Temporal Feature Enhancement Network with External Memory for Object Detection in Surveillance Video

Masato Fujitake[1], Akihiro Sugimoto[2]
[1]Dept. of Informatics, The Graduate University for Advanced Studies, SOKENDAI
[2]National Institute of Informatics Tokyo, Japan

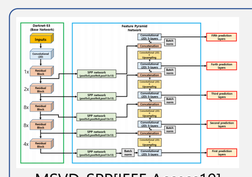
ICPR2020 25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION Milan, Italy 10 | 15 January 2021

Motivation

Task

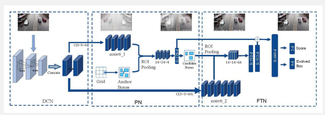
Surveillance object detection offers challenges like…
- Dense small objects detection in high resolution
- Blur
- Out of focus
- Occlusion

Problems



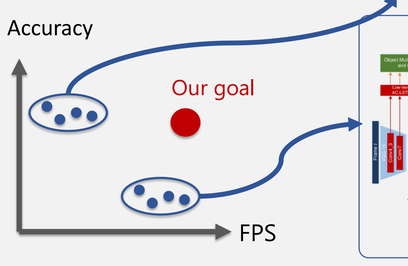
MSVD_SPP[IEEE Access19]

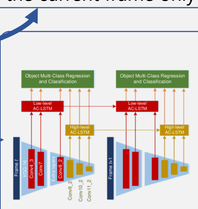
GP-FRCNN[AVSS17]



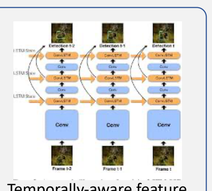
EB[ICME17]

Processing with the current frame only





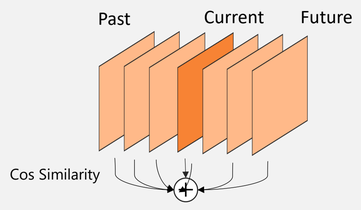
TSSD[IROS18]



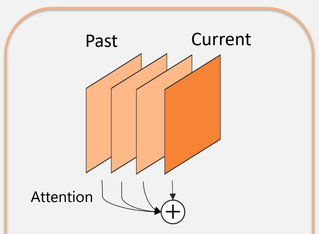
Temporally-aware feature maps[CVPR18]

Processing with time-series frames

Key Idea



Flow-guided Feature Aggregation [ICCV17]

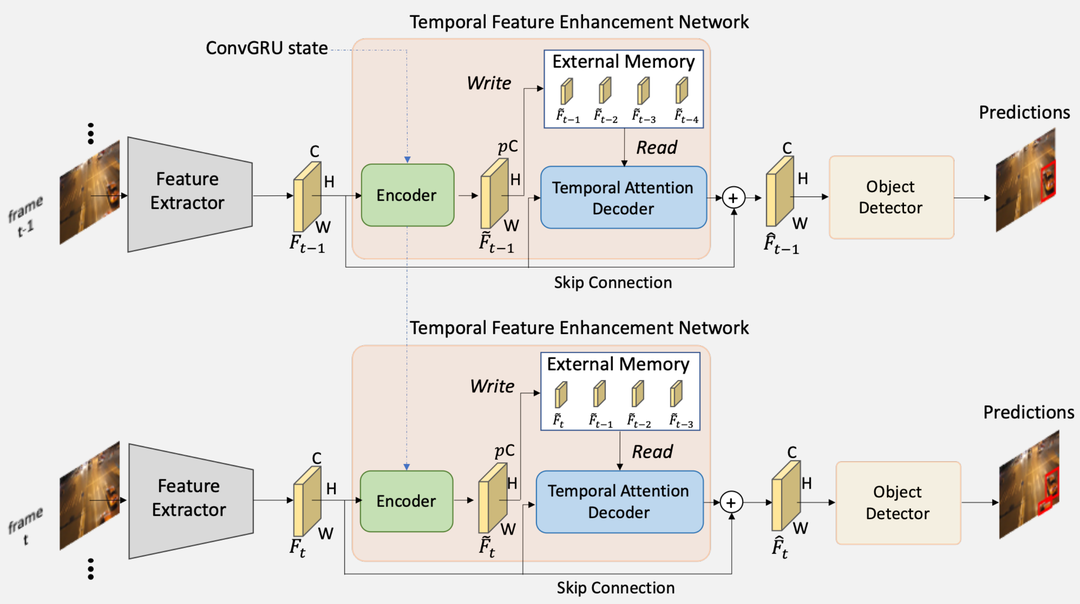


Appearance-based Feature Aggregation with Attention

- Aggregating coarse feature maps from lightweight extractors in the time direction for improvement.
- Predict aggregation weights with attention mechanism, not using the similarity of each frame.

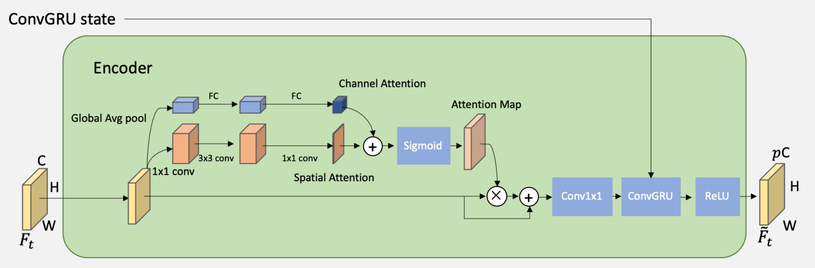
Proposed Method

Overall Architecture



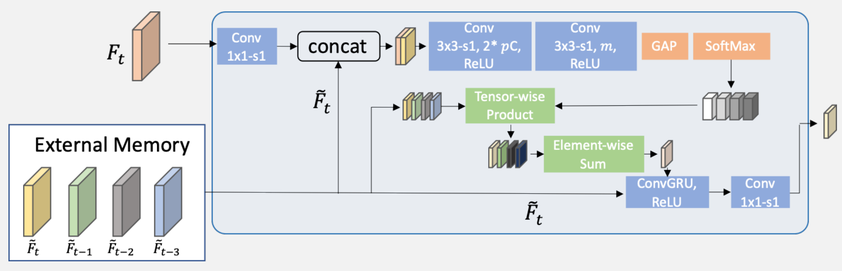
- Insert TFEN (Temporal Feature Enhancement Network) to the existing detectors.
- Improve accuracy by improving feature maps from feature extractors.

Spatiotemporal Encoder



- ConvGRU to generate temporal feature maps.
- Spatial & channel attention [3] to emphasize object areas.
- Compress channels for computational reduction

Temporal Attention Decoder & External Memory



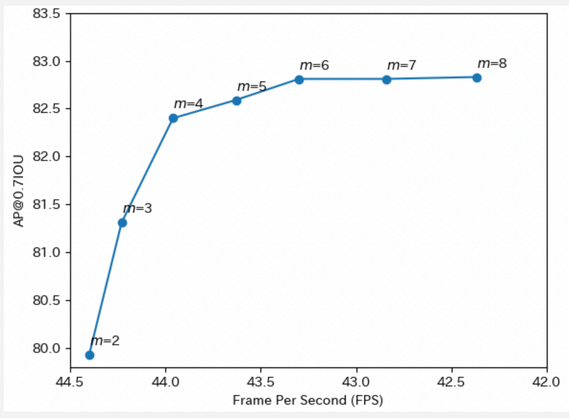
- Attentional weighting for frames in External Memory from F_t and temporal feature map $\tilde{F}_t$
- Feature aggregation in External Memory to generate the feature map based on the attention weight.

Experiment Results

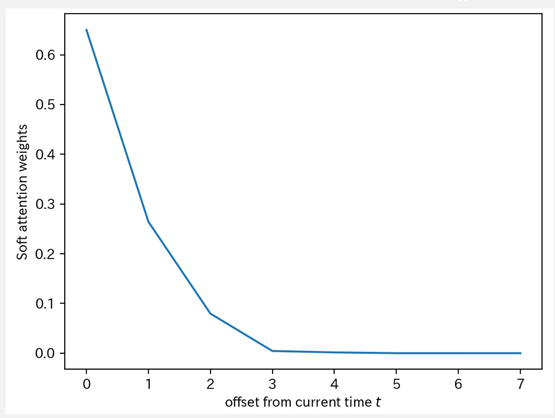
Dataset & Implementation

- Large-Scale Surveillance Camera Video Data Set (UA-DETRAC) [2]
  - With over 140,000 frames, the training and testing videos consist of 60, 40
  - Shot at a resolution of 960x540
- The evaluation AP@IoU0.7 of the test set are as follows
  - Overall
  - Difficulty level (Easy, Medium, Hard)
  - Climatic conditions (Cloudy, Night, Rainy, Sunny)
- Baseline Model (FP32)
  - Feature Extractor: MobileNetV2 [4]
  - Object Detector: Cascade R-CNN [1]

Effect of the frame number in External Memory



AP v.s. FPS under different number m of frames to be stored in the external memory



Soft attention weights used in the temporal decoder (m=8).

- From the AP& FPS trade-off figure, increasing the number of stored frames tends to improve the accuracy.
- m=4~6 would be the accuracy/speed trade-off point.
- From the soft attention weight figure, the coefficients after third frame are extremely small, so m=4 is enough.

Results

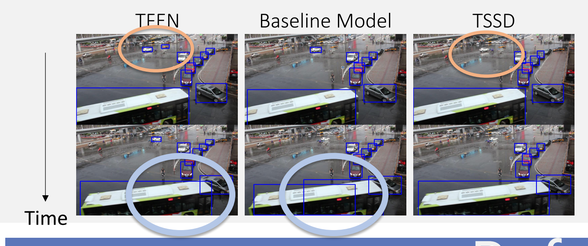
- Compared to SoTAs, the runtime is about **x3** faster with the comparable accuracy.
- Better performance in terms of accuracy, even no difference in runtime from TSSD

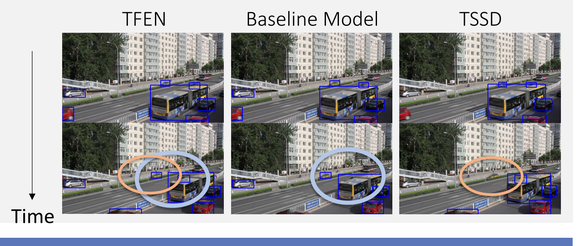
UA-DETRAC dataset AP[%] comparison (* is trained and evaluated based on official codes)

	Model	Overall	Easy	Medium	Hard	Cloudy	Night	Rainy	Sunny	FPS	GPU
Current frame only	**DPM[2010]**	25.70	34.42	30.29	17.62	24.78	30.91	25.55	31.77	0.16	-
	ACF[2014]	46.35	54.27	51.52	38.07	58.30	35.29	37.09	66.58	0.66	-
	RCNN[2014]	48.95	59.31	54.06	39.47	59.73	39.32	39.06	67.52	0.10	K40
	CompACT[2015]	53.23	64.84	58.70	43.16	63.23	46.37	44.21	71.16	0.22	K40
	Faster RCNN[2015]	58.45	82.75	63.05	44.25	62.34	66.29	45.16	69.85	11.0	Titan X
	GP-FRCNN[2017]	76.57	91.79	80.85	66.05	85.16	81.23	68.59	77.20	4.0	K40
	EB[2017]	67.96	89.65	73.12	54.64	72.42	73.93	53.40	83.73	11	Titan X
	MSVD_SPP[2019]	**85.29**	96.04	89.42	**76.55**	88.00	88.67	**78.90**	88.91	9.5	Titan Xp
	YOLOv3-SPP[2018]	84.96	95.59	**89.95**	75.34	**88.12**	**88.81**	77.46	89.46	6.5	Titan Xp
	FG-BR_Net[2019]	79.96	93.49	83.60	70.78	87.36	78.42	70.50	89.89	10	M40
Time series frames	**TSSD*[2018]**	57.16	81.06	62.07	43.14	57.59	63.87	44.98	67.73	**31.78**	2080 Ti
	TFEN(ours)	82.42	**97.40**	88.90	72.18	87.54	82.41	72.32	**90.78**	**29.11**	2080 Ti

AP performance [%] of ablation models on UA-DETRAC.

Model	Video	Temporal Attention Decoder(TAD)	Skip Connection (SK)	Spatial Attention (SA)	Temporally-aware Feature maps(TF)	Overall	Easy	Medium	Hard
Baseline	-	-	-	-	-	73.39	90.92	79.28	60.33
Model w/o TAD	✓	-	✓	✓	✓	79.26	95.96	85.83	67.42
Model w/o SK	✓	✓	-	✓	✓	72.53	91.26	78.57	59.24
Model w/o SA	✓	✓	✓	-	✓	80.93	97.17	86.08	66.44
Model w/o TF	✓	✓	✓	✓	-	79.22	95.06	84.77	65.46
(Complete) TFEN	✓	✓	✓	✓	✓	**82.42**	**97.40**	**88.90**	**72.18**





Conclusion

- Proposing the first temporal attention based external memory network for the live stream of video.
- Demonstrating the real-time performance with the comparable accuracy of SoTAs.

Reference

[1] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *CVPR*, 2017, pp. 6154–6162.
[2] S. Lyu, M.-C. Chang, D. Du, W. Li, Y. Wei, M. Del Coco, P. Carcagn`ı, A. Schumann, B. Munjal, D.-H. Choi *et al.*, "Ua-detrac 2018: Report of avss2018 & iwt4s challenge on advanced traffic monitoring," in *AVSS*. IEEE, 2018, pp. 1–6.
[3] J. Park, S. Woo, J.-Y. Lee, and I.-S. Kweon, "Bam: Bottleneck attention module," in *BMVC*, 2018.
[4] M. Sandler, A. G. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *CVPR*, 2018, pp. 4510–4520.