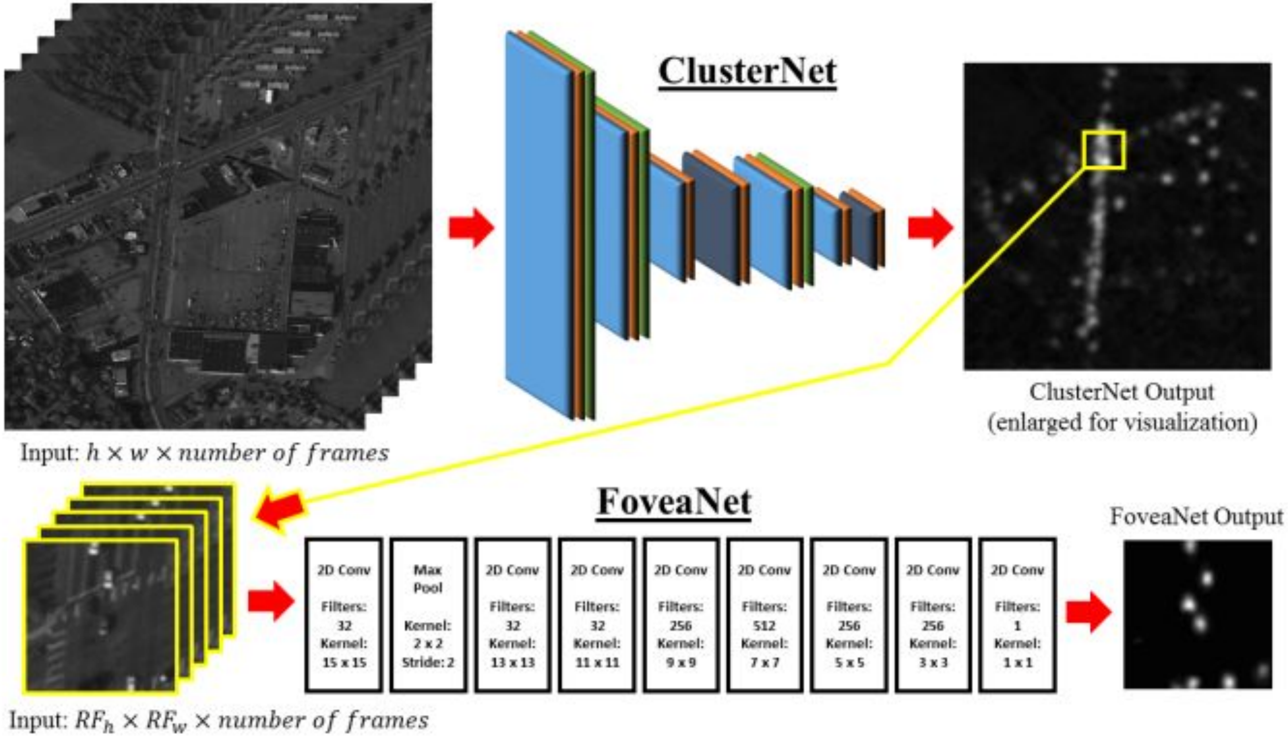
ClusterNet: Detecting Small Objects in Large Scenes by Exploiting Spatio-Temporal Information

Original work by Rodney Lalonde
Reimplementation in pytorch: Babak Ebrahimi
babak@knights.ucf.edu

Schematic of the network



Dataset: WPAFB2009

Data set consists of 8 folders: AOI 01, AOI 02, AOI 03, AOI 04, AOI 34, AOI 40, AOI 41, AOI 42

Sample image form AOI 01, size 2278*2278

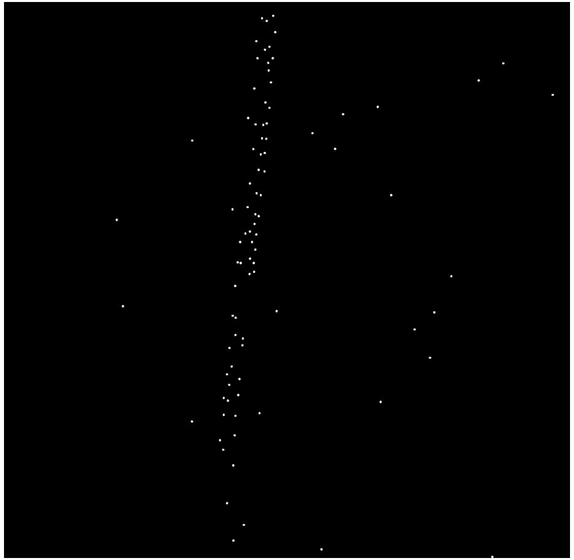
Sample image form AOI 02, size 2278*2278

Dataset

- Data set consists of 8 folders: AOI 01, AOI 02, AOI 03, AOI 04, AOI 34, AOI 40, AOI 41, AOI 42 in .pgm format
- Area of interests (AOI) 01, 02, 03, 04 and 42 each one has 1025 images and AOIs 34, 40 and 41 each one has 512 images
- For training we need two set of ground truths:
  - First set of ground truths for Clusternet (upper network)
  - Second set of ground truths for Foveanet (lower network)

Sample Input Image, size 2278*2278

Sample Clusternet groundtruth, size 72*72



Sample Foveanet Groundtruth, size 2278*2278

Network Structure: Cluster net (lower network)

- Number of input frames=5
- Use middle frame ground truth as the network label during training and testing
- Input size 2278*2278
- Output size 72*72
- Training size=800
- Testing size=225
- Batch Size=8
- Loss function = MSELoss
- Optimizer=Adam
- Threshold set to 0.5 for each pixel
- learning_rate = 0.001

Foveanet network Structure

layer#	Layer	# of Input channels	# of Output channels	Kernel size	Stride	Padding
1	Conv2d +batchnorm2d+PRelu	5	16	3	2	2
1	Maxpool2d			2	2	
2	Conv2d +batchnorm2d+PRelu	16	32	3	2	2
2	Maxpool2d			2	2	
3	Conv2d +batchnorm2d+PRelu	32	64	3	1	2
4	Conv2d +batchnorm2d+PRelu	64	32	1	1	0
5	Conv2d +batchnorm2d+PRelu	32	64	3	1	1
5	Maxpool2d			2	2	
6	Conv2d +batchnorm2d+PRelu	64	128	3	1	1
7	Conv2d +Sigmoid	128	1	1	1	0

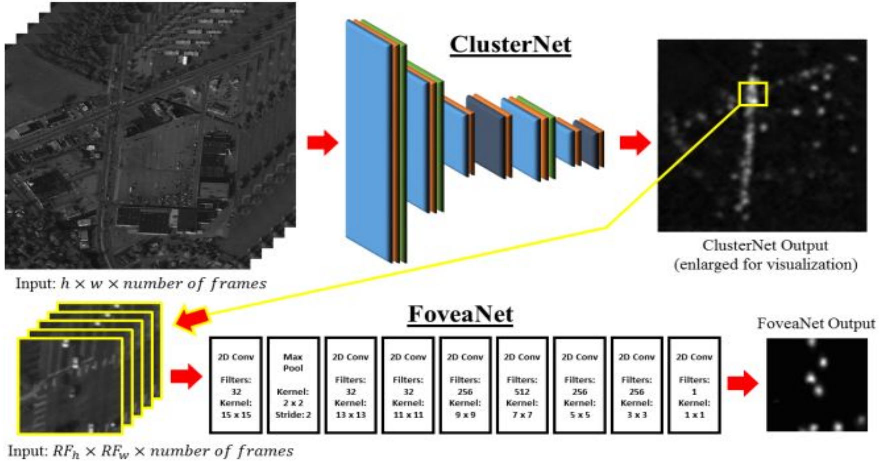
Network Structure: Fovea net (lower network)

- Number of input frames=5
- Use middle frame ground truth as the network label during training and testing
- Input size 261*261
- Output size 130*130
- Training size=800
- Testing size=225
- Batch Size=32
- Loss function = MSELoss
- Optimizer=Adam
- Threshold set to 0.5 for each pixel
- learning_rate = 0.0001

Clusternet network Structure

layer#	Layer	# of Input channels	# of Output channels	Kernel size	Stride	Padding
1	Conv2d +batchnorm2d+Relu	5	16	15	1	7
1	Maxpool2d			2	2	
2	Conv2d +batchnorm2d+Relu	16	32	13	1	6
3	Conv2d +batchnorm2d+Relu	32	64	11	1	5
4	Conv2d +batchnorm2d+Relu	64	256	9	1	4
5	Conv2d +batchnorm2d+Relu	256	512	7	1	3
6	Conv2d +batchnorm2d+Relu	512	256	5	1	2
7	Conv2d +batchnorm2d+Relu	256	128	3	1	1
8	Conv2d +Sigmoid	128	1	1	1	0

How to train and test the network?



Step 1: (pre-processing) Create the groundthuth for training data. Input pairs are 2278x 2278 grayscale images & and their corresponding binary outputs are 72*72 where the dots represents moving cars in our training data.

Step 2: Train the ClusterNet with input of size 2278 and output of size 72*72

Step 3: Train the Fovea net with Groundtruth (2278*2278) grid pieces (130*130) and corresponding receptive fields in original images (261*261). (only use the corresponding receptive filed of those pieces of the groundtruth grid which has intensity mor than the threshold)

Step 4: For testing feed the input images to the clusternet and compute the corresponding receptive fields(261*261) of grid pieces(4*4) with intensity more than threshold, Then feed the computed corresponding receptive fields to the Fovea net

Step 5: (post processing) Stick the FoveaNet outputs to make the original ground truth size result (output resized after handling overlapped area)

Step 6: Compute precision and Recall

Results of clusternet:

Sample input image (size 2278*2278)

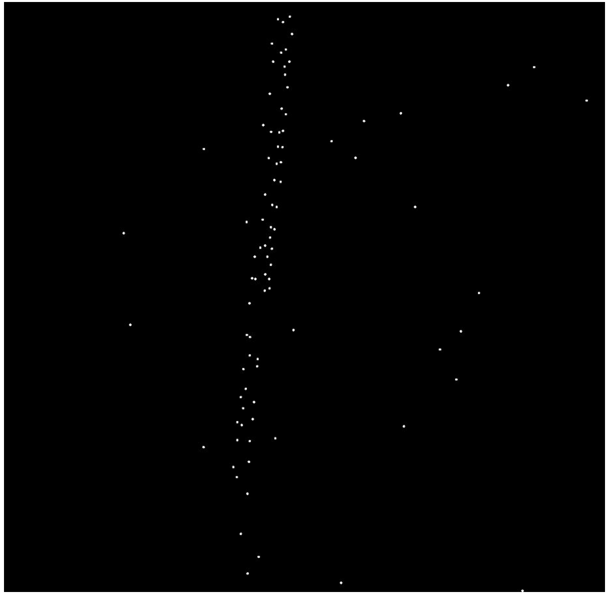
Sample Ground truth:

Sample generated output:

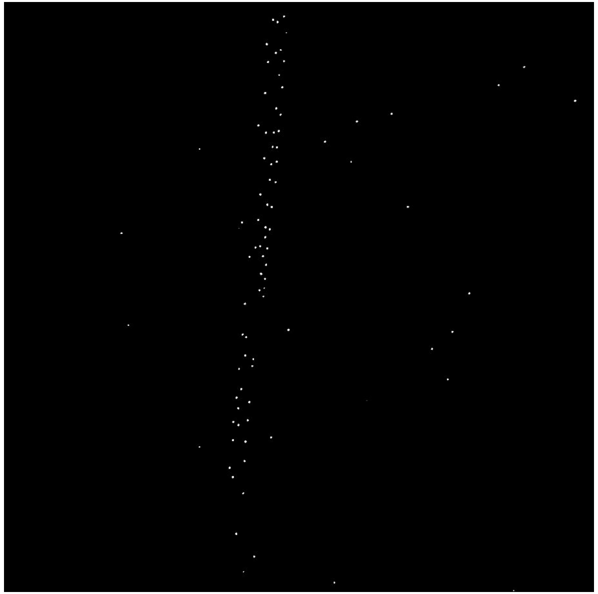
Results of Foveanet:

Sample input image (size 2278*2278)

Sample Groundtruth image (size 2278*2278)



Sample genrated output (size 2278*2278)



Results of Testing on AOI 01:

- 800 Images used for training
- 225 Images used for testing
- Threshold=0.5

- **Precision: 0.980892969**
- **Recall: 0.920958782**
- **F1 measure: 0.94774426**

references

- Fully convolutional deep neural networks for persistent multi-frame multi-object detection in wide area aerial videos, R LaLonde, D Zhang, M Shah - arXiv preprint arXiv:1704.02694, 2017 - arxiv.org