Weakly Labeled Action Recognition and Detection

Waqas Sultani
Advisor: Dr. Mubarak Shah
Motivation

- Humans are social species and have the ability to recognize actions.

- Humans can detect actions irrespective of who is performing actions as well as recognize actors irrespective of actions they are performing.

- In computer vision, we want to develop the same capability of action recognition for computers.
Applications
Action Recognition

Recognizing action in video,
Without estimating location of the actor

Diving
Action Detection

Recognizing action in video,
As well as Estimate location of the actor
Action Localization

Localizing Known Action in the video.

Throw Discus
Weakly Labeled Action Recognition

Estimate discriminative and foreground regions to facilitate action recognition using video level labels.
Weakly Labeled Action Detection

Train action detector using only video level labels.
Outline

- Human Action Recognition across Action Datasets (CVPR 2014)
- Automatic Action Annotations using Weakly Labeled Videos (CVIU 2017)
- Video Action Localization Using Web Images (CVPR 2016)
- Unsupervised Action Proposal Ranking (CVIU 2017)
Human Action Recognition across Action Datasets

CVPR 2014
Motivation

Recognition Accuracy drops across the datasets!

Cross dataset recognition

<table>
<thead>
<tr>
<th>Training</th>
<th>Testing</th>
<th>Accuracy (avg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>UCF50</td>
<td>UCF50</td>
<td>70 %</td>
</tr>
<tr>
<td>UCF50</td>
<td>HMDB51</td>
<td>55.7 %</td>
</tr>
<tr>
<td>Olympic Sports</td>
<td>Olympic Sports</td>
<td>71.8 %</td>
</tr>
<tr>
<td>Olympic Sports</td>
<td>UCF50</td>
<td>16.67 %</td>
</tr>
</tbody>
</table>

★ Training and Testing is done on similar actions across the datasets
Analysis

- Action classifiers learns backgrounds.

- Datasets have discriminative backgrounds.

Experiments

- Background Motion Features.

- Global Scene descriptors.
Experiment

Action recognition using Background Features

- Two recent popular datasets:
  - UCF YouTube
  - UCF Sports

- Extract STIP (HOG, HOF)
  - 50% Spatial-Temporal overlaps.
  - Foreground Features: 50% overlap with bounding box
  - Background Features: Less than 50% overlap with bounding box.
Action recognition using Background Features

<table>
<thead>
<tr>
<th>STIP Features</th>
<th>UCF YouTube</th>
<th>UCF Sports</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foreground</td>
<td>59.80 %</td>
<td>71.92 %</td>
</tr>
</tbody>
</table>
Action recognition using Background Features

<table>
<thead>
<tr>
<th>STIP Features</th>
<th>UCF YouTube</th>
<th>UCF Sports</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foreground</td>
<td>59.80 %</td>
<td>71.92 %</td>
</tr>
<tr>
<td>Background</td>
<td>55.27 %</td>
<td>73.97 %</td>
</tr>
</tbody>
</table>
Our Approach
Foregroudfocused Representation
Foreground Focused Representation

- Action localization
- Binary foreground/Background Segmentation
- Very challenging and difficult, akin to introducing a new problem to solve the first.

Instead
- Estimate the confidence in each pixel being a part of the foreground, and use it to obtain video representation
Obtain score of each pixel being the foreground using:

- Motion Gradients
- Color Gradients
- Saliency
- Spatial-Temporal Coherence using 3D-MRF

Final weights

Input: Video
Foreground Focused Representation

- **Motion Gradients**

\[
f_m(x, y) = \left\| \begin{bmatrix} u_x & u_y \\ v_x & v_y \end{bmatrix} \right\|_F * g
\]

\[
= \sqrt{u_x^2 + u_y^2 + v_x^2 + v_y^2} * g,
\]

- **Color Gradients**

\[
f_c(x, y) = \sqrt{L_x^2 + L_y^2 + a_x^2 + a_y^2 + b_x^2 + b_y^2} * g,
\]
Foreground Focused Representation

- Visual Saliency
  - Actor receives most of attention and hence represent salient part of the video
- Frames based Saliency
Coherence of Foreground Confidence using 3DMRF

- Initial aggregate of confidence map
  \[ \hat{f}_a = \log (f_m (f_c + f_s) + 1) \]
- The score is max-normalized for each frame of a video
- The quality of labeling is given by:

\[
E(\omega) = \sum_{\psi_p \in V} D_p(\omega_p) + \sum_{(p,q) \in V} V(\omega_p - \omega_q).
\]

\[
D_p(\omega_p) = (\hat{f}_a(p) - \omega_p)^2
\]

\[
V(\omega_p - \omega_q) = \min((\omega_p - \omega_q)^2, \kappa)
\]
Coherence of Foreground Confidence using 3DMRF

• Inference

  • The message the node $p$ send to $q$ is given by

    $$\min_{\omega_p} \left( D_p(\omega_p) + V(\omega_p - \omega_q) + \sum_{s \in N_p \setminus q} m_{s \rightarrow p}^{t-1}(\omega_p) \right)$$

  • The belief vector of node $q$ is given by

    $$b_q^t(\omega_q) = D_q(\omega_q) + \sum_{s \in N_q} m_{s \rightarrow q}^t(\omega_q)$$
Traditional Bag of words

- Codebook
- Histogram
- Similarity measure, i.e. Kernel
Traditional Bag of words

Biking

Histogram

Foreground words  Background words
Weighted Codebook

Biased Codebook towards foreground features

- The confidence of each descriptors as being on foreground in given by:

\[ w_i = \sum_{(x,y) \in P_i} f_a(x,y) / |P_i| \]

- The goal of clustering is to minimize the following energy function:

\[ \arg\min_C \sum_{j=1}^{K} C(i,j)w_i \| x_i - z_j \|^2 \]

\( C \) is the \(|X| \times k\) unknown membership matrix
**Weighted Histogram**

To reduce the contribution of background features, use weights for each features of being foreground during quantization.
Traditional bag of words

Histogram

kmeans

Foreground words  Background words

Biking
Weighted bag of words

Weighted Histogram

Weighted-kmeans

Foreground words

Background words

Biking
Weighted bag of words

Biking

Weighted Histogram

Weighted-kmeans

Foreground words  Background words
Weighted bag of words (Limitations)

Weighted Histogram

Weighted-kmeans

Foreground words  Background words
Weighted bag of words (Limitations)

Weighted Histogram

Weighted-kmeans

Foreground words
Background words
Foreground confidence based Histogram decomposition

- Categorize spatiotemporal regions corresponding to different weights in ‘R’ classes

Features partitions based on weights
Foreground confidence based Histogram decomposition

- Categorize spatiotemporal regions corresponding to different weights in ‘R’ classes
- Compute Histograms for each region separately

Features partitions based on weights
UCF50 Video

Weighted Histograms

HMDB51 Video

Weights Partitions

Weighted Histograms

Weights Partitions

Final Similarity

Weighted Summations

Histogram Intersection
Experimental Results

- Datasets used:
  - UCF50, HMDB51, Olympic Sports

- Features used:
  - STIP

- UCF50 Vs. HMDB51
  - Common actions
    - We choose actions which are visually similar:
      - Biking, Golf Swing, Pull Ups, Horse Riding, Basketball

- UCF50 Vs. Olympic Sport
  - Common actions:
    - Basketball, Pole Vault, Tennis serve, Diving, Clean and Jerk, Throw Discus
Quantitative Results

Pull Ups

Histogram Intersection= 0.2744

Weighted Histogram Intersection=0.5454

Weighted Histogram Decomposition =0.5586
Quantitative Results

Golf Swing

Histogram Intersection= 0.1684

Weighted Histogram Intersection=0.2740

Weighted Histogram Decomposition =0.3089
Quantitative Results

Biking

UCF 50  HMDB51

Histogram Intersection = 0.1035

Weighted Histogram Intersection = 0.1142

Weighted Histogram Decomposition = 0.1295
# Quantitative Results

<table>
<thead>
<tr>
<th>Training</th>
<th>Testing</th>
<th>Un-Weighted</th>
<th>Weighted</th>
<th>Histogram Decomposition</th>
</tr>
</thead>
<tbody>
<tr>
<td>UCF50</td>
<td>HMDB51</td>
<td>55.70</td>
<td>60.00</td>
<td>68.70</td>
</tr>
<tr>
<td>HMDB51</td>
<td>UCF50</td>
<td>63.3</td>
<td>64.00</td>
<td>68.67</td>
</tr>
<tr>
<td>Olympic Sports</td>
<td>UCF50</td>
<td>16.67</td>
<td>32.29</td>
<td>47.91</td>
</tr>
</tbody>
</table>
Summary

• Cross-dataset action recognition without using labels or features from the test set

• We have experimentally demonstrated the detrimental effect of background scenes on action recognition dataset

• Proposed new process of improving bag of words for action recognition
Limitations

- Give rough estimate of location of actor
- Action detector provide precise actor location
- Action detection needs thousands of annotations.
  - Deep learning
  - Large action datasets such as THUMOS14 dataset

Manual Annotations is prohibitive!
Automatic Action Annotations using Weakly Labeled Videos

Journal of Computer Vision and Image Understanding (CVIU), 2017
Spatiotemporal Bounding box annotation of action in a video is difficult

Requires:

- Many human annotators
- Hundred of hours
- Expensive annotation interfaces
- Prone to error
Video level annotation of action in a video is easy to obtain

Requires:

• Few human annotators
• Less time
• Simple annotation interfaces
• Less prone to error
Goal

- Automatic Spatio-Temporal Annotations using video level labels only.

- Action Detection using automatic annotations
Proposed Approach
1. Action Proposals

- Candidate Action locations
- Employ unsupervised hierarchal clustering of Improved Dense Trajectory features.
**Action Proposals**

**Foreground Score**

- Motion Score
- Visual Saliency
- Spatio-temporal coherence using 3D-MRF:
Action Proposals

Foreground Score
Action Proposals

Proposal Subset Selection

• Mean of foreground score within proposal as its initial proposal score.
• Remove high overlapping proposals using subset selection method.
2. Exploiting Multiple Videos

Action representative proposal should be:

• Salient (High foreground score in the video).

• Have high similarity with salient proposals in other videos.
Similarity measures between Proposals

- Global Similarity
- Fine Grained Similarity
- Shape Similarity
Global Similarity

• Improved dense Trajectory features histograms in all proposals

• Pairwise similarity between proposals is given by:

\[ S_{ij}^f = \exp \left( -\gamma \sum_{k=1}^{k=d} \frac{(h_{ik} - h_{jk})^2}{(h_{ik} + h_{jk})} \right), \]
**Fine-Grained Similarity**

- Cluster raw improved dense trajectory features in each proposal.
- Find one to one correspondence between clusters using Hungarian algorithm.
- Final similarity between proposals is the aggregate of their local (clusters) similarities
Proposal Shape Similarity

- Shape of proposal windows (aspect ratio) over time itself carries useful information about an action.

\[
    r_i = \frac{w_i}{h_i}
\]

\[
    \Lambda_{px} = [r_1, r_2, \ldots, r_n],
\]

\[
    \Lambda_{py} = [r_1, r_2, \ldots, r_m],
\]

- Find similarity between two series using Dynamic Time Warping
Find the best action representative proposal using Generalized Maximum Clique Graphs (GMCP)
GMCP Matching
GMCP Matching
GMCP Matching
GMCP Matching

- Find the maximum clique, the most action representative proposal in each video using Local Neighborhood Search.
**GMCP Graph**

- Each proposal (node) in video is connected to all proposals in other videos
- No two proposals within the same are connected

\[
\Omega_{p_i} = \text{Initial Action score} \\
\Pi_{ij} = \text{Shape similarity} \\
\Gamma_{ij} = \text{Fine-grained Similarity} \\
\Theta_{ij} = \text{Global Similarity}
\]

\[
\sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} \left( \alpha \Omega_{p_i} + (\Theta_{ij} + \Gamma_{ij} + \Pi_{ij}) \eta_{p_i} \times \eta_{p_j} \right)
\]
Experimental Results

Datasets

- UCF-Sports
- sub-JHMDB
- THUMOS’13

Experiments

- Initial Proposal Ranking
- Annotation Quality
- Action Detection
Qualitative Results
Sub-JHMDB Dataset

- Catch
- Golf Swing
- Pick
- Climb
- Jump
- Pull Ups

Proposed

Ground Truth
Quantitative Results
# Initial Proposal Ranking

## Mean Average Best Overlap (MABO)

<table>
<thead>
<tr>
<th>Method</th>
<th>UCF-Sports</th>
<th>Sub-JHMDB</th>
<th>THUMOS13</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non Maximal Suppression</td>
<td>51.98</td>
<td>50.74</td>
<td>33.60</td>
</tr>
<tr>
<td><strong>Ours (subset Selection)</strong></td>
<td><strong>56.01</strong></td>
<td><strong>55.25</strong></td>
<td><strong>35.26</strong></td>
</tr>
<tr>
<td>All Proposals</td>
<td>62.40</td>
<td>57.77</td>
<td>46.71</td>
</tr>
</tbody>
</table>
## Annotation Quality

### Quantitative Results

<table>
<thead>
<tr>
<th>Method</th>
<th>UCF-Sports</th>
<th>Sub-JHMDB</th>
<th>THUMOS13</th>
</tr>
</thead>
<tbody>
<tr>
<td>Co-segmentation[95]</td>
<td>53.15(23.3)</td>
<td>49.37(24.5)</td>
<td>9.56(5.48)</td>
</tr>
<tr>
<td>Co-segmentation[17]</td>
<td>42.25 (19.2)</td>
<td>47.78 (20.3)</td>
<td>21.41(12.4)</td>
</tr>
<tr>
<td>Negative Mining[59]</td>
<td>48.49(21.8)</td>
<td>61.39(22.3)</td>
<td>14.39(10.2)</td>
</tr>
<tr>
<td>CRANE [69]</td>
<td>61.18(27.9)</td>
<td>64.56(22.8)</td>
<td>14.17(10.0)</td>
</tr>
<tr>
<td><strong>Ours</strong></td>
<td><strong>83.83(34.6)</strong></td>
<td><strong>89.56(32.4)</strong></td>
<td><strong>41.69(19.1)</strong></td>
</tr>
</tbody>
</table>
# Annotation Quality

## Components Contribution

<table>
<thead>
<tr>
<th>Action</th>
<th>[I]nitial Score</th>
<th>[I]+[S]hape</th>
<th>[I]+[S]+[G]lobal</th>
<th>[I]+[S]+[G]+Fine-grained</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diving</td>
<td>64.2</td>
<td>64.2</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Golf Swing</td>
<td>5.50</td>
<td>55.5</td>
<td>66.6</td>
<td><strong>88.8</strong></td>
</tr>
<tr>
<td>Kicking</td>
<td>20.0</td>
<td>50.0</td>
<td>65.0</td>
<td>65.0</td>
</tr>
<tr>
<td>Lifting</td>
<td>0.0</td>
<td>0.0</td>
<td>33.0</td>
<td><strong>83.3</strong></td>
</tr>
<tr>
<td>Riding Horse</td>
<td>33.3</td>
<td>25.0</td>
<td>25.0</td>
<td><strong>100</strong></td>
</tr>
<tr>
<td>Run</td>
<td>20.0</td>
<td>20.0</td>
<td>40.0</td>
<td><strong>80.0</strong></td>
</tr>
<tr>
<td>Skateboarding</td>
<td>50.0</td>
<td>41.6</td>
<td>41.6</td>
<td><strong>91.6</strong></td>
</tr>
<tr>
<td>Swing Bench</td>
<td>10.0</td>
<td>15.0</td>
<td>25.0</td>
<td><strong>90.0</strong></td>
</tr>
<tr>
<td>Swing SideAngle</td>
<td>38.4</td>
<td>38.4</td>
<td>61.5</td>
<td><strong>76.9</strong></td>
</tr>
<tr>
<td>Walk</td>
<td>16.6</td>
<td>58.3</td>
<td>25.0</td>
<td><strong>62.5</strong></td>
</tr>
<tr>
<td><strong>Avg</strong></td>
<td><strong>25.8</strong></td>
<td><strong>36.8</strong></td>
<td><strong>48.3</strong></td>
<td><strong>83.8</strong></td>
</tr>
</tbody>
</table>
Action Detection

sub-JHMDB

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Area Under ROC Curve (AUC)</td>
<td>37.6</td>
<td>37.4</td>
<td>40.3</td>
</tr>
<tr>
<td>mean Average Precision (mAP)</td>
<td>28.0</td>
<td>28.6</td>
<td>32.4</td>
</tr>
</tbody>
</table>
Action Detection

THUMOS13

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Area Under ROC Curve (AUC)</td>
<td>33.0</td>
<td>33.1</td>
<td>35.4</td>
</tr>
<tr>
<td>mean Average Precision (mAP)</td>
<td>5.4</td>
<td>5.3</td>
<td>6.9</td>
</tr>
</tbody>
</table>
Action Detection

UCF-Sports

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Area Under ROC Curve (AUC)</td>
<td>41.53</td>
<td>32.21</td>
<td>41.59</td>
</tr>
<tr>
<td>mean Average Precision (mAP)</td>
<td>23.54</td>
<td>15.0</td>
<td>25.27</td>
</tr>
</tbody>
</table>
Limitation

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

CVPR 2016
Action Localization

Localizing Known Action in the video.

Throw Discus
Action images capture key poses
Action images capture key poses

Swing Side-Angle
Web Images
Key idea

Locations containing key poses in video frames can localize an action.
Video Clip + Action Proposals → Action localization

Rank Video proposals using Image proposals

Web Images → Remove Noisy Images → Action Proposal in Images
1. Action Proposals
2. Remove Noisy Web images

- Noisy Images removal using Random Walk
3. Image Action Proposals

Action Images
• **Unsupervised Image Action Proposals**

  • Extract region proposals for all images.

  • Find Similar images using nearest neighbor on GIST features

  • Find common region across multiple images
• **Unsupervised Image Action Proposals**

Action Images
4. Rank Video Proposals using Image Action Proposals

- Construct each video proposals as linear combination of image proposals using sparse representation.

- Consider a video $v$, containing $k$ number of action proposals.

$$P_v = [p_v^1, p_v^2, \ldots, p_v^k]$$

- Compute Deep features from each key bounding boxes of each proposal.

$$\Pi^f \in \mathbb{R}^{d \times n}$$

- Compute Deep features from images proposals in all images

$$\Upsilon^f \in \mathbb{R}^{d \times m}$$
4. Rank Video Proposals using Image Action Proposals

\[
\min_C \| \Pi^f - \Upsilon^f C \|^2_F + \lambda_2 \| C \|_1
\]

\[
\text{Sparsity}
\]

\[
\min_C \| \Pi^f - \Upsilon^f C \|^2_F + \lambda_2 \| C \|_1 + \lambda_1 \| C - \bar{C} \|^2_F
\]

\[
\text{Consistency}
\]

\[
R_p = \text{Proposals reconstruction score}
\]
5. Video Proposals that has low reconstruction error and is motion salient represents action location

\[ \mathcal{R}_p = \text{Proposals reconstruction score} \]

\[ \eta_p = \text{Motion Score} \]

\[ \Lambda_p = (1 - \mathcal{R}_p) + \eta_p \]

Reconstruction score  Motion score
Experimental Results

Trimmed Datasets

- UCF-Sports
- THUMOS13

Un-Trimmed Datasets

- THUMOS14 (4 Actions)
Trimmed Action Datasets
Running

Skateboarding

Horse Riding

Proposed

Ground Truth

UCF-Sports
Walking

Swinging Side

Lifting

Proposed

Ground Truth

UCF-Sports
Swinging

UCF-Sports
## Quantitative Results

### UCF Sports

<table>
<thead>
<tr>
<th>Method</th>
<th>CRANE</th>
<th>Negative Mining</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Localization Accuracy</td>
<td>65.41[1]</td>
<td>63.01[2]</td>
<td>92.72</td>
</tr>
</tbody>
</table>


# Components Contribution

## UCF Sports

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Images (without noisy images removal)</td>
<td>65.25</td>
</tr>
<tr>
<td>Images (After noisy images removal)</td>
<td>68.16</td>
</tr>
<tr>
<td>Images Proposals (without regularization constraints)</td>
<td>83.70</td>
</tr>
<tr>
<td>Reconstruction</td>
<td>88.56</td>
</tr>
<tr>
<td>Reconstruction + motion</td>
<td>92.72</td>
</tr>
</tbody>
</table>
Untrimmed Action Dataset – THUMOS14
Summary

- Web Images for Action localization in videos
- Video proposal reconstruction using sparse representation
- Annotations of 35,000 frames from THUMOS14
Limitations

- Accuracy of action detection and annotation is upper bounded by quality of action proposals

- Video level labels are needed to discover good action proposals
Unsupervised Action Proposal Ranking through Proposal Recombination

Journal of Computer Vision and Image Understanding (CVIU), 2017
**Proposed Approach**

Proposal 1

Proposal 2

Proposal 3

Bounding Box
**Proposed Approach**

- **Proposal 1**
- **Proposal 2**
- **Proposal 3**

**Sub-proposals**

**Unary Score:** Actionness + Motion

**Binary Score:** Appearance + Overlap

**Selected Paths**
Graph Formulation

- Node Score
  - Actionness
  - Motion Score

- Edge Score
  - Shape Similarities
  - Appearance Similarities
Node Score

Actionness
Node Score

Motion Score
**Edge Score**

**Shape Similarity**

\[ \Psi^o = \frac{\text{Area}(b_{i,1} \cap b_{j,1})}{\text{Area}(b_{i,1} \cup b_{j,1})} \]

**Appearance Similarity**

- Euclidean distance between Mean HOG
Graph Formulation

Score of any path in the graph

\[ E(P) = \sum_{f=1}^{F} (\Phi_{p_f}^f + \lambda \cdot \Psi_{(p_f, p_{f+1})}^f) \]

Maximum Scored path through Dynamic Programming

\[ P^* = \arg \max_P E(P) \]
Experimental Results
# Proposal Ranking using Different Action Proposal

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Proposal</th>
<th>Methods</th>
<th>$t=0.1$</th>
<th>$t=0.2$</th>
<th>$t=0.3$</th>
<th>$t=0.4$</th>
<th>$t=0.5$</th>
<th>$t=0.6$</th>
<th>MABO</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Proposal</th>
<th>Methods</th>
<th>$t=0.1$</th>
<th>$t=0.2$</th>
<th>$t=0.3$</th>
<th>$t=0.4$</th>
<th>$t=0.5$</th>
<th>$t=0.6$</th>
<th>MABO</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Proposal</th>
<th>Methods</th>
<th>$t=0.1$</th>
<th>$t=0.2$</th>
<th>$t=0.3$</th>
<th>$t=0.4$</th>
<th>$t=0.5$</th>
<th>$t=0.6$</th>
<th>MABO</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Proposal</th>
<th>Methods</th>
<th>$t=0.1$</th>
<th>$t=0.2$</th>
<th>$t=0.3$</th>
<th>$t=0.4$</th>
<th>$t=0.5$</th>
<th>$t=0.6$</th>
<th>MABO</th>
</tr>
</thead>
</table>
Components Contribution

UCF Sports

<table>
<thead>
<tr>
<th>Components</th>
<th>MABO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Motion Score</td>
<td>40.4</td>
</tr>
<tr>
<td>Actionness</td>
<td>28.9</td>
</tr>
<tr>
<td>Actionness + Motion Score</td>
<td>45.5</td>
</tr>
</tbody>
</table>
Proposal Ranking
MSR-II
Action Detection

UCF Sports

THUMOS13
Conclusion

- Action datasets have discriminative backgrounds, which may affect the generalizability of action classifiers
  - A new process for obtaining per pixel confidence
  - Demonstrated the use of foreground focused action representation in cross-dataset action recognition

- Automatic annotation of action by exploiting similarity across multiple videos.
  - Manual spatio-temporal annotation is laborious.
  - Shape and feature similarities across multiple videos to capture action localization in each video
Conclusion

• Action Localization using Web Images
  • Reconstruction error as a measure of similarity between web images and video frames.
  • Experiments on Un-trimmed videos.

• Unsupervised Action Proposal Ranking
  • Unsupervised method for improving action proposals.
  • New proposals discovery through proposal recombination
  • Improved action detection results.
Future work

• Foreground Focused Representation
  • Supervised Motion and Saliency estimation
  • Fusion with deep learning

• Automatic Action Annotation
  • Deep Representation
  • Fast optimization Method
Future work

- Action localization using web images
  - Using images and videos jointly
  - Temporal Boundaries of Action

- Unsupervised Proposals Ranking
  - Use more deeper network

- Action Interaction and Online
Publications and Patents

- **Automatic Action Localization in Weakly Labeled Videos.** Waqas Sultani and Mubarak Shah, In *CVIU 2017*

- **Unsupervised Action Proposal Ranking through Proposal Recombination.** Waqas Sultani, Dong Zhang and Mubarak Shah, In *CVIU 2017*

- **Automatic Pavement Object using Superpixel Segmentation Combined with Conditional Random Field.** Waqas Sultani, Sooush Mukhtari and Hae-Bum Yun, In IEEE Transaction on Intelligent Transportation System (ITS) *2017*

- **What if we do not have multiple videos of the same action? Video Action localization using Web Images.** Waqas Sultani and Mubarak Shah, In *CVPR 2016*

- **Human action recognition across dataset using foreground focused Histogram decomposition.** Waqas Sultani and Imran Saleemi, In *CVPR 2014*


- **Abnormal Traffic Detection using Intelligent Driver Model,** Waqas Sultani, Jin Young Choi, *ICPR, 2010.*
Thank You!