Event Detection in Videos Using a Graph Convolutional Network

Daiki Mukai[1], Kimiaki Shirahama[1], Takashi Matsubara[2] and Kuniaki Uehara[3]
1. Kindai University, 2. Osaka University, 3. Osaka Gakuin University

Motivation

An event is characterized by specific object appearances and their spatial-temporal relations

Need for capturing spatial-temporal relations among objects involved in an event

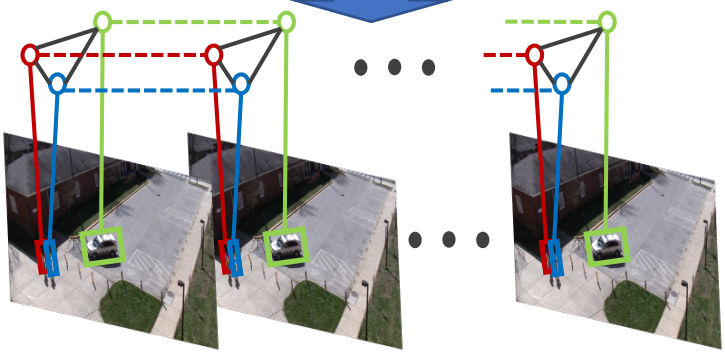
Use Spatial-Temporal Graph Convolutional Network (ST-GCN)

The visual feature of an object is abstracted into a high-level one by integrating features of spatially and temporally close objects.

Graph Creation

Segment of 20 frames sampled every 5 frames

- M2Det for object detection
- SE-ResNeXT-101 for feature extraction



- Spatial connections: Objects in the same frame are connected if they are close to each other
- Temporal connections: Two objects in consecutive frames are connected if their features are very similar

ST-GCN for Event Detection

Convolution needs the order of the data

$w = (w_1, w_2, \cdots, w_9)^T$

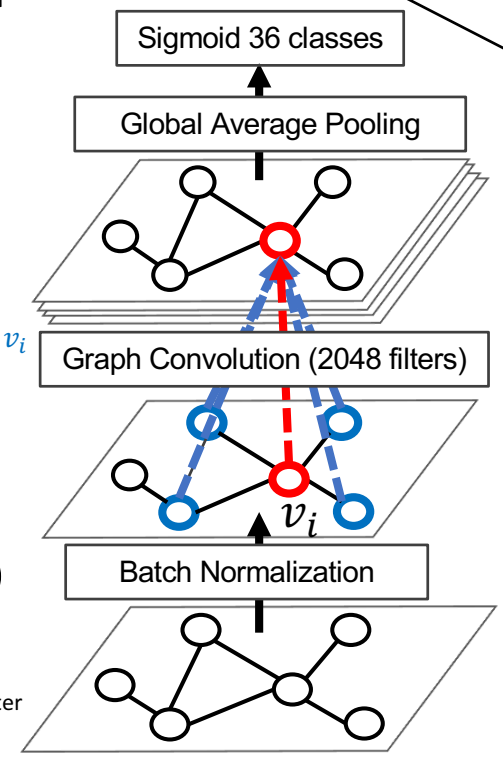
No order of nodes in a graph

Define the partitions for a node v_i

- $B_1 = \{v_i\}$
- $B_2 = \{v_j \mid d(v_i, v_j) = 1\}$

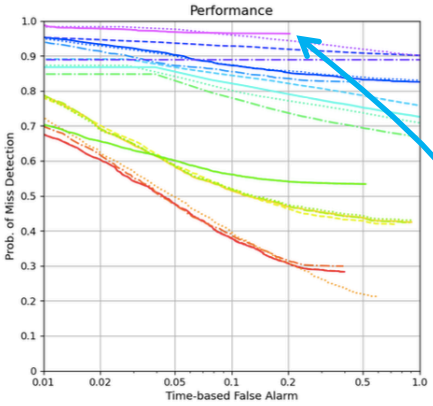
$$f_{out}(v_i) = f_{in}(v_i)w(p(v_i)) + \sum_{v_j \in B_2(v_i)} \frac{1}{|p(v_j)|} f_{in}(v_j)w\left(p(v_j)\right)$$

Abstract v_i's feature using a filter for itself and the another filter for neighboring nodes



Optimize filters so as to achieve accurate event detection

Result Overview

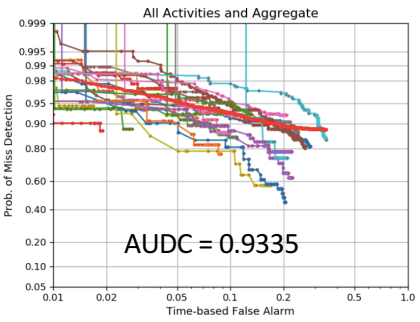


- 5,247 segments for training
- 21,489 segments for testing

The partial AUDC of our method is 0.9682

The main reason is to only output the event occurrence with the highest probability for a segment.

Discussion



Result of detecting multiple event occurrences for one segment

Many objects in a frame

Global average pooling mixes features of relevant and irrelevant objects to an event

Adopt an attention mechanism to selectively weight each object's features