Introduction

Relation understanding is crucial for holistic understanding of video data, which can support complicated applications like automatic captioning and multimedia question answering. In this work, we aim to detect visual relations between objects, including the dynamic relations and temporally changing relations. Formally, a visual relation instance in a video is represented by a relation triplet \(<subject, predicate, object>\) with the trajectories of the subject and object (see figure).

Object Tracklet Proposal

Relation Prediction

Definition of visual relation in videos

Evaluation Results

- Visual relation tagging evaluates the performance of relation prediction (no need to predict the object trajectories), which’s useful for retrieval and visual QA.
- The baselines are adapted from image based methods by using our tracklet proposal, feature extraction and relational association methods.
- Zero-shot setting requires to predict unseen visual relations w.r.t. training

VidVRD Dataset

- The first video visual relation dataset, based on ILSVRC16-VID dataset.
- Object trajectory annotation: additional 5 object categories, person, ball, sofa, skateboard and frisbee.
- Visual relation annotation: 3,219 types of relation triplets in which 258 only appears in test set. The test set has 432 unseen instances.
- Download link: https://lms.comp.nus.edu.sg/research/VidVRD.html
- Loading and evaluation codes: https://github.com/xdshang/VidVRD-helper