Flood Severity Estimation from Online News Images and Multi-Temporal Satellite Images using Deep Neural Networks

Benjamin Bischke[1,2], Simon Brugman[3], Patrick Helber[1,2]
[1]German Research Center for Artificial Intelligence (DFKI), Germany [2]TU Kaiserslautern, Germany
[3]Radboud University, Netherlands
{benjamin.bischke,patrick.helber}@dfki.de
{simon.brugman}@cs.ru.nl

ABSTRACT

This paper provides a description of our approaches for flood severity estimation in our participation at the Multimedia Satellite Task at MediaEval 2019. We use state-of-the-art deep neural networks for image classification, object detection and human pose estimation in order to estimate the water level from online news images. On the multi-temporal city-centered satellite sequences, we show that derived water indices which are often used for flood detection can be learned with neural networks. By relying on recurrent networks, we want to move forward the state-of-the-art in flood impact assessment by motivating for models that are well known in computer vision but generally not often used by remote sensing researchers.

1 INTRODUCTION

Many approaches in emergency response for flooding events are based on satellite imagery and focus on flood extend mapping. In this work, we study the enrichment of satellite imagery with complementary information from online news by focusing on flood severity estimation. We first consider the task of water level estimation during flooding events. Such information is particularly important for emergency response, but at the same time difficult to extract from satellite imagery. Reasons for this, are the need of a high resolution elevation model and a fast access to satellite imagery. The latter aspect is often difficult to establish since images from optical sensors can often not be used due to the presence of clouds and adverse constellations of non-geostationary satellites at particular points in time. Additionally, we study flood severity estimation by exploiting multi-temporal satellite images that are increasingly available nowadays. While many approaches in the past are based on indices and pre-defined thresholds that work well for particular regions of the world, we look at new methods from deep learning that are able to detect changes in multi-temporal images that are attributed due to flooding events. Our approaches build upon the dataset that was released by the Multimedia Satellite Task 2019 [1].

2 APPROACH

2.1 Multimodal Flood Level Estimation

For the estimation of water level from online news images, we use a multi-stage approach. In the first step, we use a convolutional neural network (CNN) to classify images with respect to the two classes of flood-related and non-flood related images. Therefore, we use a ResNet18 [6] pre-trained on ImageNet [3] which we finetuned on the images of the MFLE dataset. In the following step, we only consider those images that were classified as flood-related. We use a object detector to identify persons and a classifier to determine if the water level is above or below the knee of the detected person. As object detector, we use Faster R-CNN [7] with a ResNet101 [6] as backbone which was pre-trained on the Pascal VOC 2007 dataset [4]. We employ the model on the filtered images and crop patches of persons. For each extracted patch, we compute a feature vector that reflects the body pose of the depicted person. The motivation of using the body pose as a feature vector for estimating the water level is the following: If the knees or lower body parts are occluded by water, this is also reflected in the feature vector with no predicted coordinates for these body joints or with a very low confidence only. We use Openpose [2] for pose estimation and compute as feature vector the normalized coordinates of the predicted body joints as well as the corresponding confidence scores of the model. In the case that the image crop depicts more than one person in the crop, we select the one which is most centered. To finally classify the crops into water level above or beyond water level, we trained a Support Vector Machine classifier with radial basis function as kernel. If there is at least one crop that was classified as above knee level, we assign this label also to original image, otherwise we continue with the next person patch.

Evaluations on our internal validation dataset revealed that the approach of classifying the body pose as a proxy to estimate the water level estimation leads to high recall but low precision. This is because the lower legs are often occluded by other objects (e.g. other persons, cars, boats) that are not water. In order to reduce the number of False Positives, we extract the lower part of the person crop and classified this region into the classes in the two classes water and non water. This water detector is based on a ResNet18 [6] model which we fine-tuned on small patches of water and non water occluded persons.

2.2 City-centered satellite sequences

The satellite images of the CCSS dataset are already pre-processed and atmospherically corrected. However, so that the images can be processed with standard deep learning frameworks, we multiply the pixel values in all bands by a factor of $1e-4$ to map the values from 16 bit to a floating number and normalize each band with mean and standard deviation. We supressed incomplete images in a sequence with the tag $FULL-DATA-COVERAGE$ equals to false.

For classifying the change in the images that is attributed to flooding events, we use a sequence classification approach with LSTM models. Since we are dealing with images, we employ a

MediaEval'19, 27-29 October 2019, Sophia Antipolis, France

Table 1: We report the F1-Scores of the MFLE task for the testset and our internal validation set. We can see that, the water patch classifier (run 2) has only a low impact (run 1).

	Run 1	Run 2	Random	Dev. Dist.
Dev. set	73.09%	74.38%	59.51%	51.42%
Test set	74.27%	74.82%	59.17%	51.80%

Table 2: We report the F1-Scores of the CCSS task for the testset and our internal validation set. We can see that, the Conv-LSTM performs marginally better than our baselines.

	Run 1	Run 2	Run 3	Random	Dev. Dist.
Dev. set	93.82%	91.35%	92.59%	51.85%	58.88%
Test set	92.10%	93.50%	96.29%	49.32%	56.10%

Convolutional LSTM (ConvLSTM) [9] to learn the temporal dependencies between the images. The ConvLSTM uses 32 hidden units and is trained on sequences of variable lengths. We use the pre-trained network ResNet18 as encoder for extracting from raw images the feature maps before the average pooling layer and pass these feature maps to the ConvLSTM. Since ResNet18 was trained on images with only three channels, we only pass the RGB bands of the Sentinel-2 satellite imagery to the network.

In the second step, we also experiment with adding two convolutional layers before the input of the ResNet18 that compress the 12 input channels to 3 channels using 2D convolutions. Therefore, we upsample all 20 and 60 meter bands of the Sentinel 2 images of the dataset to resolution of the 10 meter bands via bilinear interpolation and perform a channel-wise stacking.

Our third approach builds on the observation that the remote sensing community has been using indexes [5], such as the Normalized difference water index (NDWI), while other researchers use Convolutional Neural Networks (CNNs) for these tasks [8]. Using indexes has the benefit that the transformed bands can be visualised and interpreted by humans, at the cost of having been selected and optimized by experts for the task at hand. The CNNs do not offer this approach, however can be trained with labelled data. We unify both approaches and propose a network where the indices are represented as layers in the CNN. The architecture consists of two convolutional layers with 1x1 convolutional kernels, and a *log*- and *exp* function as activation function after these layers respectively. In this architecture, there is an analytical solution for finding the weights that correspond to popular indexes as NDWI, NDVI, ARVI, NDRE. We use two of these layers as well as the activation functions as an alternative for the second approach to convert the 12 input channels to 3 channels.

3 RESULTS AND ANALYSIS

The development sets for all subtasks are split into an internal train and validation set with a 70/30 ratio. We make source code for both subtasks available under this link[1].

3.1 Multimodal Flood Level Estimation

For the MFLE subtask, we submitted the following runs: (1) Classification using the multi-stage pipeline, (2) Same as (1) but without the final water classifer, (3) using random guessing with the distribution of the development set. We can see in Table 1, that multi-stage approach performs marginally better than random guessing. We can also see that the water level classifier adds only a minor contribution to pose-based water level classifier. Since we are using a multi-stage system and we want to quantify the influence of the different classifiers and perform an ablation study on the internal validation set. Our results show the following insights: When considering those images that have the label 'above knee level' as relevant class, we can see that the flood classifier results in a high recall and high precision. Similarly for the person detection with Faster R-CNN [7], we obtain a high recall and high precision. For the classification with Openpose however, the recall is high but the precision is low. There are multiple reasons for this. We observed that (1) the pose estimation fails in certain conditions e.g. for women in a skirt and (2) noticed errors of the prediction from Openpose due to reflections on the water surface. By looking at the failure cases, we additionally noticed that is important to filter non-standing persons that are detected by Faster R-CNN [7], e.g. persons that are not standing or partially visible persons close to the image border.

3.2 City-centered satellite sequences

For the CCSS subtask, we submitted the following runs: (1) using a ConvLSTM with RGB bands as input, (2) using a ConvLSTM with all 12 bands as input and two 1x1 convolutions that reduce 12 to 3 bands, (3) same as in (2) but with *log* and *exp* as activation functions. As baselines we compare the approaches against (4) random guessing and (5) random guessing with the distribution of the development set. In Table 2, we report the scores for all five runs. We can see that all runs based on the ConvLSTM yield high scores for both sets. Additionally, we can see that the score for RGB (run 1) is slightly better on the dev. set than the other runs while all bands (run 2) and all bands reduced with the internally learned indices (run 3) resulted in the highest scores on the testset. Since the scores for the first three runs are all very high, we will extend this work in the future with an additional testset.

4 CONCLUSION & FUTURE WORK

Summarizing this work, we presented for the MFLE task an approach based on state-of-the-art computer vision models for water level estimation from online images. In this approach we employed the model Openpose [2] and showed how existing approaches can be used to support disaster response. Nevertheless, we also identified limitations and future directions to consider (reflections, skirts, persons on image borders). For the second subtask we showed that ConvLSTMs are a powerful model to detect changes of a particular class in multi-temporal satellite imagery. Additionally we explored the possibility to represent traditional remote sensing indices directly with neural networks. We will follow up this idea in the future, as such models can be very helpful to combine insights of Remote Sensing (indices) with recent advances of Deep Learning.

ACKNOWLEDGMENTS

This work was supported BMBF project DeFuseNN (01IW17002) and the NVIDIA AI Lab (NVAIL) program.

[1] https://github.com/bbischke/MMSat19Submission

REFERENCES

[1] Benjamin Bischke, Patrick Helber, Simon Brugman, Erkan Basar, Zhengyu Zhao, Martha Larson, and Konstantin Pogorelov. 2019. The Multimedia Satellite Task at MediaEval 2019. In *Working Notes Proceedings of the MediaEval 2019. MediaEval Benchmark (MediaEval-2019), October 27-29*. Sophia Antipolis, France.

[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *CVPR*.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[4] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision* 88, 2 (2010), 303–338.

[5] Bo-Cai Gao. 1996. NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. *Remote sensing of environment* 58, 3 (1996), 257–266.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[7] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*. 91–99.

[8] Tim GJ Rudner, Marc Rußwurm, Jakub Fil, Ramona Pelich, Benjamin Bischke, Veronika Kopačková, and Piotr Biliński. 2019. Multi3Net: Segmenting Flooded Buildings via Fusion of Multiresolution, Multisensor, and Multitemporal Satellite Imagery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 702–709.

[9] SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*. 802–810.