Flood Event Analysis base on Pose Estimation and Water-related Scene Recognition

Khanh-An C.Quan[1], Tan-Cong Nguyen [2], Vinh-Tiep Nguyen[1], Minh-Triet Tran[3]
[1]University of Information Technology, VNU-HCM
[2]University of Social Sciences and Humanities, VNU-HCM
[3]University of Science, VNU-HCM
15520006@gm.uit.edu.vn,ntcong@hcmussh.edu.vn,tiepnv@uit.edu.vn,tmtriet@fit.hcmus.edu.vn

ABSTRACT

In this paper, we describe our approach for the Multimedia Satellite Task: Emergency Response for Flooding Events at the MediaEval 2019 Challenge. Specifically, for the Multimodal Flood Level Estimation subtask, we employ a combination of ResNet-50 trained on Places365 dataset for features extractor, OpenPose for pose estimation and Mask R-CNN for segmentation to predict if an image has at least one person standing in water above the knee. Our approach has achieved the highest results for Multimodal Flood Level Estimation subtask.

1 INTRODUCTION

In this Multimedia Satellite Task, we take part in two subtasks: Image-based News Topic Disambiguation (INTD) and Multimodal Flood Level Estimation (MFLE). We propose using EfficientNet features [8] for training a water-related image classifier in the first subtask. For the second task, we use both EfficientNet and ResNet-50 features. Then, we employ Faster R-CNN[7] to detect if there are people in the image. We also combine binary mask from Mask R-CNN [3] and pose from OpenPose [2] to predict whether the image contains at least one person standing in water above the knee. We also implement a language model for article's content and title contains the image. To evaluate our method, we use F1 score. Full details of the challenge tasks can be found in [1].

2 APPROACH

2.1 INTD subtask

Firstly, the input image will be segmented to get the background. After that, we use EfficientNet architecture to extract image features of both the original image and background image. We use the extracted features on multiple convolution layer and concatenate them together. By using the original image and the background image we have two extracted image features of the same size. Finally, we concatenate these two features together and feed into fully-connected layers to estimate the final result.

2.2 MFLE subtask

For the second task, our proposed method contains four stages: water-related scene recognition, person detection, pose estimation and prediction based on paired mask and pose. Our system pipeline for this subtask shown in the Figure 1.

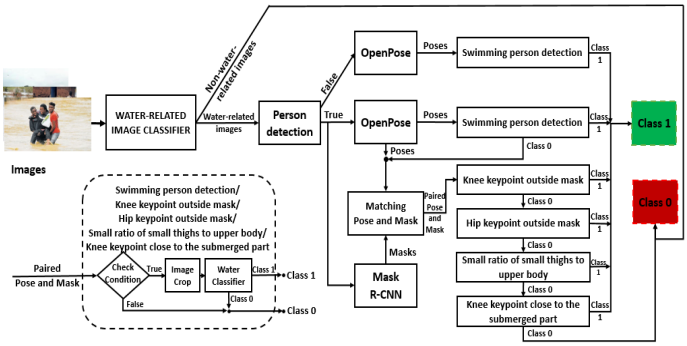

Figure 1: Overview of our MFLE pipeline

For the first stages, we label all the images of the training set into two categories: water-related and non-water-related scene. Then, we use the result of the average pooling layer from ResNet-50 trained on Places365[9] as visual features to combine with a neural network to classify whether the image scene related to water or not. We also employ the visual model from the first task (described in detail in section 2.1) on the non-water-related images to ensure that the water-related images are not omitted. All water-related images will be carried over to the next stage and the remaining images will be labeled as class 0.

For the second stage, we use Faster R-CNN to eliminate images that do not contain a person inside on water-related images. Both positive and negative images will be estimated pose by OpenPose to detect the swimming person in the next stage.

In the third stage, we use the OpenPose with COCO output format [5] to estimate the poses of all people (figure 2). We also calculate the pose bounding box based on the output keypoints. After that, we train a WaterClassifier network to predict the label of water/non-water. We crop ⅓ area from the bottom of all images in the training set vertically and divide manually into water/non-water image. Then we extract visual features using ResNet-50 trained on Places365 and train a neural network to predict the label of water/non-water.

For the last stage, we will make a prediction based on the paired mask and pose. Firstly, to detect all the swimming persons that contain in the image we extract poses with shoulders upwards only (including arms). In most swimming case, OpenPose gives very well result. After extract poses, we crop 50 x 100 pixels areas below the pose bounding box that calculated from the previous step then feed into WaterClassifier to predict whether a person is swimming or not. According to the observation, we realized that

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

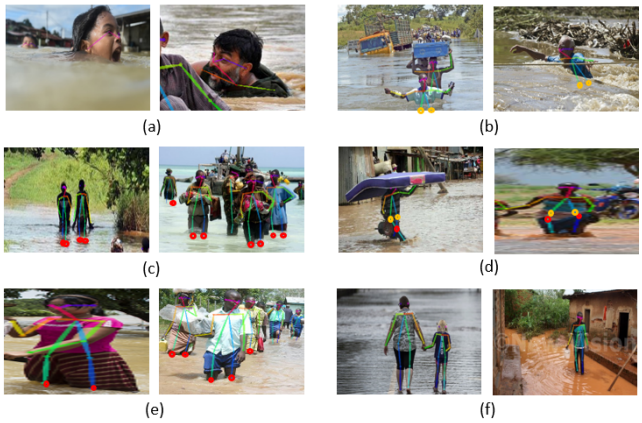

Figure 2: All cases of the result predicted by OpenPose (Red/yellow circle illustrates the knee/hip keypoint meet the conditions of each case). (a) Swimming person with pose from the neck upward, (b) Hip keypoint outside of the person, (c) Knee keypoint outside of the person, (d) Keypoint fit with the person but the ratio of thighs is very small compared to the upper body, (e) Knee keypoint close to the submerged part, (f) Normal pose of the person not submerged

in some cases swimming persons that only have head and upward cannot be detected or misclassified by Faster R-CNN. Therefore, we also applied this to the negative result from the person detection stage to make sure we do not miss any swimming person.

We also use Mask R-CNN to get binary mask and bounding box (bbox) of each person in the images. Then, we conducted a pairing between pose and binary mask of each person in all the water-related images that have at least one person. We calculate the IoU score of bounding box mask and bbox pose of all pose and binary mask pairs included in the image. After that, we match pose and mask with IoU score from high to low with each pose having only one mask and vice versa. We also eliminate cases where the person is on a vehicle or boat removing the paired mask and pose from the image. After matching pose and mask of each person in each image, we conduct resolve special flooded cases: Knee keypoint outside of the person (Figure 2.c), Hip keypoint outside of the person (Figure 2.b), Keypoint fit with the person but the ratio of thighs is very small compared to the upper body (Figure 2.d), Knee keypoint close to the submerged part (Figure 2.e) by crop a rectangular area suitable rectangle for each case. All the rectangular areas cropped from the above cases will be extract features using ResNet-50 trained on Place365 as the input of the WaterClassifier to predict whether the person's knee above the water or not. All remaining images not classified with water or that do not meet the above cases are classified in class 0.

For this subtask, we also implement language model base on the article's content and title. We employ both LSTM and CNN to extract features of preprocessed text. Then, we use GloVe [6] to represent each word by a 300-dim vector. In the first module, we use Bidirectional LSTM [10] with 2 layers with 512 nodes of each. In the second module, we use CNN [4] 3 layers with increasing kernel size of 3,4,5. After the title and content of the article are put into these two modules, we summarize their output on the output layer and feed in the full connected layers to classify.

Runs	F1-Score
Run 1	**0.8850**
Run 2	0.8603
Run 3	0.8757

(a)

Runs	F1-Score
Run 1	**0.8831**
Run 2	0.5341
Run 3	0.7484
Run 4	0.8746
Run 5	0.7419

(b)

Figure 3: Results of (a) INTD subtask and (b) MFLE subtask.

3 RESULTS AND ANALYSIS

3.1 Submitted runs

For the INTD subtask, we have submitted 3 runs, as below:

- **Run 1:** Randomly split the train set with 9:1 ratio into train and val set. After training, we also run more some epochs on the entire training and validation set before predicting on the test set.
- **Run 2:** Same model as **Run 1** with additional photos in the training set of MFLE subtask.
- **Run 3:** Same model as **Run 2** but adjusted some threshold.

For the MFLE subtask, we have submitted 5 runs, as below:

- **Run 1:** Model described at Section 2.2.
- **Run 2:** Text model described at the end of the Section 2.2.
- **Run 3:** Combine the results **Run 1** and **2** with class 1 only.
- **Run 4, Run 5:** Same as **Run 1** and **Run 3** but adjusted some threshold of visual model.

3.2 Results and Analysis

Figure 3.a presents results of three runs for the INTD subtask. In the first Run, the model has 0.887 score. But in run 2 and Run 3, by using extra dataset from Subtask 2, the performance is reduced, may be due to distribution of two datasets are different.

As shown in the Figure 3.b, Run 1 obtains the best result for the MFLE subtask. The text features at Run 2 does seem to achieve average results. The results of Run 4 and Run 5 are only adjusted at some threshold so there is no big difference compared with Run 1 and Run 3.

4 CONCLUSION AND OUTLOOK

In this paper, we employ a combination of ResNet-50 trained on Places365 dataset and EfficientNet for features extractor, OpenPose for pose estimation and Mask R-CNN for segmentation to predict an image has at least one person standing in water above the knee. Our methods show potential results and achieve the highest rank at the MFLE subtask at the challenge. For the future works, we think we can improve both water-related image classifier and water classifier to increase accuracy.

ACKNOWLEDGMENTS

Research is supported by Vingroup Innovation Foundation (VINIF) in project code VINIF.2019.DA19. We would like to thank AIOZ Pte Ltd for supporting our team with computing infrastructure.

REFERENCES

[1] Benjamin Bischke, Patrick Helber, Erkan Basar, Simon Brugman, Zhengyu Zhao, and Konstantin Pogorelov. The Multimedia Satellite Task at MediaEval 2019: Flood Severity Estimation. In *Proc. of the MediaEval 2019 Workshop* (Oct. 27-29, 2019). Sophia Antipolis, France.

[2] Zhe Cao, Gines Hidalgo, Tomá imon, Shih-En Wei, and Yaser Sheikh. 2016. Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), 1302–1310.

[3] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. 2017. Mask R-CNN. *2017 IEEE International Conference on Computer Vision (ICCV)* (2017), 2980–2988.

[4] Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882* (2014).

[5] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *ECCV*.

[6] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[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 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (Eds.). Curran Associates, Inc., 91–99.

[8] Mingxing Tan and Quoc V Le. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *arXiv preprint arXiv:1905.11946* (2019).

[9] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2017).

[10] Peng Zhou, Zhenyu Qi, Suncong Zheng, Jiaming Xu, Hongyun Bao, and Bo Xu. 2016. Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling. *arXiv preprint arXiv:1611.06639* (2016).