated with image pyramid to detect the rock zones under different scales[[10]]. The shadow detection is mainly based on the shining angle of the sunlight [[11]], and it has the advantages of real-time and high stability.

The studies of slope and discontinuities estimation are relatively few. Early slope estimation techniques for landing of detectors are based on distance information or digital elevation model (DEM) from laser scanning radar[[12]]. But compared with passive vision sensors such as optical cameras, such active vision sensors have the backwards of high cost, heavy weight. So it’s meaningful to estimate slope based on optical vision method. The discontinuities detection is relatively simple, there nearly no research to study on it specifically.

In this paper, only optical camera is considered for the hazards detection during planetary landing phase. And the image sequences captured by the camera and their corresponding optical flow [[13]] results are used for the detection algorithms. The Fig.2 shows the way of image sequence captured during landing phase.

A spatiotemporal network is proposed in this paper to uniformly process all types of hazards. And a metric fusion method is also proposed to fuse the spatial features and temporal features, thus form stable spatiotemporal features embedding. In training phase, the features of hazards are learned without pertinently modelling and the classes are correctly distinguished based on triplet loss. In testing phase, full-size images are fed into the network to get the end-to-end hazards distribution results in which the locations and classes are clearly marked. The evaluation results prove the effectiveness of the proposed network.

2. Spatiotemporal networks architecture
The architecture of proposed spatiotemporal networks is shown in Fig. 3. It’s a triplet of two-stream architecture.

![Figure 3](image)

**Figure 3.** The framework of proposed networks. The upper part is the architecture of training phase and the lower is the testing phase.
Each two-stream branch consists of a spatial convolution network and a temporal convolution network, which takes the colour image sequences and optical flow image sequences as inputs, respectively. The spatial part of the two-stream branch carries information about contents and scenes, and the temporal part conveys the motion and edge information. The features learned by these two parts are fused in the metric space to collaboratively represent the spatial-temporal information. And the triplet loss is utilized to make these features more discriminative.

2.1. Convolutional network

The spatial and temporal convolution networks in this paper are the same layer configurations as in Fig.4. It is basically a simple structure for feature learning, which consists of three convolution layers, two max-pooling layers and ReLU activation function. Supposing the spatial image sequence is represented as $s' = \{s'_i | s'_i \in \mathbb{R}^D \}_{i=1}^{T_1}, i = p,a,n$. $T_1$ is the length of image sequence and D is the dimension of images. $p$ means the positive sample branch, $a$ is the anchor sample branch and $n$ is the negative sample branch. Each spatial stream is a CNN architecture which can be denoted as $f(s)$, and the output can be denoted as $f(s')$.

The only difference between these two streams is that the spatial stream operates on the colour image sequence, but the temporal one processes the optical flow image sequence.

The temporal image sequence is expressed by $o' = \{o'_i | o'_i \in \mathbb{R}^D \}_{i=1}^{T_2}, i = p,a,n$. $T_2$ is the length of optical flow image sequence. Thus the output of each temporal stream can be denoted as $f(o')$.

![Convolutional network layers.](image)

2.2. Metric fusion

Each branch of the triplet net is a two-stream architecture, in which the features learned by spatial stream and temporal stream are supposed to be fused to represent a spatiotemporal information. Our method is to embed the temporal feature $f(o')$ into the metric space of spatial feature. The mapping relations can be expressed by: $H(f(s'), f(o')) : (\mathbb{R}^{D_1}, \mathbb{R}^{D_2}) \mapsto \mathbb{R}^{D_1}$, $i = p,a,n$, where $D_1$ is the dimension of fused feature.

Specifically, we take use of polar coordinates transformation $R(\star)$ to calculate this mapping relationship. Thus, spatial feature $f(s')$ and temporal feature $f(o')$ are all supposed to be transformed into polar coordinates space, whose radius and polar angles can be expressed by:

$$r(f(s')), \theta(f(s')) = R(f(s'))$$

$$r(f(o')), \theta(f(o')) = R(f(o'))$$

Where $r(\star)$ is the radius and $\theta(\star)$ is the polar angle.

The mapping result from temporal feature space to spatial feature space can be formulated as following:

$$f'(o') = R^{-1}\left(\frac{r(f(s'))}{r(f(o'))}, \theta(f(s'))\right)$$

Where $f'(o')$ is the metric in spatial feature space and $R^{-1}(\star)$ is the polar inverse transformation.

So the final feature fusion result of the two-stream branch is calculated as
\[ H' = \frac{1}{2} \left( R(f(s')) + f'(o') \right) \]  (4)

2.3. Triplet loss

Within the triplet of \( \{H^p, H^n, H^a\} \), we usually expect features of the positive sample \( H^p \) is more similar to anchor sample \( H^a \) than the features of the negative sample \( H^n \). So the triplet loss is adopted in this paper to pull similar pairs and push dissimilar pairs away. And the expression is formulated as following [[14]]:

\[ L_{tp} = \frac{1}{N} \sum_{i} \left[ \|H^p_i - H^n_i\|_2^2 - \|H^p_i - H^a_i\|_2^2 + \alpha \right] \]  (5)

Where \([\cdot], = max(\cdot, 0)\), \(\alpha\) is a margin that is enforced between positive and negative pairs, \(\|\cdot\|\) is the metric distance. But the number of possible triplets grows cubically with the number of examples, it’s infeasible to process them all and the training converges slowly. So it’s crucial to select triplets to ensure fast convergence. Thus the strategy we adopt is online triplet selection [[15]] method that can randomly select hard negative for each positive pair within a mini-batch.

3. Implementation details

3.1. Network training

The training procedure takes six image sequences as inputs at a time, since the training framework is a triplet of anchor sample, the positive sample and the negative sample. And the meanwhile, each sample branch contains a spatial ConvNet stream and a temporal ConvNet stream. The spatial ConvNet processes the colour image sequence to learn the contextual features while the temporal one takes the optical colour image sequence as input to capture the motion features. Each image sequence just contains one kind of terrain. The training phase aims to make all features more discriminative and classify different target correctly.

3.2. Network testing

The testing phase aims to obtain a hazard map directly through the end-to-end network. There are many image sequences with several scenes taking as test inputs. Each image contains complex terrains, in order to generate a hazard map, we need to extract different terrains to identify the corresponding class and mark them. The mathematical morphology and connected components labelling are utilized to operate on the optical flow image to obtain candidate patches. Then the corresponding patches in colour image and optical flow image are fed into the spatiotemporal net to get the attribute of each patch. All patches can get their labels after ranking with training results. The label of background is set to -1, the labels of terrains are set to 0, 1, 2, 3, respectively. So there is a hazard map obtained, in which the locations and classes of the terrains are marked.

4. Evaluation and discussions

4.1. Dataset

We test the networks on our own dataset which is called Dynamic Planet Terrains (DPT) dataset. The colour image sequence samples are captured in different light intensity, different angles and different height from System Tool Kit (STK). And the corresponding optical flow image sequences are calculated by dense flow method. Parts of samples are listed in Fig.5 below. The number of sequence (NoS) of each type is 40 in train samples and 20 in validation samples. The number of frames (NoF) in each image sequence varies from 15 to 176, in which the image size is 182x184. The detail information of the DPT dataset is shown in table 1.
## 4.2. Evaluation of the proposed method

The experiments are conducted with PyTorch and OpenCV environment. The spatiotemporal network is trained on the DPT dataset to learn the spatiotemporal features of the six-stream inputs. To evaluate the effectiveness of the proposed network, the classification accuracies of the following three schemes are compared:

I. Single CNN stream  
II. Spatiotemporal Net+triplet loss (all possible triplets)  
III. Spatiotemporal Net+triplet loss (online triplet selection)

### Table 2. The comparison of classification accuracy

| Scheme | crater | discontinuities | rock | slope | Average |
|--------|--------|-----------------|------|-------|---------|
| I      | 0.62   | 0.71            | 0.81 | 0.96  | 0.76    |
| II     | 0.76   | 0.81            | 0.89 | 0.94  | 0.85    |
| III    | 0.93   | 0.94            | 0.95 | 0.98  | 0.95    |

From the comparison results in the table 2, it can be clearly seen that CNN has not enough ability to divide all targets apart only by the extracted features. The triplet loss can pull similar targets and push dissimilar targets away, so the method II and III have the better classification ability. But the selection method of triplet is the key of improving the effectiveness of algorithm. And it shows that online triplet selection within a mini-batch in method III is more efficient than process them all in method II.
Fig. 6 shows the triplet embedding result in training phase. The learned metrics are close to each other within class and far away from different classes. As is shown in Fig.7, if spatial stream and temporal stream obtain correct metrics, the metric fusion result is undoubtedly correct. But if one of the stream gets wrong results, like the temporal result of slope in Fig. 8, it’s embedded into the ranges of rock wrongly. Though the metric result of another spatial stream is also far from the centre of the correct class, our metric fusion method also has the ability to calibrate the fusion result to a correct metric result. It may infer that the spatiotemporal net model has been trained as a good terrain classifier.

In the test phase, the input of the algorithm is the sequence of 1226 \times 734 full-size actual images that contains different terrains which is shown in Fig.9(a). The optical flow images in Fig.9(b) are calculated based on the visual image sequence. And the image patches in Fig.9(c) are cropped from colour and optical flow images and then fed into the spatiotemporal network to get their classes. We mark all hazards in the original image with different colours and draw white rectangular boxes according to the labels result. From the hazard map in Fig.9(d), the spatiotemporal net assigns the correct locations and classes of all terrains in the whole image.
Figure 9. The testing results of the network: (a) Full-size colour image; (b) Optical flow image calculated by dense flow; (c) Parts of the cropped candidate patches; (d) hazard map generated by the network.

5. Conclusion
The proposed vision-based spatiotemporal network is a triplet of two-stream architecture, which needs 6 inputs at once. The spatial steam operates on the colour image sequence for contextual features and the temporal stream utilize optical flow to obtain motion and edge information. The metric fusion method can make a good balance between these features and form a good spatiotemporal feature. The network is finally trained as a good terrain classifier via triplet loss. And the hazard map can be generated in the test phase through predicting the candidate patches in the full-size images. The experiment results show that the spatiotemporal network with online triplet selection has a better classification performance and the generated hazard map based on full-size image obtains the precise locations and correct classes of terrains. Future work is to improve the computational performance on practical on-board equipment. And the hyperspectral imaging is taken into consideration for the multispectral information fusion.

6. References
[1] Huertas, Cheng Y, and Madison R 2006 Passive imaging based multi-cue hazard detection spacecraft safe landing IEEE Aerospace Conference
[2] Chen J, Cui P, Cui H 2011 Automated detection and classification for craters based on geometric matching International Symposium on Photoelectronic Detection and Imaging
[3] Sawabe Y Matsunaga T and Rokugawa S 2006 Automated detection and classification of lunar craters using multiple approaches Advances in Space Research 21-27
[4] Ding W 2010 Automatic detection of craters in planetary images: An embedded framework using feature selection and boosting Proceedings of the 19th ACM international conference on Information and knowledge management 749-758
[5] Di K 2014 A machine learning approach to crater detection from topographic data Advances in Space Research 2419-2429
[6] Zhu S Cui P and Hu H 2012 Hazard detection and avoidance for planetary landing based on Lyapunov control method Intelligent Control and Automation 2822-2826
[7] Emam E 2015 Automatic crater detection using convex grouping and convolutional neural networks International Symposium on Visual Computing 213-224
[8] Campbell T 2017 A deep learning approach for optical autonomous planetary relative terrain navigation 27th AAS/AIAA Space Flight Mechanics Meeting
[9] Lunghi P Ciarambino M and Lavagna M 2015 Vision-based hazard detection with artificial neural networks for autonomous planetary landing 13th ESA/ESTEC Symposium on Advanced Space Technologies in Robotics and Automation 1-8
[10] Thompson D R and Castano R 2007 Performance comparison of rock detection algorithms for autonomous planetary geology Aerospace Conference 1-9
[11] Huertas A Cheng Y and Matthies L H 2008 Automatic hazard detection for landers
[12] Johnson A E Klumpp A R, Collier J B 2002 Lidar-based hazard avoidance for safe landing on Mars Journal of guidance control and dynamics 1091-1099

[13] Hassner T and Liu C 2016 Dense Image Correspondences for Computer Vision Springer International Publishing 3-14

[14] Schroff F Kalenichenko D and Philbin J 2015 Facenet: A unified embedding for face recognition and clustering Proceedings of the IEEE conference on computer vision and pattern recognition 815-823

[15] Hermans Beyer L and Leibe B 2017 In defense of the triplet loss for person re-identification arXiv preprint arXiv:1703.07737

Acknowledgement
This work is jointly supported by the National Natural Science Foundation of China (No.61771400), the Key Research and Development Program of Shaanxi Province (No.2017GY-066) and the Seed Foundation of Innovation and Creation for Graduate Students in Northwestern Polytechnical University (ZZ2018061).