**Multi-region two-stream R-CNN for action detection**

**Introduction**

**Motivation:**
- Previous work shows improvement with better proposal methods [1].
- State-of-the-art CNN-based action classification relies on multi-frame optical flow [2].
- Object recognition is improved by multi-region features [3].

**Contribution:**
- We introduce a multi-region Region Proposal Network (RPN).
- We introduce optical flow to significantly improve action detection.
- We embed a multi-region scheme in the Faster R-CNN model.

**Two-stream R-CNN for action detection**

Pipeline of our Two-stream Faster R-CNN (TS R-CNN)

Input: a single RGB frame and multi-frame optical flow

**Training:**
- RGB/appearance (A) and optical flow/motion (M) are trained separately with a VGG16 model pretrained on ImageNet.
- When training with multi-frame optical flow, we duplicate the VGG16 filters of the first layer multiple times and use the bounding boxes of the middle frame as ground truth.

**Testing:**
- Combine A & M models by developing a RoI fusion layer.
- Average the softmax scores from A & M streams.
- The bounding box regressors are applied to corresponding RoIs of each stream (the red and blue solid bars).

**Multi-region two-stream faster R-CNN**

Architecture of our multi-region TS R-CNN (MR-TS R-CNN)

- Introduce a multi-region generation layer next to the RPNs.
- Introduce a mask-supported RoI pooling layer for border regions.

**Additional regions:** Upper half, bottom half, border.

**Training:** For both A & M streams, we fine-tune the network (only fully connected layer) of the original regions separately for each region.

**Linking and temporal localization**

Given two regions, \( R_t \) and \( R_{t+1} \) from consecutive frames, the linking score for class \( c \) is:

\[
L_c(R_t, R_{t+1}) = \frac{1}{k_{t+1}} \sum_{k=t}^{t+1} s_c(R_k) = \lambda \frac{|I_t \cap I_{t+1}|}{|I_t|} + \beta \omega(A_t, A_{t+1})
\]

where \( \omega(A_t, A_{t+1}) \) is the IOU overlap, \( \beta \) is a scalar, \( \omega(A) \) is a threshold function.

**Action proposal evaluation**

Comparison of different frame-level proposals on UCF-Sports (left) and JHMDB (right).

- Multi-frame optical flow RPN is better than single-frame optical flow.
- Best results are obtained by combining all the proposals of flow-5 and appearance RPNs.

**Experimental results**

Frame-mAP evaluation of TS R-CNN with varying number of frames and combination strategies on UCF-Sports (left) and JHMDB split 1 (right).

- Multi-frame flow improves action detection and saturates above 5 frames.
- Increasing the number of RGB frames does not improve the performance.
- Our score fusion with a RoI fusion layer is better than a naive NMS method.

- Some actions are better represented by individual part R-CNN models.
- Multi-region TS R-CNN model improves the Orig R-CNN model.

**Reference**


