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

Masato Fujitake\textsuperscript{1}, Akihiro Sugimoto\textsuperscript{2}

\textsuperscript{1}Dept. of Informatics, The Graduate University for Advanced Studies, SOKENDAI
\textsuperscript{2}National Institute of Informatics Tokyo, Japan

**Motivation**

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

**Task**

- 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.

**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 $F^t_t$.
- Feature aggregation in External Memory to generate the feature map based on the attention weight.

**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**

- 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.

**Experiment Results**

**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

**Reference**