Temporal 3D Fully Convolutional Network for Water-hazard Detection

Juntao Li
College of Engineering and Computer Science
Australian National University
Canberra, Australia
u6342214@anu.edu.au (now juli2122@uni.sydney.edu.au)

Chuong Nguyen*, Shaodi You
Data61 - Cyber Physical Systems
CSIRO
Canberra, Australia
{chuong.nguyen, shaodi.you}@csiro.au

***Abstract*—Water on road is a potential hazard for road vehicles. In complex traffic, it is difficult for drivers to safely share the road with other vehicles and pedestrians while avoiding water puddles and pot holes. Running into these could potentially damage electronic components and corrode metal parts of the car, or even lead to loss of control. Similarly, driver-less cars also need to detect water puddles and plan safe path around them or slow down if it deems to be safe to do so. Such detection needs to be both accurate and fast enough for realtime assistance or planning. We present a new temporal 3D fully convolutional network called T3D-FCN for water detection that exploit temporal information to achieve accuracy comparable to the-state-of-the-art accuracy while requires less computation resources. We also show that by adding temporal information the detection performance significantly improves with a small additional computation as compared to single image detection using a similar network architecture.**

***Index Terms*—Water puddle detection, Road hazard detection, Driverless cars, Deep learning, 3D CNN**

I. Introduction

Water on road can pose serious problems to vehicles and people sharing roads. Running over water puddles can cause damages to electronics and mechanical components of the vehicle, or loss of control and potential serious accidents. Planning ahead to avoid water puddles requires accurate detection of water and occupying area.

Water detection for cars has been studied extensively in literature. A wide range of detection methods include radars [19], infrared sensor [7], thermal sensor [2], colour [17], texture [23], stereo cameras [9], and polarised cameras [15], [21] have been used in the past. Recently deep learning method has been proposed by [8] to detect water using Reflection Attention Units (RAU) which captures the nature of reflection where a mirror image appears along vertical direction of the original object. Although the detection performance by [8] is high, high computational budget is also required making it difficult for realtime onboard processing for autonomous driving and advanced driving assistance.

Recent works on multi-class detection of water and other classes include [22] with ERF-PSPNet on RGB and depth images, and [3] with SegNet [1] on RBG and polarised images. These works have been based only on single image frame to perform detection.

While Han [8] utilised the effect image of reflection to improve performance of water detection, we argue that water puddle appearance changing at different angles and distances as a car moves can also be utilised to reveal critical information about a water puddle. As shown in [15], the appearance of water changes is a strong cue for detection. From a moving camera, such changes can be observed more clearly from series of images than from a single image. As a result, we believe that using temporal information can improve the accuracy of water detection.

In this work, we propose a new temporal 3D convolutional network (T3D-FCN) that exploits inter-frame temporal information to achieve better detection performance. This work is inspired by C3D network [18] for action recognition for extracting temporal information. We also borrow ideas from FCN [14] where fully convolutional layers are used successively with skips or shortcuts to improve segmentation resolution. Particularly, we come up with novel pooling techniques called temporal pooling techniques to deal with the shortcut and add lower abstract layers to the higher ones.

Our tests show that our proposed temporal 3D network improves detection performance significantly as compared with a similar network architecture without using temporal information. Our network also achieves comparable detection performance with the state-of-the-art network [8] but at a more realtime frame rate and much less memory and CUDA cores. Please note that our T3D-FCN is different from others [5], [6], [10], [20] in that our input data is 3D-spatial-temporal while others are fully 3D-spatial input data. Our work is also different from [16] in that their network lacks skips or shortcuts to improve the resolution of the output.

II. Method: T3D CNN with temporal pooling for water detection

We propose a new network architecture called T3D-FCN as an extension of C3D [18] and FCN [14] to extract temporal connections between continuous video frames to enhance the performance water detection. Figure 1 shows our proposed network architecture. The source code is available at [11].

The input to this network is a stack of 8 RGB video frames of 240 × 320 pixels. The input is passed forward through 3D convolutional blocks, each block has 2 convolutional layers,

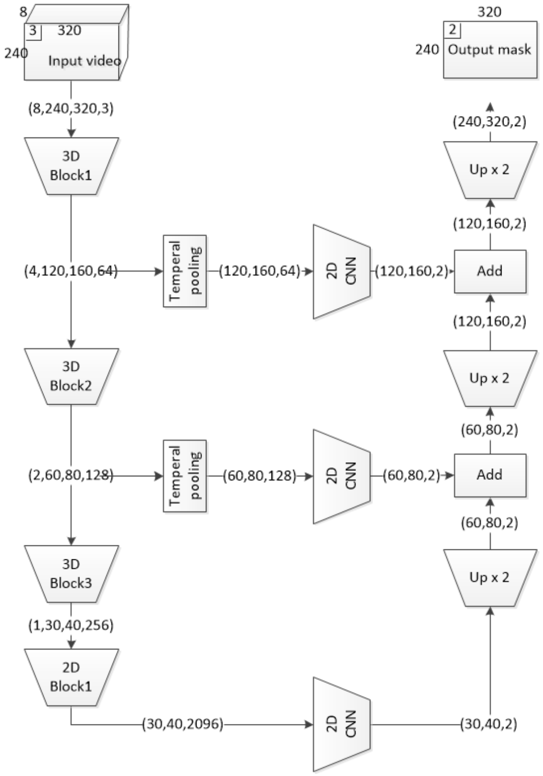

Fig. 1. T3D-FCN structure, each block contains 2 3D-convolutional layers and 1 3D-pooling layer, up-sampling is using deconvolutional layers. The numbers with brackets are the shape after each operation, the format is [[time, width, height, channel]. Brackets with only 3 numbers mean there is no temporal dimension. The number inside each small square is the channels or number of filters of that area.

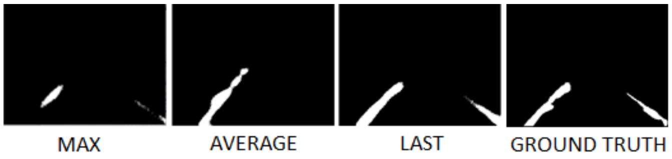

Fig. 2. Temporal pooling to reduce channels for shortcuts. Results water mask of 3 different temporal pooling methods Average, Max and Lasts. Last pooling method is used and recommended in this paper.

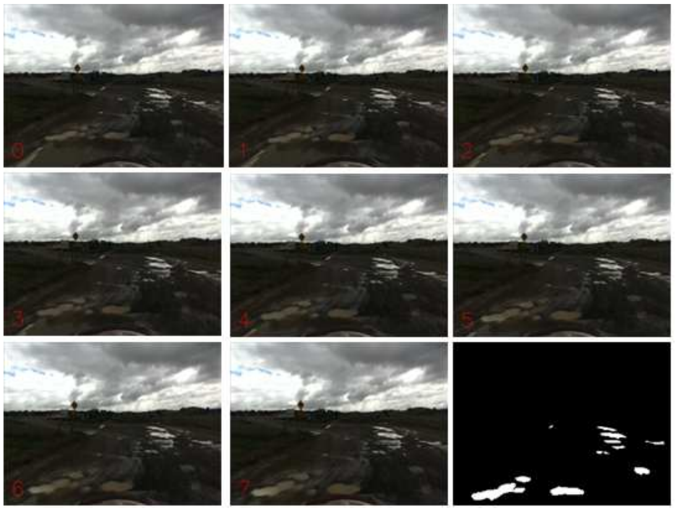

Fig. 3. Example of 8 continuous video frames used as input to our T3D-FCN network and the corresponding ground-truth water mask of the 8th frame used for training.

a batch normalisation and a 3D pooling layer. After each block, the size of the feature map is reduced in half in all 3 dimensions while the number of feature channels increases. 3D filters size of $3\times3\times3$ and 3D pooling size of $2\times2\times2$ are used in these blocks. After 3D Block 3, the temporal dimension is reduced to 1, while the number of filters increases to 256. The output 2D feature map from 3D Block 3 represent a summary of both spatial and temporal features of the 8 frames video. Then this is passed through two 2D convolutional layers with 2096 kernels of size 7×7 and 1×1, and another 2D convolutional layer with 2 kernels of size 1×1 where each pixel in the feature map is assigned with a class (either water pixel or non-water pixel). After that, the feature map is deconvolved to up-sample 2 times. Similar to FCN [14], in order to capture more spatial resolution, the up-sampled feature map needs to be combined with the higher-resolution lower-level feature maps from 3D Blocks 1 and 2.

To create shortcuts to improve the output resolution, we need to remove the temporal dimension from feature maps from 3D Blocks 1 and 2, and reduce the number of features to match with that of the outputs from the up-sampling layers (before adding them together). We propose to use one of the three temporal pooling functions (Average, Max and Last) to remove temporal dimension, and a 2D convolutional layer to reduce the number of features. Average pooling takes the average along the temporal dimension, while Max pooling takes the maximum, and Last pooling takes only the last temporal feature map. Different pooling functions result in different detection results as shown in Figure 2. From overall examination of the three outputs, the Last pooling gives the best result. This is further validated in the experimental section of the paper.

III. Experiments and Results

A. Experiments

The dataset used in this experiment is Puddle-1000 dataset introduced by Han et al [8]. There are in total 11455 continuous coloured stereo images divided into 2 datasets: On Road (ONR) dataset and Off Road (OFR) dataset. Out of these, 985 (approximately one for every 10 frames) of left images are annotated and provided with corresponding water masks. For on-road (ONR) dataset, the networks are trained on 272 videos and tested on 85 videos, for off-road (OFR) dataset, the networks are trained on 530 videos and tested on 98 videos. The focal loss [13] and stochastic gradient descent (SGD) are used to train the network.

In this paper, water is detected from the current and 7 previously consecutive frames, without any future frames. As a result, during training and testing for a given ground truth masks, 8 frames are selected from the current frame to the past 8th frame as shown in Figure 3.

We use a laptop with a GeForce GTX 980M GPU to perform all the trainings and testings. To fit the network into 4GB RAM of the GPU, all the images are resized to 240 × 320 pixels which is also convenient for pooling and up-sampling. All the water masks are One-hot encoded for SoftMax function.

To demonstrate the effectiveness of 3D convolution, we also tested another network called 2D-FCN that has similar architecture with the proposed T3D-FCN, except that the input is a single image, and 3D convolution layers are replaced by 2D convolution ones, and there are no temporal pollings. This network, running on the same GTX 980M GPU, allows us to quantify the effect of the additional temporal information and the additional computational cost of the proposed T3D-FCN.

B. Results

Figure 4 5 shows examples of water detection result for on-road and off-road data sets. Although the predictions roughly match with the ground truths, the former shows some blurriness and pepper-noise. Some fine details are missing from the predictions.

Table I shows the performance of our proposed T3D-FCN network with different pooling functions on on-road data set. Last pooling gives the best F1-measure and precision and inference time.

TABLE I
POOLING COMPARISON ON ON-ROAD DATASET

Poolings	F1-measure	Precision	Recall	Time (sec/frame)
Average (ours)	0.648	0.774	0.578	0.23103
Max (ours)	0.680	0.710	**0.662**	0.23001
Last (ours)	**0.684**	**0.788**	0.616	**0.22975**

Tables II and III show the comparisons of our proposed T3D-FCN network, its 2D-FNC reference network, and the result from FCN-8s-FL-RA, FCN-8s and DeepLab [8] and GMM & polarisation [15].

TABLE II
PERFORMANCE COMPARISON ON ON-ROAD DATASET

Methods	F1-measure	Precision	Recall	Time
T3D-FCN-LAST (ours)	0.68	**0.79**	0.62	0.23
2D-FCN (ours)	0.51	0.51	0.49	**0.20**
FCN-8s-FL-RAU [8]	**0.70**	0.68	**0.72**	0.32
FCN-8s [14]	0.57	0.59	0.55	0.06
DeepLab [4]	0.22	0.37	0.16	0.06
GMM & polar. [15]	0.31	0.19	0.90	NA

For both data sets, we can see a significant jump in performance from 2D-FCN to T3D-FCN by extending from single image to 8 continuous images. This suggests that changes of water appearance in viewing angle and therefore across multiple frames in time provides strong cue for detection. Furthermore, the significant performance improvement only requires a small increase in computation time (from 0.20 to 0.23 second/frame).

TABLE III
PERFORMANCE COMPARISON ON OFF-ROAD DATASET

Methods	F1-measure	Precision	Recall	Time
T3D-FCN-LAST (ours)	0.73	**0.87**	0.63	0.23
2D-FCN (ours)	0.56	0.68	0.52	**0.20**
FCN-8s-FL-RAU [8]	**0.81**	**0.87**	**0.77**	0.32
FCN-8s [14]	0.64	0.78	0.55	0.06
DeepLab [4]	0.45	0.71	0.33	0.06
GMM & polar. [15]	0.28	0.17	0.85	NA

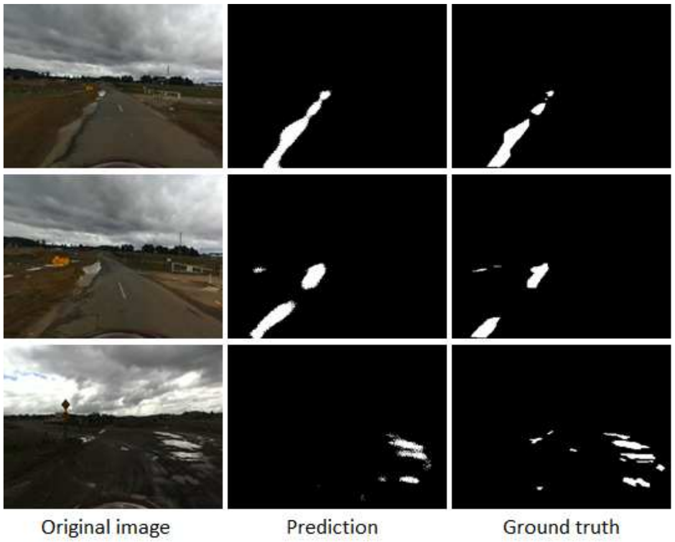

Fig. 4. Water puddle detection results of on-road data set. This video shows detection results on this data set at [12].

In addition, for both data sets, our proposed T3D-FCN performs near on-bar with the state-of-the-art method FCN-8s-FL-RAU [8] reported on the same data sets while requiring less time (0.22 vs 0.32 second), memory (4GB vs 12GB) and less CUDA cores (1536 Maxwell cores vs 3584 Pascal cores). This shows that our networks can improve its performance using a different strategy other than reflection as in [8]. This suggests that further improvement could be achieved by combining reflection and temporal information into a single deep learning framework.

As the original data [15] of Puddle-1000 dataset includes stereo image pairs to compute 3D depth of the scene, the water detection can be visualised in 3D. As it is difficult to compute directly the depth of a water puddle, it is assumed that they lie on the ground plane and have the same depth. The 3D puddles are then overlaid onto the 3D computed scene to help a user or operator to know their actual distances. An example is shown in Figure 6 where blue regions in image view and bird-eye view are detected puddles. The side view shows rays pointing from an external stereo camera (mounted on top of the car) to detected puddles, some of which extend outside the side view.

IV. CONCLUSION

By including temporal information across 8 continuous frames, our proposed T3D-FCN network can detect water at an accuracy close to state of the art while requires less memory

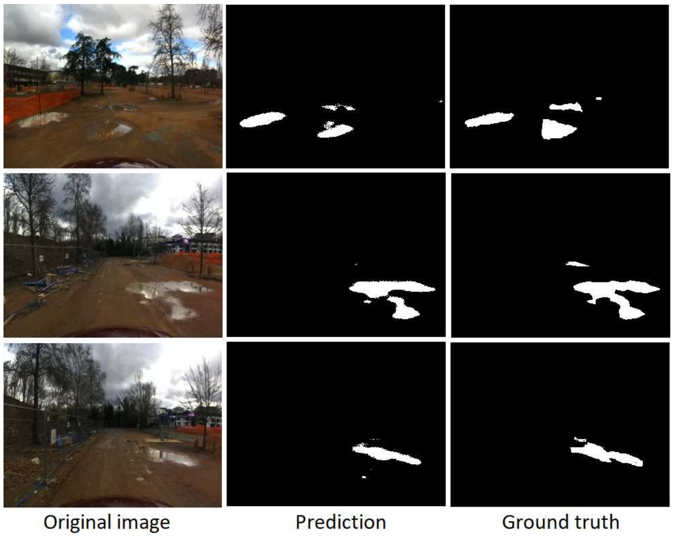

Fig. 5. Water puddle detection results of off-road data set.

and computational resources. In the network architecture, we propose 3 temporal pooling functions and show that the Last pooling function produces the best result. Our 3D-FCN can be trained on a good laptop GPU and detects water at near real-time frame rate. This is crucial for practical applications such as driverless navigation and advanced driving assistance. We have shown that by adding temporal information the detection performance is significantly improves with a small additional computation as compared to single image detection. Future work is to combine both reflection and temporal information for further improvement as these cues for water detection are complementary to each other.

References

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[2] Massimo Bertozzi, Rean Isabella Fedriga, and Carlo DAmbrosio. Adverse driving conditions alert: investigations on the swir bandwidth for road status monitoring. In *International Conference on Image Analysis and Processing*, pages 592–601. Springer, 2013.

[3] Marc Blanchon, Olivier Morel, Yifei Zhang, Ralph Seulin, Nathan Crombez, and Désiré Sidibé. Outdoor scenes pixel-wise semantic segmentation using polarimetry and fully convolutional network. 2019.

[4] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, April 2018.

[5] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016.

[6] Jose Dolz, Christian Desrosiers, and Ismail Ben Ayed. 3d fully convolutional networks for subcortical segmentation in mri: A large-scale study. *NeuroImage*, 170:456–470, 2018.

[7] Hiroshi Fukamizu, Masaji Nakano, Kunio Iba, Taro Yamasaki, and Kenji Sano. Road surface condition detection system, September 1 1987. US Patent 4,690,553.

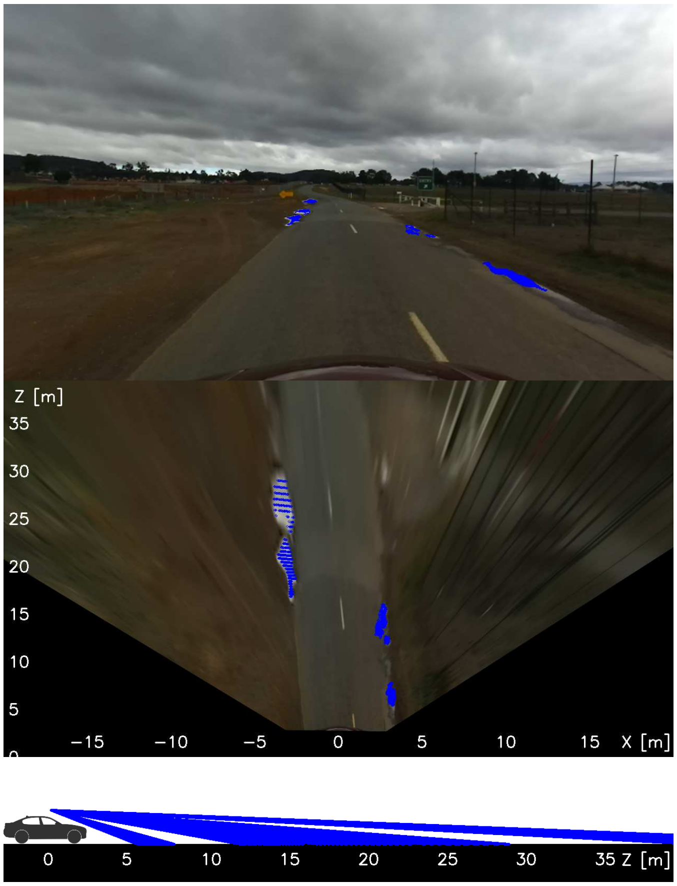

Fig. 6. 3D visualisation water puddle detection (blue) from on-road data set shown in image view (top), bird-eye view (middle) and side view (bottom).

[8] Xiaofeng Han, Chuong Nguyen, Shaodi You, and Jianfeng Lu. Single image water hazard detection using fcn with reflection attention units. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 105–120, 2018.

[9] Jisu Kim, Jeonghyun Baek, Hyukdoo Choi, and Euntai Kim. Wet area and puddle detection for advanced driver assistance systems (adas) using a stereo camera. *International Journal of Control, Automation and Systems*, 14(1):263–271, 2016.

[10] Bo Li. 3d fully convolutional network for vehicle detection in point cloud. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1513–1518. IEEE, 2017.

[11] Juntao Li and Chuong Nguyen. Github - temporal-3d-fully-connected-network-for-water-hazard-detection. https://github.com/Junta0Li/Temporal-3D-Fully-Connected-Network-for-Water-Hazard-Detection, September 2019.

[12] Juntao Li and Chuong Nguyen. Result of t3d-fcn classification - on road. https://youtu.be/YRF_GuvYeug, September 2019.

[13] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[14] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[15] Chuong V Nguyen, Michael Milford, and Robert Mahony. 3d tracking of water hazards with polarized stereo cameras. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5251–5257. IEEE, 2017.

[16] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning deep spatio-temporal dependence for semantic video segmentation. *IEEE Transactions on Multimedia*, 20(4):939–949, 2017.

[17] Arturo L Rankin, Larry H Matthies, and Paolo Bellutta. Daytime water detection based on sky reflections. In *2011 IEEE International Conference on Robotics and Automation*, pages 5329–5336. IEEE, 2011.

[18] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[19] Ville V Viikari, Timo Varpula, and Mikko Kantanen. Road-condition recognition using 24-ghz automotive radar. *IEEE transactions on intelligent transportation systems*, 10(4):639–648, 2009.

[20] Bo Wang, Yang Lei, Sibo Tian, Tonghe Wang, Yingzi Liu, Pretesh Patel, Ashesh B Jani, Hui Mao, Walter J Curran, Tian Liu, et al. Deeply supervised 3d fully convolutional networks with group dilated convolution for automatic mri prostate segmentation. *Medical physics*, 2019.

[21] Bin Xie, Huadong Pan, Zhiyu Xiang, and Jilin Liu. Polarization-based water hazards detection for autonomous off-road navigation. In *2007 International Conference on Mechatronics and Automation*, pages 1666–1670. IEEE, 2007.

[22] Kailun Yang, Luis M Bergasa, Eduardo Romera, Juan Wang, Kaiwei Wang, and Elena López. Perception framework of water hazards beyond traversability for real-world navigation assistance systems. In *2018 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 186–191. IEEE, 2018.

[23] Yibing Zhao, Yunxiang Deng, Chi Pan, and Lie Guo. Research of water hazard detection based on color and texture features. *Sensors & Transducers*, 157(10):428, 2013.