Motion U-Net: Multi-cue Encoder-Decoder Network for Motion Segmentation

Gani Rahmon[1], Filiz Bunyak[1], Guna Seetharaman[2], Kannappan Palaniappan[1]
[1] University of Missouri-Columbia, MO, USA
[2] U.S Naval Research Laboratory, Washington, DC, USA

1. Overview

Introduced a novel hybrid moving object detection system, Motion U-Net (MU-Net),

- integrates motion, change, and appearance cues,
- an encoder-decoder deep convolutional neural network for robust moving object detection.

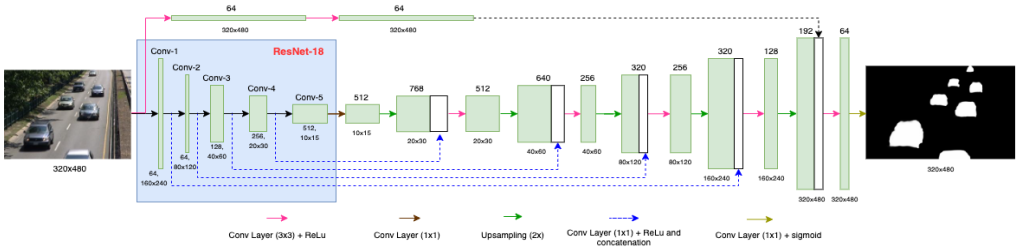
Proposed two versions of MU-Net:

- MU-Net1 uses only a single RGB frame without temporal information,
- MU-Net2 uses a three channel input stream consisting of RGB frame converted to grayscale, motion cue (flux tensor) and change cue (background subtraction).

2. Motion U-Net Networks

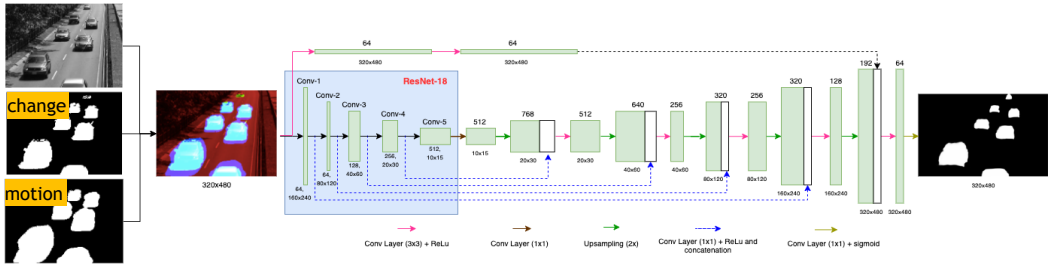
1.MU-Net1: Single-stream Spatial-only Detection Using Semantic Segmentation

- Input: Single RGB frame; **Output:** Binary mask; **Backbone:** ResNet-18



2. MU-Net2: Single-stream Early Fusion for Spatio-temporal Change Detection

-Input: RGB frame converted to grayscale, change (BG subtraction) and motion (flux tensor) masks; **Output:** Binary mask; **Backbone:** ResNet-18



- **Change:** estimated using adaptive multi-modal background subtraction (OpenCV Background SubtractorMOG2).
- **Motion:** estimated using our efficient flux tensor (J_F) motion computation scheme that represents the temporal variation of the optical flow field within local 3D spatio-temporal volume.

$$J_F = \begin{bmatrix} \int_\Omega \left\{\frac{d^2 I}{dx\,dt}\right\}^2 dy & \int_\Omega \frac{d^2 I}{dx\,dt}\frac{d^2 I}{dy\,dt} dy & \int_\Omega \frac{d^2 I}{dx\,dt}\frac{d^2 I}{dt^2} dy \\ \int_\Omega \frac{d^2 I}{dy\,dt}\frac{d^2 I}{dx\,dt} dy & \int_\Omega \left\{\frac{d^2 I}{dy\,dt}\right\}^2 dy & \int_\Omega \frac{d^2 I}{dy\,dt}\frac{d^2 I}{dt^2} dy \\ \int_\Omega \frac{d^2 I}{dt^2}\frac{d^2 I}{dx\,dt} dy & \int_\Omega \frac{d^2 I}{dt^2}\frac{d^2 I}{dy\,dt} dy & \int_\Omega \left\{\frac{d^2 I}{dt^2}\right\}^2 dy \end{bmatrix}$$

$$trace(J_F) = \int_\Omega ||\frac{d}{dt}\nabla I||^2 dy$$

3. Experimental Results

Comparison to top-performing methods on CDnet

Methods	Overall (CDnet-2014 videos)				
	Rank	Re	PWC	P	F
FgSegNet_v2	1	0.989	0.040	0.982	0.985
FgSegNet_S	2	0.989	0.046	0.975	0.980
FgSegNet	3	0.984	0.056	0.976	0.977
BSPVGAN	4	0.954	0.227	0.950	0.947
BSGAN	5	0.947	0.328	0.923	0.934
Cascade CNN	6	0.951	0.405	0.899	0.921
MU-Net1		0.927	0.209	0.941	0.915
MU-Net2		0.945	0.235	0.941	0.937

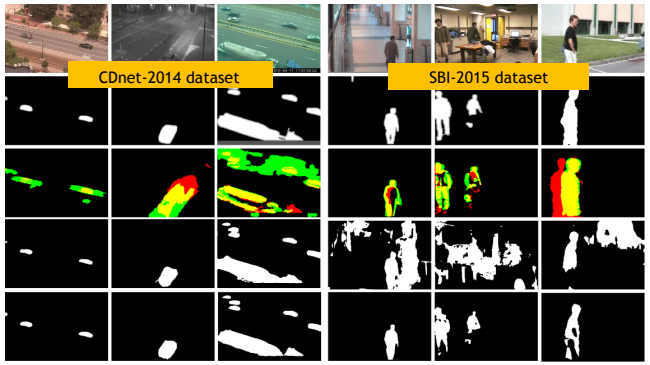
Generalization capabilities of MU-Net models on unseen SBI-2015 videos

Methods	Overall (SBI-2015 videos)			
	Re	PWC	P	F
MU-Net1	0.809	15.835	0.288	0.378
Motion U Change	0.836	11.088	0.411	0.519
MU-Net2	0.730	2.683	0.848	0.763
FgSegNet _v2(50%)	0.242	5.772	0.815	0.352

Training parameters and time of MU-Net with FgSegNet_v2

Methods	# models	GPU	Train Time
FgSegNet_v2	53	GTX 1080 Ti	29 days
MU-Net1	1	GTX 1080 Ti	4 hrs
MU-Net2	1	Tesla V100	4.5 hrs

Methods	Network Size (# parameters)
FgSegNet_v2	489 M (53 * 9,225,161)
MU-Net1	17.8 M (1 * 17,799,809)
MU-Net2	17.8 M (1 * 17,799,809)



Input Images

Ground Truth Mask

Red: motion mask
Green: change mask

MU-Net1 respons

MU-Net2 response

- Decoupled, unsupervised motion (flux tensor) and change (BG subtraction): leads to reduced network complexity, training times, and need for training data.
- Motion U-Net able to learn fusion, object level reasoning, and semantic analysis using only 8% of the CDnet-2014 labeled video frames.
- Great performance improvement on unseen data compared to appearance-only MU-Net1, and the top ranked FgSegNet_v2 (from 37% and 35% to 76%).