What if we do not have multiple videos of the same action? –
Video Action Localization Using Web Images

Wagas Sultani  
Mubarak Shah

**Motivation:**
- Bounding box annotations are cumbersome, require hundreds of hours and subject to human biases.

**Key Contributions:**
- Demonstrate the feasibility of images to achieve spatio-temporal action localization
- Action localization using video proposal sparse reconstruction with motion saliency
- First to report spatio-temporal action localization results on (the part of) untrimmed THMOS'14 dataset.
- Spatial Annotations of more than 32,000 frames.

**Key IDEA:**
Web images related to key poses in video frames can localize an action.

**Challenges:**
- Large number of candidate action locations in videos
- Noisy Web Images
- Distracting Image Backgrounds
- How to use images to localize action in videos?
- Lack of spatiotemporal annotations for Un-Trimmed Dataset

**Proposed Pipeline:**
- Video clip
- Action proposals in each clip
- Top-Ranked Action Proposal
- Rank video proposals using image proposals
- Outliers Removal
- Action proposals in each image

**Method Details:**
- **Noisy Image Removal:** Random walk
  \[ p(i,j) = \frac{e^{-\gamma||v(i) - v(j)||}}{\sum_{k} e^{-\gamma||v(i) - v(k)||}} \]
  \[ r(v,j) = \beta \sum_{i} r(v,i)(1 - \beta) + \beta v_j \]

- **Image Proposals**

- **Video Proposals**

- **Video Proposal Reconstruction using Image Proposals:**
  \[ \min \alpha \left\| \mathbf{C} - \mathbf{C}_1 \right\|_2^2 + \lambda_1 \left\| \mathbf{C} \right\|_2^2 + \lambda_2 \left\| \mathbf{C} \right\|_1 \]

**Experiments:**
- **Trimmed Action Datasets:**
  - UCF Sports

**Un-Trimmed Action Dataset:**
- THUMOS'14, 4 Actions: Baseball Pitch, Golf Swing, Tennis Swing, Throw Discus.

**Overall Results:**

**References:**