

Unsupervised Learning of Scene Categories on the Lunar Surface

Thorsten Wilhelm¹, Rene Grzeszick², Gernot A. Fink² and Christian Wöhler¹

¹Image Analysis Group, Department of Electrical Engineering, TU Dortmund University, 44227 Dortmund, Germany

²Pattern Recognition in Embedded Systems Group, Department of Computer Science, TU Dortmund University, 44227 Dortmund, Germany

Keywords: Remote Sensing, Scene Detection, CNN, Unsupervised Learning.

Abstract: Learning scene categories is a challenging task due to the high diversity of images. State-of-the-art methods are typically trained in a fully supervised manner, requiring manual labeling effort. In some cases, however, these manual labels are not available. In this work, an example of completely unlabeled scene images, where labels are hardly obtainable, is presented: orbital images of the lunar surface. A novel method that exploits feature representations derived from a CNN trained on a different data source is presented. These features are adapted to the lunar surface in an unsupervised manner, allowing for learning scene categories and detecting regions of interest. The experiments show that meaningful representatives and scene categories can be derived in a fully unsupervised fashion.

1 INTRODUCTION

Unsupervised learning of scene categories is a challenging and worthwhile task. The advantage of unsupervised learning is easily at hand. The necessity of annotations is eliminated, which vastly reduces the human effort necessary to reach a meaningful division of the analyzed image data. In this work it will be shown, that it is possible to achieve a meaningful categorization by entirely relying on unsupervised learning algorithms.

A large amount of publicly available image data, which was previously not analyzed with respect to scene categorization, are satellite images of the lunar surface. These images can be seen as a special form of natural scene images. The Lunar Reconnaissance Orbiter (LRO) is a satellite orbiting the moon since 2009 with the objective to analyze the lunar surface with various scientific instruments, including laser altimeters and cameras. The quest of the LRO is diverse and includes finding possible landing sites, and constructing high resolution maps of the moon. The data are publicly available and hosted by the National Aeronautics and Space Administration (NASA). In detail, we use the data provided by the Wide Angle Camera (WAC) global mosaic described in (Speyerer et al., 2011), which has a spatial resolution of 100 meters per pixel and covers the complete surface of the Moon, to look for repeating scenes on the lunar surface. A possible set of scenes is depicted in Fi-

gure 1 with typical scenes being *plains, mountains, highlands, valleys* and *craters*.

Annotations of the lunar surface are scarce and restricted to the most prominent features like craters with a large diameter, or the large plains of the near-side of the moon called lunar mare. It is therefore hardly possible to use state-of-the-art image classification approaches like Convolutional Neural Networks (CNNs) which are typically trained in a fully supervised manner.

One option in such cases is to employ crowd sourcing in order to obtain scene or object annotations (Patterson and Hays, 2012; Perona, 2010). This requires a large number of human annotators and distributes the annotation effort. In contrast, unsupervised scene learning tries to reveal information without the need for annotations. As a result, this task becomes challenging, especially due to the diversity of scene images typically requiring top-down knowledge. It is usually achieved by either a pure clustering approach with handcrafted features like GIST (Oliva and Torralba, 2001) or HoG (Dalal and Triggs, 2005). Deep Learning approaches apply Deep Belief Networks (DBN) (Lee et al., 2009) in which several Restricted Boltzmann Machines (RBM) are stacked. The network is then trained by Contrastive Divergence (CD), which is comparable to gradient descent (Lee et al., 2009).

Another challenge is that in contrast to SAR images of terrestrial surface, the number of information

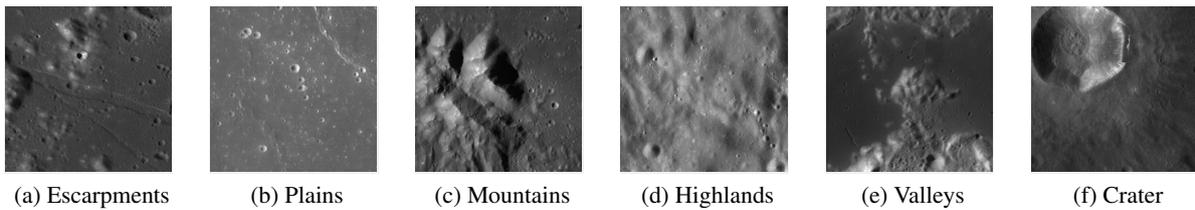


Figure 1: Examples of possible scene categories on the lunar surface extracted from LROC WAC global mosaic (Speyerer et al., 2011). The range varies from valleys to mountains to more Moon specific scenes like the crater heavy plains. Each image depicts an area of roughly 920km^2 of the lunar surface.

cles is limited (Hänsch, 2014) to grey level intensities. Furthermore, due to the low resolution of the WAC images, the number of pixels that is available for context is also limited.

Leveraging the capabilities of CNNs, a novel method which combines the strength of pre-trained networks with unsupervised learning is proposed. The method is able to transfer the feature representation obtained from pre-trained convolution kernels of a CNN to the novel lunar data, which is then used to derive object-like detectors and finally group the object appearances into meaningful scene categories.

2 METHOD

For recognizing different scenes on the lunar surface, it is desirable to have a very descriptive feature representation for a given patch p on the surface. State-of-the-art feature representations can be obtained using CNNs. However, as labeled data is hardly available, learning a feature representation using a CNN is not possible. On the other hand it is also very difficult to design a meaningful handcrafted feature representation. In the following it will be shown how to leverage learned feature representations from a CNN trained on a different data source for unsupervised learning. An overview of an unsupervised detector and scene learning approach are shown in Fig. 2 and Fig. 3, respectively. From an intermediate layer of the CNN a feature representation for yet unknown patches of the lunar surface is derived. Similar to the Bag-of-Features principle these patches are then clustered in an unsupervised manner using spherical k-means clustering. This yields a set of representatives that can be found on the lunar surface, but may as well occur on any image (e.g. simple edges). For the detection, a set of negative samples is used in order to compute the distribution of the representatives on arbitrary images. Choosing a quantile of least occurring representatives, a set of representatives that is most discriminative for the lunar surface is chosen. These representatives are then used as candidates for

detecting different types of lunar surface, e.g. craters or mountains. For the scene learning, the occurrences of the representatives are clustered again on scene level, yielding a set of similar scenes. Furthermore, both scene and detection representatives can be annotated by a human user in order to train a detector or typical scenes with minimal annotation effort.

2.1 Feature Representation

The VGG16 CNN architecture is used as a basis for generating the feature representation (Simonyan and Zisserman, 2014). The convolution part of the network is designed of stacked 3×3 convolution layers and each two convolution layers are followed by a 2×2 max pooling layer. Hence, the context is enlarged by two pixels or multiplied by two in the layers respectively. For example, the context of the first two convolution layers is 3×3 and 5×5 pixels. The next step is a max pooling procedure yielding a context of 10×10 pixels. Hence, after five stages of convolution and pooling layers a context of more than 200 pixels is obtained. For recognition tasks, the convolution part is then followed by fully connected layers. As the resolution of the images is low, the patch size that can be used for context is also limited and information must be obtained within a context of much less pixels. Here, the activations of a learned convolution filter are chosen as the feature representation for the patch m (cf. (Razavian et al., 2014)). The convolution layer can be chosen according to a desired size for the patch m .

A huge advantage of these *off-the-shelf* features is that such a representation can be pre-trained on a separate dataset (i.e., ImageNet (Deng et al., 2009)). Given the assumption that the training set contains a large variability at least a subset may be of interest for the task at hand. In the following, a new set of representatives will be learned based on the feature representation derived from the convolution layer.

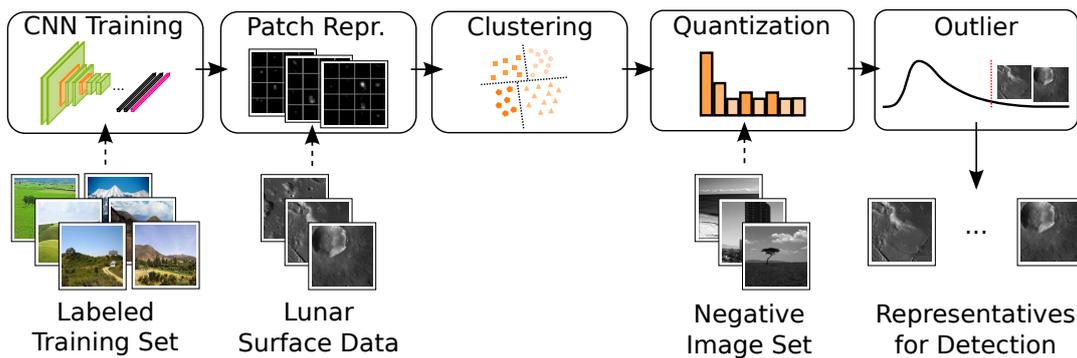


Figure 2: Overview of the proposed detection method. A Convolutional Neural Network is trained on a labeled dataset (i.e., ImageNet (Deng et al., 2009)). From an intermediate layer of the Network a feature representation for yet unknown patches of the lunar surface is computed. Similar to the Bag-of-Features principle these patches are then clustered in an unsupervised manner. Then a set of negative samples is used in order to compute the distribution of these cluster centroids on arbitrary images. Choosing a quantile of least occurring representatives, a set of representatives that is most discriminative for the lunar surface is chosen. These representatives can then be chosen for detection.

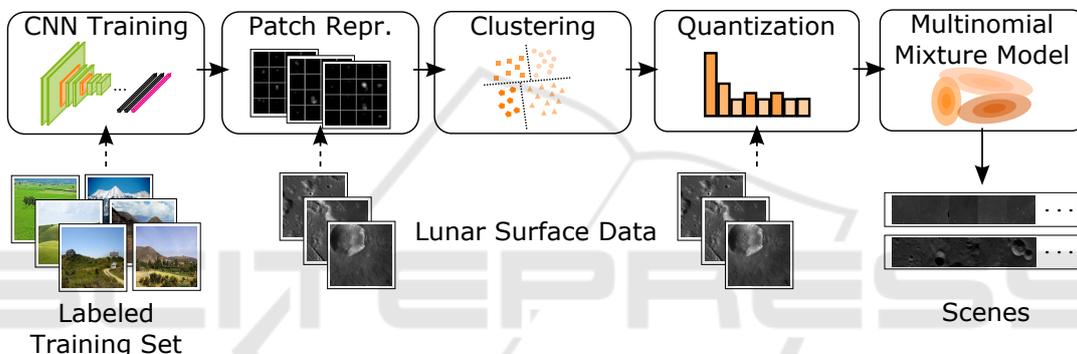


Figure 3: Overview of the scene learning. As for the detection, a CNN is trained on a labeled dataset and features are derived from an intermediate layer. Clustering these features and quantizing the representation at every pixel in an image then given a new representation for an image, as is done for image classification in Bag-of-Features approaches. These are then clustered on scene level using a multinomial mixture model in order to obtain different sets of scenes.

2.2 Unsupervised Learning of Filter Masks

Following the Bag-of-Features principle (Csurka et al., 2004) the activation of the different convolutions are clustered in order to obtain a new representation in an unsupervised manner. A set of patches m is randomly drawn and the activations are derived from the CNN. As the distribution of the activation does not necessarily follow those of an Euclidean space, the cosine distance (Baeza-Yates et al., 1999)

$$d(i, j)_{cos} = 1 - \frac{\mathbf{f}_i \cdot \mathbf{f}_j}{\|\mathbf{f}_i\| \|\mathbf{f}_j\|} \quad (1)$$

is known to work well, where \mathbf{f}_i and \mathbf{f}_j are the feature representations for patch i and j . Spherical k-means clustering is employed to the data (Zhong, 2005), computing a set of centroids \mathbf{c} . The centroids of the clustering process yield a combination of the learned filters from the pre-trained filter masks that describe

the new dataset of patches from the lunar surface well.

The next step is to find those representatives that are especially descriptive for the lunar surface and not just arbitrary images. Hence, a set of negative image \mathbf{I}_n is chosen. For these images a set of negative features \mathbf{n} is computed at each possible location in the image. Hence, a large number of negative patches is computed. All patches are assigned to the set of centroids \mathbf{c} by hard quantization:

$$\operatorname{argmin}_j 1 - \frac{\mathbf{n}_i \cdot \mathbf{c}_j}{\|\mathbf{n}_i\| \|\mathbf{c}_j\|} \quad \forall j \quad (2)$$

This gives a distribution over all centroids. From these a quantile γ is chosen as the most discriminative patch representations.

2.3 Detection

After a set of filter masks is determined, detection is carried out by computing the Pearson product-

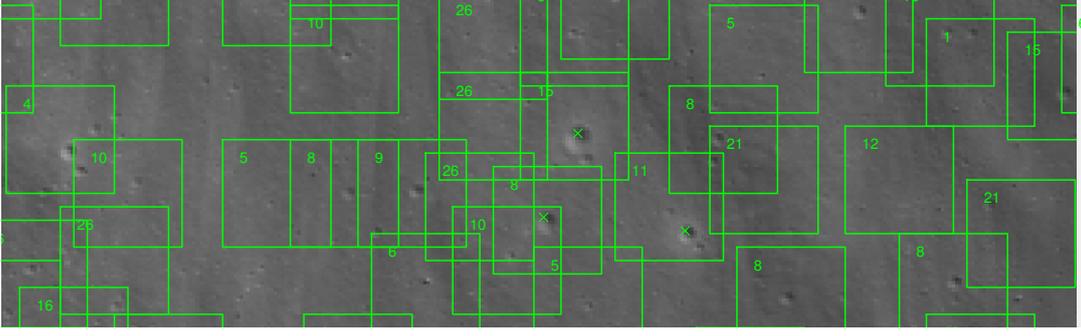


Figure 4: Subset of the detections on the NAC dataset used for crater evaluation. Ground truth crater positions are marked by a green cross. Detections of the crater filter are shown by green bounding boxes. The filter number is noted inside the bounding boxes for comparison. Note how filter eight frequently detects small craters, which are not marked by the ground truth.

moment correlation coefficient r (Fisher, 1915). For two datasets \mathbf{x} and \mathbf{y} with n samples r is computed by

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}. \quad (3)$$

The correlation coefficient offers the advantage of a low computational burden even for large datasets, and is independent of linear transformations, which makes the detection robust against possibly occurring linear transformations.

We assume that a filter detects something in an image if the corresponding correlation coefficient between the image representation and the filters at any given position surpasses a certain threshold θ_r , which may be set globally or individually for every filter.

Non maximum suppression (NMS) (Felzenszwalb et al., 2010) is used to reduce the amount of overlapping detections, which naturally occur around basins of attractions in the image. Due to the unsupervised nature of the training process, similar filter masks may have a high correlation coefficient at similar locations in the image. By using NMS we restrict ourselves to those filter masks with the highest detection confidence in a specific region. The amount of overlap can be adjusted by a threshold θ_{nms} .

2.4 Unsupervised Scene Learning

Based on the detections we try to derive an underlying scene, which can be interpreted as a latent variable. Therefore, histograms of the detections are computed for every image I_i . The feature representation of the images is the same as described in Section 2.1. Based on the histograms of the detections a Multinomial Mixture Model (MMM) with a fixed number of components is used to model different object distributions among different images.

The multinomial distribution has the probability density function (pdf) (Murphy, 2012, p. 35)

$$\mathcal{M}_d(\mathbf{x}; \mathbf{p}) \sim \frac{\Gamma(\sum_{i=1}^N x_i + 1)}{\prod_{i=1}^N \Gamma(x_i + 1)} \prod_{i=1}^N p_i^{x_i}, \quad (4)$$

where d indicates the dimension of the distribution, Γ the Gamma function, \mathbf{p} a vector of probabilities and the observed counts \mathbf{x} . The resulting mixture model is then

$$p(\mathbf{X}|\Theta) = \prod_{i=1}^C \pi_i \mathcal{M}_d(\mathbf{X}; \theta_i), \quad (5)$$

with C as the number of components and the elements $\boldsymbol{\pi}$ are the corresponding mixture weights. The parameters of the whole model are summarized in a set of parameters Θ .

The model will be estimated in a Bayesian fashion, so that the mixture weights and the latent parameters \mathbf{z} indicating the class memberships of every datum are given appropriate probability distributions as well. We assume that the mixture of latent variables \mathbf{z} follows a categorical distribution with

$$p(\mathbf{z}; \boldsymbol{\pi}) = \prod_{i=1}^C \pi_i^{z_i}, \quad (6)$$

and the mixture weights $\boldsymbol{\pi}$ are given by a Dirichlet distribution, because it is the conjugate distribution (Raiffa, 1974) to a categorical distribution. This means that the posterior distribution is given in analytical form and there is no need for computationally demanding posterior inferences for this variable. The dirichlet distribution has the pdf

$$p(\boldsymbol{\pi}; \boldsymbol{\alpha}) = \frac{\Gamma(\sum_{i=1}^C \alpha_i)}{\prod_{i=1}^C \Gamma(\alpha_i)} \prod_{i=1}^C \pi_i^{\alpha_i - 1}, \quad (7)$$

with the hyperparameter $\boldsymbol{\alpha}$ controlling the shape of the distribution, which may be used to model the prior

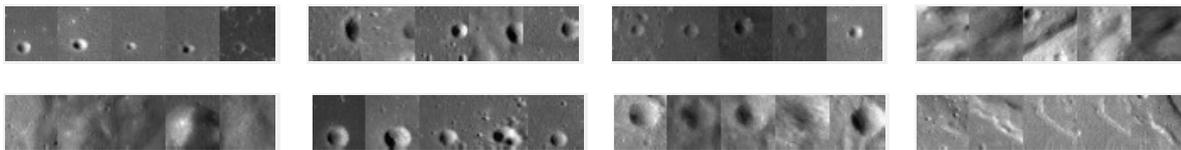


Figure 5: Visualization of the five nearest neighbors according to the cosine similarity for some of the computed centroids. The range changes from craters of varying shape and position to pure plains and hill-like structures.

knowledge available about the mixture weights. Note the analytical similarity to Eq. 6, due to the conjugate nature of the distributions. Further, the probabilities \mathbf{p}_i of every multinomial \mathcal{M}_d are given a Dirichlet prior for the same reason. To estimate Θ we use Markov Chain Monte Carlo (MCMC) with conjugate posterior updates (Congdon, 2014).

3 EVALUATION

The presented approach has been evaluated on a set of images of the lunar surface to derive a meaningful scene representation. Note that only a very rough ground truth describing the properties of the surface is available. For example, the most prominent craters have been annotated or some of the larger mountains. To give some idea of how accurate the derived filter masks are, we evaluated the detections on an entirely different dataset, where annotations for some craters are available. This is described in Section 3.2. In the following we describe some details of the implementation.

3.1 Implementation Details

For the clustering process 400 randomly drawn patches are extracted from each positive image I_p . The patch size is set to 32×32 pixels and the filter response from the conv3.2 layer of the VGG16 net centered at the patch is used for generating the feature representation of the patch.

The set of negative images I_n is generated based on the 15 Scenes dataset (Lazebnik et al., 2006). The dataset has been chosen as it shows arbitrary scenes and greyscale images, which is an important property of the lunar surface images. All images in the dataset have been used as negative samples.

A set of 500 representatives \mathbf{c} has been used in the clustering process. A subset of these is depicted in Fig. 5, where the five image patches with the lowest cosine distance are presented. It can be seen that the variation of typical lunar elements is captured well by the learned centroids.

3.2 Detection

To evaluate the accuracy of the learned filter masks an annotated ground truth will be used. Since the availability of possible annotations is scarce, we restrict ourselves to evaluate the accuracy of the most prominent feature on the lunar surface, craters. In detail, we use annotations provided by (Fisher, 2014) in which craters with diameters varying from 5 to 41 meters are marked. The spatial resolution of the analyzed orbital images amounts to 0.5 meters per pixel, the excerpts are part of Narrow Angle Camera (NAC) (Chin et al., 2007) image M126961088LE. The analyzed region is an area around the crater *Hell Q* which has a diameter of 4 km and is among lunar scientists an interesting field for the study of secondary impact craters, which need to be considered when estimating the age of a surface area based on crater counts (Fisher, 2014).

Note that our approach alleviates the necessity to include the light direction, the surface albedo, or any other information apart from the grayscale image into the detection process. Therefore the detection is more robust with respect to different illumination conditions on the planetary surface and can readily be applied to similar problems on different planetary surfaces. Illumination changes are a frequent issue occurring in conjunction with orbital images, because the data are usually acquired in a bush-broom manner, where different areas of the planet are scanned in different time steps, during which the position towards the sun naturally changes.

Further, the filter masks are trained on the entirely different WAC global mosaic which has a reduced spatial resolution of roughly 100 meters per pixel. However, we assume that craters appear on every possible scale on the surface and yet remain comparable across different scales. Of course, this poses another challenge for the learned filter masks apart from the possible change in illumination with respect to the training data.

Table 1 provides a summary of the detection results. Additionally, the same experiments have been done with patches of the greyscale image data as features. While the true positive rate is encouraging, the false positive rate seems unusually high. The reason

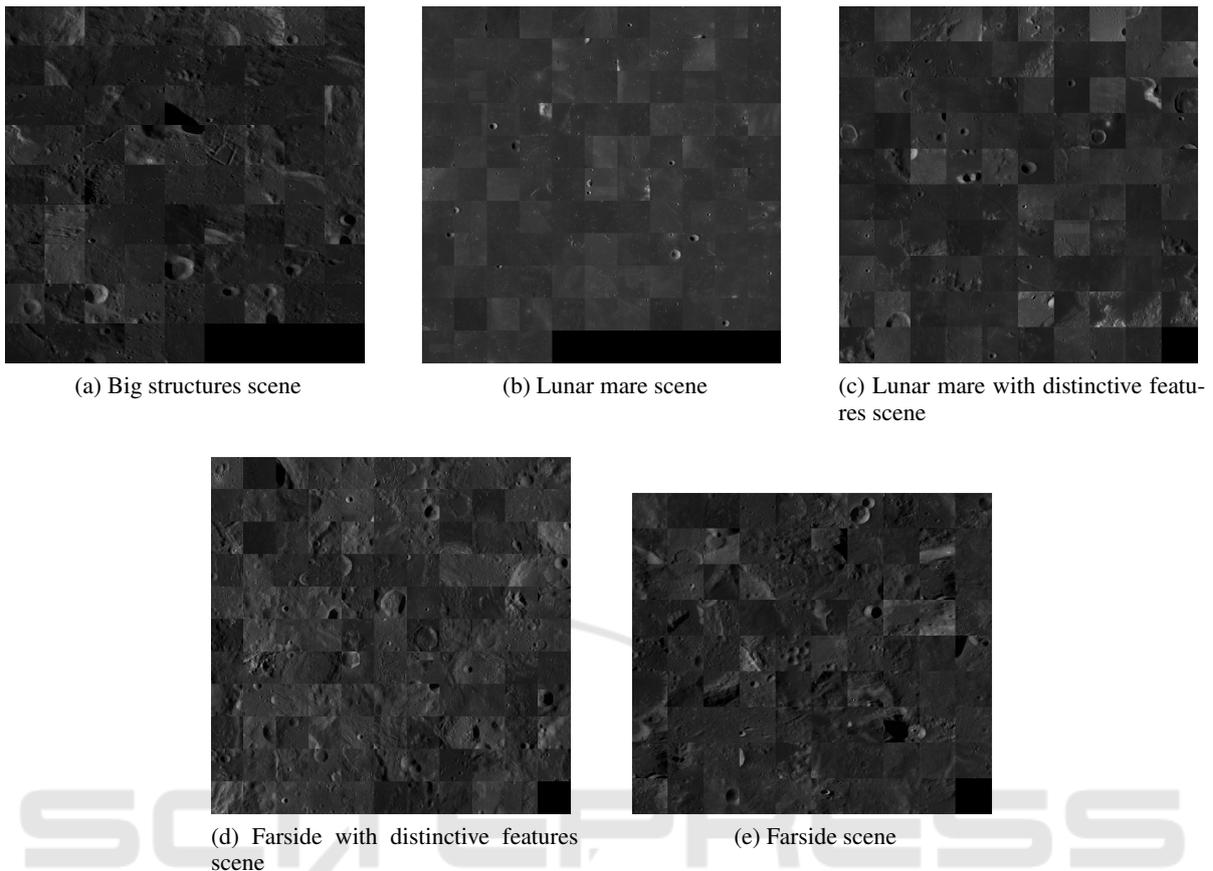


Figure 6: Visualization of the derived categorization based on the object detections assuming that five different scenes are present in the analyzed data. The resulting scene categories meaningfully divide the data in lunar mare and lunar highland scenes. A further distinction between these two scenes is made by differentiating between the presence and absence of distinctive lunar features like large craters, mountains, valleys, or the number of small craters present in the scene.

for this behavior is that our algorithm is not able to distinguish between craters below or above a certain threshold, but detects every possible crater-like structure which is similar to a centroid. In Fig. 4 a subset of the analyzed area is depicted. It is worth noting that all ground truth positions are covered by a bounding box. While this explains the good recognition results, the reason for the high false positive rate is evident as well. A large number of craters, especially the smallest ones, are not comprised by the ground truth. However, our algorithm detects a vast majority of the them as well. In (Fisher, 2014) it is stated that only craters with a diameter ranging from 5 m to 41 m are marked. Obviously, our algorithm detects craters with a diameter less than 5 m.

3.3 Scene Learning

Apart from craters, other characteristic scenes are prominent on the lunar surface. This includes ridges, escarpments, crater chains, mountains, valleys, lunar

mare, or lunar highlands. An overview is depicted in Fig. 1.

Our second goal was to show that it is possible to derive meaningful scenes in an unsupervised manner based on the object detections of the previous section. The mixture model described in Section 2.4 is used to achieve this goal. The idea is that every scene has a unique distribution over objects which are present in the image. A lunar mare scene would for instance have a high probability of exhibiting small craters and plain area patches. In contrast, a scene from the far side of the Moon would have a high probability at features which describe a rough surface and small probabilities on the plain area features.

To stress our uncertainty regarding the mixture weights, we choose to set every entry of the hyperparameter α to one, and do the same for the prior on \mathbf{p} , the feature probabilities of every multinomial. The mixture weights $\boldsymbol{\pi}$, the latent variables \mathbf{z} indicating the class membership, and the \mathbf{p}_i are estimated with MCMC. The resulting scene categorization is shown

Table 1: Evaluation of the crater classifier for various thresholds θ_r used to compute the correlation coefficient between every datum and every centroid c_i . True positive rate (TPR), false positive rate (FPR), false negative rate (FNR), and specificity (SPC) are presented. Besides the results with CNN features, results for the same analysis with greyscale values as features (Px) are presented as a baseline for comparison. Best values are marked in bold. The relatively high false positive rate of the CNN is due to the fact that craters smaller than 5 m are not comprised by the ground truth.

θ_r	TPR		FPR		FNR		SPC		P	N
	CNN	Px	CNN	Px	CNN	Px	CNN	Px		
0.65	0.962	0.202	0.032	0.004	0.038	0.798	0.968	0.996	890	1 143 750
0.70	0.923	0.140	0.028	0.003	0.078	0.860	0.972	0.997	890	1 143 750
0.75	0.836	0.096	0.021	0.002	0.164	0.905	0.979	0.998	890	1 143 750
0.80	0.570	0.054	0.010	0.001	0.430	0.946	0.990	0.999	890	1 143 750
0.85	0.136	0.012	0.005	0.001	0.864	0.981	0.995	0.999	890	1 143 750

in Fig. 6. However, annotations are too scarce to evaluate the accuracy. Therefore we restrict ourselves to a visual inspection.

The found categorization can be divided into two major categories, lunar mare and highland. While the former is most dominant in Fig. 6b, the latter is summarized in Fig. 6e. The MMM further derived a scene which summarizes the boundary between lunar mare and lunar highland and is depicted in Fig. 6c. This fine distinction is worth noting and underlines the success of the presented approach. The remaining scenes describe either large structures, like ridges or parts of bigger craters, or contain scenes where the far side of the Moon is shown with distinctive features.

4 CONCLUSION

A novel approach towards unsupervised scene learning has been described. Based on a pre-trained CNN, state-of-the-art feature representations are adapted to images of the lunar surface. The resulting feature representations have been clustered with spherical k-means in a Bag-of-Features approach to extract object-like detectors capturing frequently occurring patterns in the dataset. The accuracy of a subset of the detectors is evaluated on an annotated dataset of craters on the lunar surface. Based on the learned object detections a scene representation is learned in a Bayesian fashion. The resulting categorization meaningfully divides the analyzed data into typical lunar scenes, like lunar mare, lunar highlands, and the border regions between both.

ACKNOWLEDGMENT

The annotation data of the Hell Q region were provided by Kurt Fisher. This work has been funded by the

Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Projekt number 269661170.

REFERENCES

- Baeza-Yates, R., Ribeiro-Neto, B., et al. (1999). *Modern information retrieval*, volume 463. ACM press New York.
- Chin, G., Brylow, S., Foote, M., Garvin, J., Kasper, J., Keller, J., Litvak, M., Mitrofanov, I., Paige, D., Raney, K., et al. (2007). Lunar reconnaissance orbiter overview: The instrument suite and mission. *Space Science Reviews*, 129(4):391–419.
- Congdon, P. (2014). *Bayesian statistical modelling*, volume 704. John Wiley & Sons.
- Csurka, G., Dance, C., Fan, L., Willamowski, J., and Bray, C. (2004). Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, volume 1, pages 1–2. Prague.
- Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 886–893. IEEE.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE.
- Felzenszwalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645.
- Fisher, K. (2014). An age for hell q from small diameter craters. In *Lunar and Planetary Science Conference*, volume 45, page 1421.
- Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficients in samples from an indefinitely large population. *Biometrika*, 10(4):507–521.
- Hänsch, R. (2014). *Generic object categorization in PolSAR images-and beyond*. PhD thesis, TU Berlin.
- Lazebnik, S., Schmid, C., and Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recog-

- nizing natural scene categories. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 2169–2178.
- Lee, H., Grosse, R., Ranganath, R., and Ng, A. Y. (2009). Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 609–616. ACM.
- Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. The MIT Press.
- Oliva, A. and Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175.
- Patterson, G. and Hays, J. (2012). Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Perona, P. (2010). Vision of a Visipedia. *Proceedings of the IEEE*, 98(8):1526–1534.
- Raiffa, H. (1974). Applied statistical decision theory.
- Razavian, A., Azizpour, H., Sullivan, J., and Carlsson, S. (2014). CNN features off-the-shelf: an astounding baseline for recognition. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 806–813.
- Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.
- Speyerer, E., Robinson, M., Denevi, B., et al. (2011). Lunar reconnaissance orbiter camera global morphological map of the moon. In *Lunar and Planetary Science Conference*, volume 42, page 2387.
- Zhong, S. (2005). Efficient online spherical k-means clustering. In *IEEE International Joint Conference on Neural Networks (IJCNN)*, volume 5, pages 3180–3185. IEEE.