CONVOLUTIONAL GATED RECURRENT NETWORKS FOR VIDEO SEGMENTATION

Mennatullah Siam , Sepehr Valipour*, Martin Jagersand, Nilanjan Ray*

University of Alberta

{mennatul,valipour,mj7,nray1}@ualberta.ca

ABSTRACT

Semantic segmentation has recently witnessed major progress, but most of the previous work focused on improving single image segmentation. In this paper, we introduce a novel approach to implicitly utilize temporal data in videos for online segmentation. This design receives a sequence of consecutive video frames and outputs the segmentation of the last frame. Convolutional gated recurrent networks are used for the recurrent part to preserve spatial connectivities in the image. This architecture is tested for both binary and semantic video segmentation tasks. Experiments are conducted on the recent benchmarks in SegTrack V2, Davis, Camvid, and Synthia. Using recurrent fully convolutional networks improved the baseline network performance in all of our experiments. Namely, 5% and 3% improvement of F-measure in SegTrack2 and Davis respectively, 5.7% and 1.6% improvement in mean IoU in Synthia and Camvid. Thus, RFCN networks can be seen as a method to improve any baseline segmentation network by embedding them into a recurrent module that utilizes temporal data.

***Index Terms*—** Video Semantic Segmentation, Recurrent Networks

1. INTRODUCTION

Semantic segmentation, which provides pixel-wise labels, has witnessed tremendous progress recently. As shown in [2][4][1][16], they provide dense predictions and partition the image to semantically meaningful parts. The work in [2] presented the first method for end-to-end training of semantic segmentation. It yields a coarse heat-map followed by in-network upsampling to get dense predictions. In [4] a deeper deconvolution network was developed with stacked deconvolution and unpooling layers. Then a trend to incorporate contextual information appeared, where the work in [16] provided a method to incorporate context using recurrent neural networks. Most published methods perform single image segmentation on video frames, failing to utilize temporal coherence. One exception, uses a combination of Recurrent Neural Networks (RNN) and CNN for RGB-D video segmentation[12]. However, their proposed architecture is difficult to train because of the vanishing gradient problem. It does not utilize pre-trained networks and it cannot process large images as the number of their parameters is quadratic with respect to the input size.

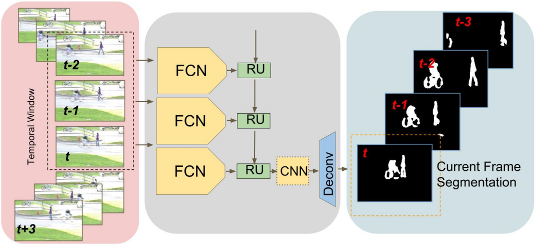

Fig. 1: Overview of the Proposed Method of Recurrent FCN. The recurrent part is unrolled for better visualisation

Gated Recurrent Architectures[7][8] alleviate the vanishing or exploding gradients problem in recurrent networks. The main bottleneck with these previous architectures is that they only work with sequences and therefore, do not preserve spatial information in images or feature maps. In[15] convolutional GRU is introduced for learning spatio-temporal features from videos and used for video captioning and action recognition. In [24] convolutional GRU was used with focus on object tracking and using 2D laser scans not raw images.

Inspired by these methods we design a gated recurrent FCN architecture to solve many of the shortcomings of the previous approaches. Contributions include:

- A novel architecture that can incorporate temporal data into FCN for video segmentation, and an end-to-end training method for online pixel-wise classification of videos.
- An experimental analysis on video binary segmentation and video semantic segmentation is presented on recent benchmarks.
- Experiments on a video collected from micro unmanned aerial vehicle (MAV) is also presented. To our knowledge no prior work on semantic segmentation is presented for MAV videos. This demonstrates different challenges like abrupt camera motion, and noisy video transmission.

Authors contributed equally

978-1-5090-2175-8/17/$31.00 ©2017 IEEE

Table 1: Proposed networks in detail. I denote input layer, C is convolution, R is Relu, P is pooling and D is deconvolution. Size denote filter size or image size according to layer type. Number of feature maps(#Chs) generated by the layer is same as the previous layer if it is not mentioned.

	Recurrent Node															
	I	C1	R1	P1	C2	R2	P2	C3	R3	C4	R4	C5	R5	C6	R6	C7	D
Size	240X360	11	-	3	5	-	3	3	-	3	3	-	3	3	3	1	20
Stride	-	4	-	-	-	-	-	-	-	-	-	-	-	-	-	8	
Pad	-	40	-	-	2	-	-	1	-	1	-	-					
#Chs	3	64	-	-	256	-	-	256	-	256	-	512	-	128	-	1	-

The paper is structured as follows. The proposed method is presented in details in section 2. Section 3 presents experimental results and discussion on recent benchmarks. Finally, section 4 summarizes the findings.

2. PROPOSED METHOD

An overview of the method is presented in Figure 1. A recurrent fully convolutional network (RFCN) is designed that utilizes the spatiotemporal information for video segmentation. The recurrent unit in the network can either be LSTM, GRU or Conv-GRU (which is explained in 2.2). A sliding window over the video frames is used as input to the network. This allows on-line video segmentation as opposed to off-line batch processing. The window data is forwarded through the RFCN to yield a segmentation for the last frame in the sliding window. Note that the recurrent unit can be applied on the coarse segmentation (heat map) or intermediate feature maps. Two main approaches are explored in our method: (1) conventional recurrent units, and (2) convolutional recurrent units 1. Specifically, three different network architectures under these two approaches are used as detailed in the following sections.

2.1. Conventional Recurrent Architecture for Segmentation

RFC-Lenet is a fully convolutional version of Lenet. Lenet is a well known shallow network. Because it is common, we used it for baseline comparisons on synthetic data. We embed this model in a recurrent node to capture temporal data. The output of deconvolution is a 2D map of dense predictions that is then flattened into 1D vector as the input to a conventional recurrent unit. The recurrent unit then outputs the segmentation of the last frame (Figure 1).

2.2. Convolutional Gated Recurrent Architecture (Conv-GRU) for Segmentation

Conventional recurrent units are designed for processing text data not images. Therefore, using them on images without any modification causes two main issues. 1) The size of weight parameters becomes very large since vectorized images are large 2) Spatial connectivity between pixels are ignored. For example, using a recurrent unit on a feature map with the spatial size of $h \times w$ and number of channels c requires $c \times (h.w)^2$ number of weights. This will cause a memory bottleneck and inefficient computations. It will also create a larger search space for the optimizer, thus it will be harder to train.

Convolutional recurrent units, akin to regular convolutional layer, convolve three dimensional weights with their input. Therefore, to convert a gated architecture to a convolutional one, dot products should be replaced with convolutions. The following equations show this modification for the GRU. The weights are of size of $k_h \times k_w \times c \times f$ where k_h, k_w, c and f are kernel's height and width, number of input channels, and number of filters, respectively. Learning filters that convolve with the entire image instead of learning individual weights for each pixel, makes it much more efficient. This layer can be applied on either final heat map or intermediate feature maps.

$$
\begin{aligned}
z_t &= \sigma(W_{hz} * h_{t-1} + W_{xz} * x_t + b_z) \\
r_t &= \sigma(W_{hr} * h_{t-1} + W_{xr} * x_t + b_r) \\
\hat{h_t} &= \Phi(W_h * (r_t \odot h_{t-1}) + W_x * x_t + b) \\
h_t &= (1 - z_t) \odot h_{t-1} + z \odot \hat{h_t}
\end{aligned}
$$

RFC-VGG in Table 1 is an example of this approach, where intermediate feature maps are fed into a convolutional gated recurrent unit. Then a convolutional layer converts its output to a heat map. It is based on VGG-F [11] network. The reason for switching to the RFC-VGG architecture is to use pre-trained weights from VGG-F. Initializing weights of our filters by VGG-F trained weights, alleviates over-fitting problems as these weights are the result of extensive training on Imagenet dataset. The last two pooling layers are dropped from VGG-F to allow a finer segmentation with a reduced network.

RFC-Dilated is the recurrent version of FC-Dilated architecture and is used in our semantic segmentation experiments. It uses dilated convolution to increase receptive field while maintaining the resolution of the segmentation. FC-dilated is an adapted version of FCN-16s as originally introduced in [2].

Two pooling are removed and conv4 and conv 5 layers are reduced to two dilated convolution layers with dilation factor 2, and 4 respectively.

3. EXPERIMENTS

This section describes the experimental analysis and results. First, the datasets are presented followed by the training method and hyper-parameters used. Then both quantitative and qualitative analysis are shown. All experiments are performed on our implemented open source library that supports convolutional gated recurrent architectures. The implementation is based on Theano [9].

The key features of this library are: **(1)** The ability to use any arbitrary CNN or FCN architecture as a recurrent node. In order to utilize temporal information. **(2)** Implementation of three gated recurrent architectures which are, LSTM, GRU, and Conv-GRU as of now. **(3)** It includes deconvolution layer for in the network upsampling and supports skip architecture for finer segmentation. A public version of the code for the library along with the trained models is published[26].

3.1. Experimental Setup

Five datasets are used in the experimental analysis: **(1)** Segtrack version 2[14]. **2)** Densely Annotated VIdeo Segmentation (Davis) [13]. **3)** Synthia[18] has over 200,000 images with different weather conditions (rainy, sunset, winter) and seasons. Since the dataset is large only a portion of it from Highway sequence is used for our experiments. **4)** Camvid[23]. **(5)** A sequence collected from AR-Drone with abrupt camera motion.

The main experiments' setup includes using Adadelta [10]. The loss function used throughout the experiments is the logistic loss, and the maximum number of epochs used for training is 500. The evaluation metrics used for the binary video segmentation is precision, recall, F-measure and IoU. As for multi-class segmentation mean class IoU is used.

3.2. Results

In the initial experiments RFC-Lenet is compared to FC-Lenet baseline on a synthetic moving MNIST dataset. The RFC-Lenet outperforms the baseline with 2.2% in F-measure. The second phase of experiments the network denoted as FC-VGG and RFC-VGG are compared. To avoid overfitting, the first five layers of the network are initialized with the weights of a pre-trained networked and only lightly tuned. Table 2 shows the results of the experiments on SegTrackV2 and DAVIS datasets. In these experiments, the data is split into half for training and the other half as keep out test set. RFC-VGG outperforms the FC-VGG architecture on both datasets with about 3% and 5% on DAVIS and SegTrack respectively. Figure 2 shows the qualitative analysis of RFC-VGG against FC-VGG. It shows that utilizing temporal information through the recurrent unit gives better segmentation for the object. This can be contributed to the implicit learning of the motion of segmented objects in the recurrent units.

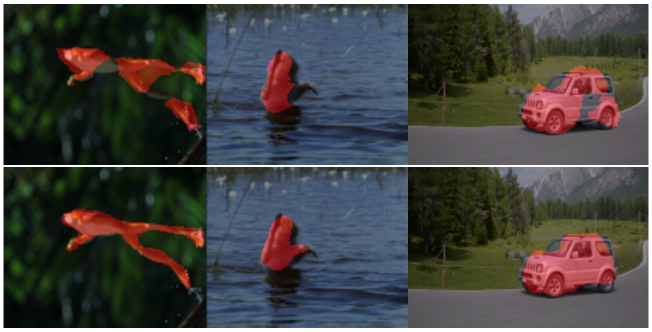

Fig. 2: Qualitative results of experiments with SegtracV2 and Davis datasets, where network prediction are overlaid on the input. The top row is for FC-VGG and the bottom row is for RFC-VGG.

The same architecture was used for semantic segmentation on synthia dataset after modifying it to support the thirteen classes. A comparison between FC-VGG and RFC-VGG is conducted. Table4 presents the results on synthia dataset and AR-Drone sequence. RFC-VGG has 5.7% over FC-VGG in terms of mean class IoU in Synthia. Figure3shows the qualitative analysis on Synthia and AR-Drone data.

Finally, experimental results on camvid dataset is shown in Table3. It showed 1.6% improvement over the baseline. Note that the performance of the RFCN network depends on its baseline fully convolutional network. Thus, RFCN networks can be seen as a method to improve any baseline segmentation network by embedding them into a recurrent module that utilizes temporal data. Note also that the FCN baseline used in our experiments is five times faster than FCN8s[2] with 0.07 versus 0.33 seconds tested on TITAN X GPU.

4. CONCLUSION

We presented a novel method that exploits implicit temporal information in videos to improve segmentation. This approach utilizes convolutional gated recurrent network which allows it to use preceding frames in segmenting the current frame. We performed extensive experiments on six datasets. We showed that embedding FCN networks as a recurrent module, consistently improved the results through out different datasets. Specifically, a 5% improvement in Segtrack and 3% improvement in Davis in F-measure over a plain fully convolutional network; a 5.7% improvement on Synthia in mean IoU, and 1.4% improvement on Camvid.

Table 2: Comparison of RFC-VGG with its baseline counterpart on DAVIS and SegTrack

		Precision	Recall	F-measure	IoU
SegTrack V2	FC-VGG	0.7759	0.6810	0.7254	0.7646
	RFC-VGG	**0.8325**	0.7280	**0.7767**	**0.8012**
DAVIS	FC-VGG	0.6834	0.5454	0.6066	0.6836
	RFC-VGG	**0.7233**	**0.5586**	**0.6304**	**0.6984**

Table 3: Semantic Segmentation Results on Camvid for RFC-Dilated compared to FC-Dilated.

	Mean Class IoU	Per-Class IoU						
		Sky	Building	Road	Sidewalk	Vegetation	Car	Pedestrian
FC-Dilated	46.7	86.3	69.1	87.8	63.7	60.8	63.6	21.4
RFC-Dilated	**48.3**	**87.5**	**69.1**	**89.4**	**69.4**	**62.0**	**64.3**	**24.3**
Super Parsing[5]	42.0	-	-	-	-	-	-	-
Segnet[3]	46.4	-	-	-	-	-	-	-

Table 4: Semantic Segmentation Results on Camvid

	FCN	RFCN
Synthia	0.755	**0.812**
ARDrone	0.857	**0.871**

5. REFERENCES

[1] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr, "Conditional random fields as recurrent neural networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1529–1537.

[2] 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*, 2015, pp. 3431–3440.

[3] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *arXiv preprint arXiv:1511.00561*, 2015.

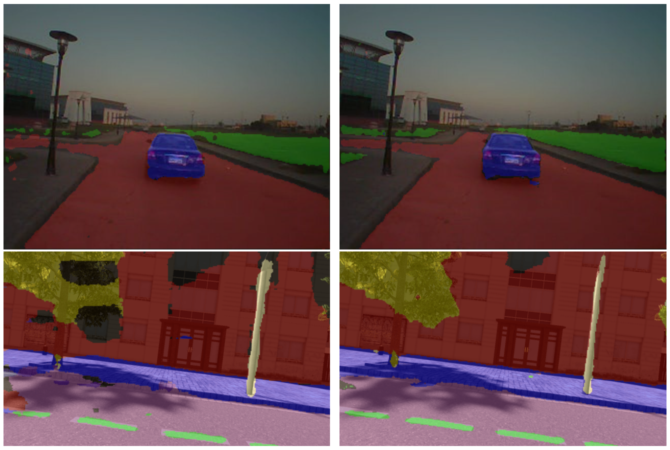

Fig. 3: Qualitative results of experiments with Synthia dataset, where network prediction are overlaid on the input. First row: Synthia with FC-VGG. Second row: Synthia with RFC-VGG.

[4] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han, "Learning deconvolution network for semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1520–1528.

[5] Joseph Tighe and Svetlana Lazebnik, "Superparsing: scalable nonparametric image parsing with superpixels,"

in *European conference on computer vision*. Springer, 2010, pp. 352–365.

[6] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[7] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[8] Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint arXiv:1409.1259*, 2014.

[9] Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, James Bergstra, Ian J. Goodfellow, Arnaud Bergeron, Nicolas Bouchard, and Yoshua Bengio, "Theano: new features and speed improvements," Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop, 2012.

[10] Matthew D Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.

[11] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[12] Mircea Serban Pavel, Hannes Schulz, and Sven Behnke, "Recurrent convolutional neural networks for object-class segmentation of rgb-d video," in *Neural Networks (IJCNN), 2015 International Joint Conference on*. IEEE, 2015, pp. 1–8.

[13] F Perazzi, J Pont-Tuset1 B McWilliams, L Van Gool, M Gross, and A Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," .

[14] Fuxin Li, Taeyoung Kim, Ahmad Humayun, David Tsai, and James M Rehg, "Video segmentation by tracking many figure-ground segments," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 2192–2199.

[15] Nicolas Ballas, Li Yao, Chris Pal, and Aaron Courville, "Delving deeper into convolutional networks for learning video representations," *arXiv preprint arXiv:1511.06432*, 2015.

[16] Francesco Visin, Kyle Kastner, Aaron Courville, Yoshua Bengio, Matteo Matteucci, and Kyunghyun Cho, "Reseg: A recurrent neural network for object segmentation," *arXiv preprint arXiv:1511.07053*, 2015.

[17] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele, "The cityscapes dataset for semantic urban scene understanding," *arXiv preprint arXiv:1604.01685*, 2016.

[18] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez, "The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3234–3243.

[19] Hongyi Zhang, Andreas Geiger, and Raquel Urtasun, "Understanding high-level semantics by modeling traffic patterns," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 3056–3063.

[20] Ondrej Miksik, Vibhav Vineet, Morten Lidegaard, Ram Prasaath, Matthias Nießner, Stuart Golodetz, Stephen L Hicks, Patrick Pérez, Shahram Izadi, and Philip HS Torr, "The semantic paintbrush: Interactive 3d mapping and recognition in large outdoor spaces," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2015, pp. 3317–3326.

[21] Vibhav Vineet, Ondrej Miksik, Morten Lidegaard, Matthias Nießner, Stuart Golodetz, Victor A. Prisacariu, Olaf Kähler, David W. Murray, Shahram Izadi, Patrick Perez, and Philip H. S. Torr, "Incremental dense semantic stereo fusion for large-scale semantic scene reconstruction," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2015.

[22] Daniel Wolf, Johann Prankl, and Markus Vincze, "Enhancing semantic segmentation for robotics: The power of 3-d entangled forests," *IEEE Robotics and Automation Letters*, vol. 1, no. 1, pp. 49–56, 2016.

[23] Gabriel J. Brostow, Jamie Shotton, Julien Fauqueur, and Roberto Cipolla, "Segmentation and recognition using structure from motion point clouds," in *ECCV (1)*, 2008, pp. 44–57.

[24] Peter Ondruska, Julie Dequaire, Dominic Zeng Wang, and Ingmar Posner, "End-to-end tracking and semantic segmentation using recurrent neural networks," *arXiv preprint arXiv:1604.05091*, 2016.

[25] Viorica Patraucean, Ankur Handa, and Roberto Cipolla, "Spatio-temporal video autoencoder with differentiable memory," *arXiv preprint arXiv:1511.06309*, 2015.

[26] "RFCNN Repository," `https://gitlab.com/sepehr.valipour/RFCNN`.