Multi-label activity recognition in extended videos using objects' spatio-temporal boundaries

“ITI-CERTH participation in ActEV and AVS Tracks of TRECVID 2021”

Konstantinos Gkountakos, Damianos Galanopoulos, Despoina Touska, Konstantinos Ioannidis, Stefanos Vrochidis, Vasileios Mezaris, Ioannis Kompatsiaris

Presenter: Despoina Touska
Problem statement

- Activity recognition and localization in surveillance scenarios
  - Processes untrimmed surveillance videos
    - Indoor or outdoor environments
    - Human, vehicles or both
  - Recognizes activity assigning a label
    - Human related
    - Vehicle related
    - Interaction between humans
    - Human-object interaction
  - Localizes activity’s spatio-temporal area
    - Time boundaries (start, end)
    - Spatial location
Surveillance scenarios challenges

- Untrimmed videos’ nature
- Camera’s large field of view
- Multiple activities simultaneously
- Multiple objects involved within each activity
- Actors perform more than one activity
  - At the same time
  - At overlapping time intervals
- Varying lengths of activities
Proposed approach

- Three-step pipeline:
  - **Detect** objects from **RGB video** frames
    - Extract bounding boxes for every object-of-interest (person, vehicles)
    - Track the detections over the time
    - Output spatio-temporal proposals of the detected objects
  - **Post-processing** the spatio-temporal proposals
    - Generate **Extended Activity Bounding Box** (EABBox) for every object
    - Construct final spatio-temporal activities proposals
  - **Classify** activities proposals
    - 3D-CNN model (3D-Resnet)
    - Assign labels to each activity proposal
Pipeline demonstration

Object detection - Tracking

Post - processing

Activity classification

Person walks

Person talks on phone
Object detection - YOLOv4

- **State-of-the-art real-time** object detector
- **43.5% AP** for MS COCO at 65 FPS (real-time) on Tesla V100
- **Pre-trained** using MS COCO dataset
  - Include objects such as “person”, “car”, “truck”
- **Fine-tuning** using the VIRAT dataset
  - 20 epochs
  - **Vehicle** and **person** the target objects
- Detected objects are described by:
  - **Bounding box**
  - **Confidence score**
- **Object tracker** based on **Euclidean distance**


Post-processing

- Extended Activity Bounding Box (EABBox) creation
- The union of the separated bounding boxes of each object
- Benefits:
  - Minimisation of the cropping effects avoiding a stretched and deform illustration of the objects
  - Acquisition of useful background information which could be helpful for activity classification
Activity classification - 3D-Resnet

- **Sample size**: (16, 112, 112) (frames, width, height)
- **Number of layers**: 50
- **Loaded weights**: Kinetics-400 dataset
- **Fine-tune** using the VIRAT dataset
- **Total epochs**: 350
- **Multi-label classification**
- **Weighted binary cross-entropy loss**
- **35 target** activities

Soft - Non maximum suppression

- Refines the classified activities proposals
- Improved version of the NMS algorithm
- Decays the detection scores of all objects as a continuous function of their overlap with other neighboring objects
- No object is eliminated in contrast with NMS
- Same computational complexity with NMS
- Implementation simplicity
Submitted systems

- **M4D_2021-baseline:**
  - Fine-tuned YOLOv4
  - Tracking with Euclidean distance
  - Post-processing
  - 3D-Resnet

- **M4D_2021-M4D_2021_S1:**
  - Fine-tuned YOLOv4
  - Tracking with Euclidean distance
  - Post-processing
  - 3D-Resnet
  - **Soft-NMS**
## Evaluation results

<table>
<thead>
<tr>
<th>System Name</th>
<th>*PARTIAL AUDC</th>
<th>MEAN-P <a href="mailto:MISS@0.15TFA">MISS@0.15TFA</a></th>
<th>MEAN-W_P MISS@ 0.15RFA</th>
</tr>
</thead>
<tbody>
<tr>
<td>M4D_2021-baseline</td>
<td>0.85484</td>
<td>0.79732</td>
<td>0.87719</td>
</tr>
<tr>
<td>M4D_2021-M4D_2021_S1</td>
<td>0.84658</td>
<td>0.79410</td>
<td>0.88521</td>
</tr>
</tbody>
</table>

*PARTIAL AUDC is the primary metric, the lower values the better results

- Slightly improvements in 2nd system
- Soft-NMS algorithm improves the results as it offers the possibility to eliminate duplicate activities which affect negatively the results
- Further improvement are observed for >0.2TFA
Experimental evaluation
Thank you

Despoina Touska
destousok@iti.gr

This work was partially supported by the European Commission under contracts H2020-786731 CONNEXIONS and H2020-833115 PREVISION