Spatio-Temporal Action Detection with Multi-Object Interaction

Huijuan Xu[1] Lizhi Yang[1] Stan Sclaroff[2] Kate Saenko[2] Trevor Darrell[1]

[1]University of California, Berkeley [2]Boston University
[1]{huijuan, lzyang, trevor}@eecs.berkeley.edu, [2]{sclaroff, saenko}@bu.edu

Abstract. Spatio-temporal action detection in videos requires localizing the action both spatially and temporally in the form of an "action tube." Most spatio-temporal action detection datasets are annotated with action tubes that contain a single person performing the action, thus the predominant action detection models simply employ a person detection and tracking pipeline for localization. However, when the action is defined as an interaction between multiple objects, such methods may fail since each bounding box in the action tube contains multiple objects. In this paper, we study spatio-temporal action detection problem with general multi-object interaction. We introduce a new dataset annotated with action tubes containing multi-object interactions, and propose an end-to-end spatio-temporal action detection model that performs both spatial and temporal regression simultaneously. Our spatial regression may enclose multiple objects participating in the action. During test time, we simply connect the regressed bounding boxes within the predicted temporal duration using a simple heuristic. We report the baseline results of our proposed model on this new dataset, and also show competitive results on the standard benchmark UCF101-24 with fast detection speed.

Keywords: Spatio-temporal action detection, Multi-object interaction

1 Introduction

Current methods for spatio-temporal action detection mostly focus on human-centric actions. The bounding boxes in corresponding datasets are annotated around the human subject. This raises the question of how actions should be defined. Should they be defined as the movement of the subject (e.g. person) or as the interaction of involved subjects and objects? Take the action "a person throws the frisbee" as a toy example, shown in Figure 1. Should the annotated action bounding box be focused on the person throwing the frisbee or include both the person and the frisbee? In this paper, we define an action as a subject-object interaction, and propose a spatio-temporal action detection dataset called S-STAR, where the bounding box annotation for the action tube includes the action subject (hand) and the objects involved in the action. We choose hand as action subject to alleviate the spatial occlusion with interacted objects.

Existing methods use the object detection and linking pipeline and are not well-suited to detecting multi-object interaction, since they rely on object detection models for spatial localization. However, reliable detectors may not be

Fig. 1. Toy example for annotating the action "a person throws the frisbee". Existing datasets annotate the action around the person subject (left), while we take the action definition of subject and object interaction and annotate the action enclosing the involved subjects and objects (right).

available for all objects, considering that objects in the real world are quite diverse and have a long-tailed distribution. In this paper, we propose a simple, elegant and effective multi-object action detection model with simultaneous spatial and temporal localization called the *Spatio-TemporAl Regression (STAR)* model. Our model regresses the spatial action bounding box containing multi-object interaction without relying on external object detectors. It also contains a temporal localization branch to regress the temporal duration in parallel with the spatial regression during training time. At test time, our STAR model takes a top-down approach by first selecting a generated temporal segment proposal and then connecting the regressed spatial action boxes within the temporal duration using a simple Intersection over Union (IoU) heuristic, forgoing the extra time-consuming temporal optimization.

2 Spatio-TemporAl Regression (STAR) Model

Our STAR model in Figure 2 takes a sequence of RGB video frames $I_t; t = 1:::T$ with dimension $\mathbb{R}^{3\times T\times H\times W}$. Then, a 3D ConvNet [6] is used to extract rich spatio-temporal feature hierarchies with the output feature map $C_{conv5b} \in \mathbb{R}^{D\times\frac{T}{8}\times\frac{H}{16}\times\frac{W}{16}}$, where D is the channel dimension. We use C_{conv5b} feature activations as the shared input to the temporal localization and action classification branch and the spatial localization branch.

Our spatial localization branch conducts the action bounding box regression at each frame and outputs binary actionness score for each regressed box. We use a region proposal network as our spatial localization branch. The number of encoding feature maps $\frac{T}{8}$ is treated as batch size dimension during the training of spatial localization, and each feature map conducts spatial localization independently. The ground truth action tube bounding boxes are mapped to one of the $\frac{T}{8}$ feature maps as spatial supervision according to nearest neighbour policy. For a certain feature map, if no ground truth action tube is temporally across that feature map, it will not contribute to the spatial localization training.

The video encoding feature map is spatially pooled and on top of which, class-agnostic temporal proposals are predicted with respect to anchor segments, and high quality temporal proposals are refined and classified into specific action classes. Our spatial regression branch only outputs a class-agnostic actionness score; the final action classification is achieved in the temporal branch, and the

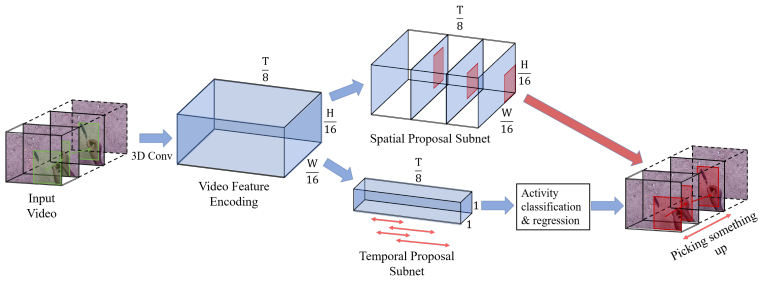

Fig. 2. STAR model architecture.

features used for final action classification come from 3D segment of interest (SoI) pooling for selected temporal proposals. We train the network by optimizing both the classification and regression tasks jointly in the two branches of our STAR model. The softmax/binary classification loss function is used for classification, and smooth L1 loss function is used for regression.

3 Experiments

In this paper, we explore the spatio-temporal action detection in videos with multi-object interaction. To tackle this problem, we introduce a new dataset called S-STAR. The videos used in our S-STAR dataset comes from the "something-something" dataset [1]. We balance the number of videos in each class and set the number of videos in each class to be 300. This results in a total of 47 classes and 14100 videos in our proposed "S-STAR" dataset. We annotate the action tube in each video with spatial bounding boxes containing the subject hand as well as interacted objects. The average temporal duration ratio of the action tubes in our S-STAR dataset is 55.6%. We report STAR model results on our annotated S-STAR dataset with multi-object interaction, as well as the widely used spatio-temporal action detection benchmark UCF101-24 [5]. The spatio-temporal action detection results are evaluated in terms of video mean Average Precision - video mAP@ , frame mAP, as well as temporal mAP.

S-STAR Dataset We split our S-STAR dataset into 80% for training and 20% for testing. We run the first baseline of our STAR model starting from Sports-1M pretrained C3D classification weights released in [6] and get the video mAP@0.5 of 16.65%, shown in Table 1. We also finetune C3D model on the something-something classification data and get 25.6% top-1 classification accuracy and 50.6% top-5 classification accuracy. Starting from the something-something finetuned C3D classification backbone, our STAR model's video mAP@0.5 result further improves by 5.77 points and arrives at 22.42%. This indicates that our model can benefit from better pretrained video classification backbone during end-to-end training. Figure 3 shows one representative qualitative result.

UCF101-24 Dataset In Table 2, our STAR model got video mAP@0.5 of 53.0% on UCF101-24 dataset, which is very competitive when comparing to other sate-of-the-art models using more advanced two-stream I3D video classification backbone, showing the generalizability of our proposed STAR model.