MINING SITES RECOGNITION VIA SCENE CLASSIFICATION USING 3D CONVOLUTIONAL NEURAL NETWORKS

André Estevam Costa Oliveira[1], *Matheus Corrêa Domingos*[1], *Valdivino Alexandre de Santiago Júnior*[1], *Thales Sehr Körting*[1], *Caio Eduardo Dias*[1], *Hércules Carlos dos Santos Pereira*[1] *and Dener César Mendes*[1]

[1]Laboratório de Inteligência ARtificial para Aplicações AeroEspaciais e Ambientais (LIAREA), Instituto Nacional de Pesquisas Espaciais (INPE)
andre.costa@inpe.br, matheus.domingos@inpe.br, valdivino.santiago@inpe.br, thales.korting@inpe.br, caio.dias@inpe.br, hercules.pereira@inpe.br and dener.mendes@inpe.br

ABSTRACT

High spatio-temporal resolution satellite imagery has become increasingly accessible, thanks to advancements in the aerospace industry which, combined with a growing computational power, has enabled the spring of novel techniques regarding recognition in remote sensing (RS) images. That being said, there is still a lack of studies around 3D convolutions for spatio-temporal data applied to classification problems in RS. Hence, this study investigates the application of 3D convolutional neural networks (3DCNNs) for scene classification in RS images, focusing on the identification of mining sites for its high environmental and social impact. We firstly developed a dataset based on MapBiomas products and Planet imagery, then we evaluated the effectiveness of 3DCNNs in capturing temporal information from a sequence of monthly captured images. Comparison between a 3DCNN and a traditional approach, using only a single image with a 2DCNN, shows the advantage of the 3DCNN solution with NIR-G-B images.

Keywords – *3DCNN, scene classification, mining, deep learning.*

1. INTRODUCTION

Mining has played a significant role in the economic development of various regions, being a central activity in the extraction of essential natural resources for various industries. Mining, in general, can be divided into two categories: large scale, carried out by large companies with industrial operations, and small scale, generally carried out by local communities or small groups, with less use of technology and infrastructure.

Large-scale operations involve large investments, which boost the regional economy, but also cause harmful environmental impacts, such as loss of ecosystems, pollution of natural resources and changes in the social dynamics of neighboring communities. These impacts require robust monitoring systems and appropriate public policies to mitigate them.

Small-scale mining, despite lower volume and technology, also generates impacts. Its disorderly expansion can cause environmental damage and is aggravated by informality, making environmental regulation and control difficult [1]. Figure 1 presents example of large-scale and small-scale mining.



Figure 1: Large-scale (left) and small-scale (right) mining.

Faced with these challenges, the creation of effective methods to monitor and evaluate mining activities, regardless of their scale, becomes imperative. In this context, the use of advanced image pattern recognition techniques is promising, as it allows for the accurate detection and monitoring of mined areas, contributing to more efficient environmental management and helping to formulate more assertive public policies.

Over the past few decades, rapid advancements in aerospace technology have led to a significant increase in remote sensing (RS) data acquisition. This data features progressively higher spatial resolution and shorter capture intervals, driving the development of various strategies aimed at maximizing its utility. A notable example is the data cube—a collection of images of the same region taken at different times—commonly used in tasks like land cover classification and change detection [2, 3].

More recently, for tasks such as clustering, classification, and regression, the most popular techniques belong to the field of machine learning. When dealing with temporal data, recurrent neural networks (RNNs), e.g., long short-term memory (LSTM), are more suitable while, for spatial data, convolutional neural networks (CNNs) and derivatives tops most benchmarks. In between the two, there are spatio-temporal approaches, which leads to hybrid methods such as ConvLSTM [4] and 3DCNNs [5].

In a 3DCNN, convolutional layers apply three-dimensional filters to the input data. These filters capture information at different scales and levels of abstraction, allowing the neural network to identify complex patterns and specific features. 3DCNNs have several applications, some of which are action recognition in videos [6], medical diagnosis [7] and object recognition in 3D images [8].

Regarding RS, temporal information is used mainly in land use and land cover (LULC) change analysis. To support such an analysis, data are naturally relevant, such as that provided by the Norway's International Climate and Forests Initiative Satellite Data Program (NICFI) which provides free access to a collection of pre-processed monthly mosaics from Planet's constellation imagery dating back to September

2020. Moreover, Brazil has a well developed project called MapBiomas [9] which maps LULC throughout its entire national territory, providing an important resource for supervised approaches.

Despite numerous approaches already being proposed to assist in LULC mapping by means of spatiotemporal analysis, including machine and deep learning methods [10], there is still a lack of studies demonstrating the feasibility of 3DCNNs in the spatiotemporal context. We believe that such an investigation is highly relevant to the RS community, as it can highlight potential directions for research supported by this technique.

This article explores the potential of 3DCNNs to improve the performance of scene classification task in RS mining images by leveraging space-time information. We firstly developed a dataset based on MapBiomas products and Planet imagery, then we evaluated the effectiveness of 3DCNNs in capturing temporal information from a sequence of (monthly captured) images. We the compared the performance of a 3DCNN and a traditional approach, using only a single image with a 2DCNN and performed a sensitivity analysis with the image bands.

2. MATERIAL AND METHODS

2.1. Study Area

Figure 2 depicts the study area. As we will explain later on, we randomly collected datacube samples all over Brazil's territory, assigning them as "mining"or "other".

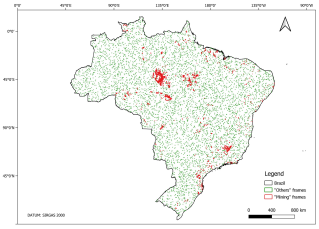


Figure 2: Study area and samples locations.

2.2. Methodology

The methodology of this study is shown in Figure 3. As for training and testing of the models, it was necessary to firstly generate the dataset, as there was no freely available dataset suitable for this study's objective. This phase involved two steps: frames generation and scenes generation. The well-known Brazil Data Cube program offers multi-spectral imagery through time of various satellites, but none of its products matched the pre-requisites for this study, which are: lengthy depth $(t-1)$ and high spatial resolution.

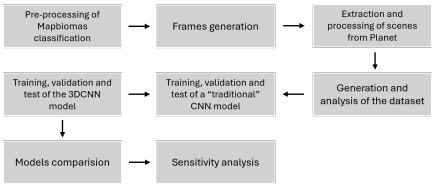


Figure 3: Methodology flowchart.

Having created the dataset, the next step involved developing the algorithms for training and testing the proposed models. The architectures explained in Section 2.4 were chosen so that any difference in performance could be reasonably attributed for the extra layer of information (the temporal dimension).

2.3. Dataset

As for the first step, which we called "frames generation", we used MapBiomas 2022 classification product as reference for mining sites. Its products come in raster format, so first there was a pre-processing step which involved vectorizing the raster and cleaning wrongfully labeled polygons. This cleaning was done using the area of each polygon: if the polygon area was lower than a threshold, we discarded it. The threshold used was 50000 m². Visual post-analysis showed all remaining polygons belonging to mining class.

A tool for QGis was developed to extract frames given an input vector, in this case, the mining sites (see Figure 4). Later, these frames were filtered so that each one should have at least 5% of overlap with a mining site, thus avoiding scenes with barely any mining site in it. The number of total frames generated was 6000 with 2560 meters of height and width. Subsequently, we generated another 6000 frames with 2560 meters of height and width, but this time randomly throughout the region of Brazil and ensuring none overlapped with a mining site. This way, we arranged labeled footprints necessary to clip the Planet imagery into scenes.

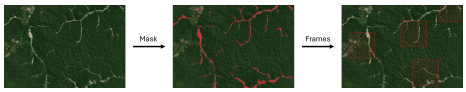


Figure 4: Frames generation process exemplified

As for the next step, i.e, extracting the scenes themselves, we utilized Google Earth Engine (GEE) services, which offers access to various resources such as the Planet imagery, already pre-processed and ready to use, through the "NICFI Satellite Data Program Basemaps for Tropical Forest Monitoring". The NICFI program offers mosaics of the tropical regions with 5m spatial resolution with a monthly temporal resolution since september 2020. Here, we developed a script to run through every mosaic and every frame generating the scenes in an iterative manner.

Ultimately, we have 12000 samples, where each sample is a datacube comprised of 36 scenes (from 01/2021 to 01/2023) of dimension 128x128x4, along with its annotation.

2.4. Models and Metrics

This study focuses on the impact of a temporal layer for the task at hand and the capacity in which a 3DCNN can leverage it. So, as to minimize any difference in performance not related to the models temporal aspect, the proposed architectures for both the 2DCNN and 3DCNN are identical, differing only in their feature extraction components. Figure 5 displays the overall structure of the base model.

The model architecture included 8 layers dedicated to feature extraction, designed to capture and transform raw data



Figure 5: 2DCNN and 3DCNN architecture.

into more abstract and discriminative representations. In the final step, 2 fully connected layers were used, responsible for connecting the extracted features to the output units.

As for 3DCNN, every convolution has a kernel of 3x3x3 with stride of 2x2x2, intercalated by a maxpooling operation with 2x2x2 kernel size and stride. Both models have as last layer a single artificial neuron with a sigmoid activation function, so as to output the probability that the input scene contains a mining site. From one model to another, convolutional layers differ only in its third dimension, i.e., the "traditional"CNN has its third dimension equal to 1, becoming a 2D convolution in practice (2DCNN).

Additionally, we included a (3D) batch normalization step between each convolution and pooling operation with a epsilon of 10^{-5} and momentum of 0.1, so as to avoid the vanishing/exploding gradients problem, given that a downside of 3D Conv is its higher number of parameters to train compared to a 2DCNN. The performance metrics utilized were accuracy, precision and recall, most suitable for binary classification.

3. RESULTS

The experiments described in this study were executed using a set of 500 samples per class for training and 100 samples for test, both randomly selected, totaling 1200 samples. We did not use the 12000 samples generated as there was not enough computing power. The number of epochs was set to 10, and the threshold used was 0,9, i.e., outputs bigger than 0.9 were considered as being of the "*mining*"class and "others"otherwise.

The dataset used in our experiments is characterized by the following configurations as shown in Table 1.

Configuration	**Details**
Image size	512 × 512 pixels
Spatial Resolution	5 meters
Bands	RGB-NIR
N° of samples (train)	1000
N° of samples (test)	200
Time Frames	36 (01/2021 - 01/2023)
Classes	"mining"and "others"

Table 1: Dataset Configuration.

Note that in this first experiment, we are using all 4 bands of the images. It is also important to notice that every performance metric shown in the next items are derived from the test set.

3.1. 2DCNN

As discussed before, specifically for training and testing of the 2D version of the CNN, we utilized a single *t*, 12/2022. The performance metrics achieved are depicted in Table 2. According to Table 2, good results were obtained even using a 2DCNN.

Metrics	Accuracy	Recall	Precision
Value	86%	94%	81%

Table 2: Performance metrics on the test set.

3.2. 3DCNN

As for the 3DCNN model, Table 3 presents the outcomes. From Table 3, we noticed a great result for precision, so we see that the model was able to recognize mining with a good degree of confidence. The recall tells us that the model correctly identified 78% of all actual positive instances. In other words, among all the instances that actually belong to the positive class, the model managed to identify 78% of them. Accuracy says that the model correctly classified 85% of the instances in the dataset.

Metrics	**Accuracy**	**Recall**	**Precision**
Value	85%	78%	90%

Table 3: Performance metrics on the test set.

But we realized that compared to the 2DCNN model, the 3DCNN approach was better only in the Precision metric. We wondered whether there is redundant band information that makes the performance of the 3DCNN to be not as expected. Thus, we acccomplished the sensitity analysis presented in the next section.

4. SENSITIVITY ANALYSIS

We conducted a sensitivity analysis regarding the bands used for training to evaluate their impact on the models' performances. While experiments shown in Sections 3.1 and 3.2 utilized all bands, here we tested using only RGB and NIR-GB. We repeated the training and testing using the same hyperparameters shown in Sections 2 and 3.

Figure 6 shows the differences between a scene depicted using only RGB bands and using NIR-GB. Table 4 presents the results for the 3DCNN while Table 5 shows the outcomes for the 2DCNN model.

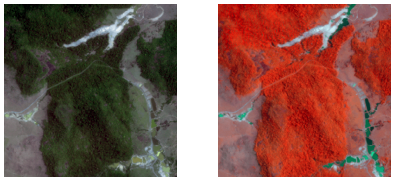


Figure 6: On the left a mining scene using only RGB bands; on the right the same scene but using NIR-GB bands.

Band variation analysis showed that 3DCNN performed better when removing the red band and considering only NIR-GB. The 3DCNN model performed considerably better across all three metrics.

Metrics	C3D (RGB-NIR)	C3D (RGB)	C3D (NIR-GB)
Accuracy	85%	82%	93%
Recall	78%	94%	96%
Precision	90%	75%	90%

Table 4: Sensitivity analysis regarding bands used for the 3DCNN model training. Metrics calculated using the test set.

Metrics	C2D (RGB-NIR)	C2D (RGB)	C2D (NIR-GB)
Accuracy	86%	85%	79%
Recall	94%	84%	80%
Precision	81%	85%	78%

Table 5: Sensitivity analysis regarding bands used for the 2DCNN model training. Metrics calculated using the test set.

5. DISCUSSION

Given the performance metrics shown in Tables 4 and 5, it can be said that the 2DCNN model wielded better results when using the red band, and especially when using all bands (RGB-NIR), with slightly better accuracy but significantly higher recall, comparing to its 3DCNN counterpart.

Unlike the two previous cases, when using the NIR-GB bands, the 3DCNN model was much superior to the 2DCNN one. This demonstrates that correct band selection (in this case, removing the red band and considering only NIR-GB), combined with an approach such as 3DCNN, can result in significantly superior performance compared to a solution using a 2DCNN.

Nonetheless, a 3DCNN architecture requires some considerations. The third dimension (time) cannot be too small: as the data decreased in dimensionality, they became so low that it was infeasible to continue the process. The optimum length of t using the proposed architecture was found to be 36. This aspect can be mitigated using a few strategies, such as lowering the third dimension stride from the kernels and padding the cube before any convolution operation.

6. CONCLUSIONS

As the quantity and quality of RS images increase, spatio-temporal data are becoming more accessible. While several methods leverage both dimensions, none, to the best of our knowledge, explore 3D convolutions for mining site recognition. This study tests the effectiveness of 3DCNNs in utilizing both spatial and temporal information for scene classification.

A completely new dataset was generated to test a 3DCNN model against a "traditional" 2DCNN. Preliminary results shows that the 3DCNN model can perform as good or better than traditional approaches, showing an increase across all performance metrics. Nonetheless, 10% of the whole dataset generated was used, due to technical limitations, which leaves open the question of how the performance metrics would change if given more data. The sensitivity analysis showed how different each model behaves given a different set of bands. Further research is needed to better understand this behavior.

Future directions include an analysis of how 3DCNN models behave with land cover classes that usually get a high confusion with mining, such as urban areas. A comparison to other hybrid architectures, such as ConvLSTM, can bring a better insight into 3DCNN effectiveness.

ACKNOWLEDGMENTS

This research was developed within the project *Classificação de imagens e dados via redes neurais profundas para múltiplos domínios* (**IDeepS**: https://github.com/vsantjr/IDeepS) which is supported by LNCC via resources of the SDumont supercomputer. This reseach was also supported by CAPES and CNPq.

7. REFERENCES

[1] Gavin Hilson. Small-scale mining and its socio-economic impact in developing countries. In *Natural resources forum*, volume 26, pages 3–13. Wiley Online Library, 2002.

[2] Sheng Ji, Li Zhang, and Yujie Li. A review of remote sensing data cube and its applications in land cover classification and change detection. *Remote Sensing*, 10(12):1906, 2018.

[3] Charles Pelletier, Daniel Marceau, and An Nguyen. Data cube: An innovative tool for monitoring land cover and change detection. *International Journal of Applied Earth Observation and Geoinformation*, 81:73–85, 2019.

[4] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W. P. J. S., M. T., J. Y., T. M., and Z. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in Neural Information Processing Systems*, pages 802–810, 2015.

[5] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4489–4497, 2015.

[6] Tamires Martins Rezende. Reconhecimento automático de sinais da libras: desenvolvimento da base de dados minds-libras e modelos de redes convolucionais. 2021.

[7] Anthony EA Jatobá, Lucas L Lima, and Marcelo C Oliveira. Pulmonary nodule classification with 3d convolutional neural networks. In *Anais do XV Workshop de Visão Computacional*, pages 67–72. SBC, 2019.

[8] Felipe Pierre Conter. Detecção de objetos e estimativa de posição 3d através da aplicação de redes neurais convolucionais. 2022.

[9] MapBiomas Project Team. Mapbiomas: Brazilian land cover and use mapping project, 2020.

[10] Kabir Gidado, Mohd Kamarudin, Nik Firdaus, Aliyu Nalado, Ahmad Saudi, Muhammad Saad, and Salisu Ibrahim. Analysis of spatiotemporal land use and land cover changes using remote sensing and gis: A review. *Article in International Journal of Engineering and Technology*, 7:159–162, 2018.