Action recognition
in the spirit of object detection

Nick Turner, Sven Dorkenwald

COS 598 - 04/23/18
Temporal activity detection

Classify

1) Action
2) Temporal window

Example from Charades
Fixed time contexts in prior approaches

Prior two-step approaches:
(1) classify action → (2) agglomerate actions
Fixed time contexts in prior approaches

Prior two-step approaches:
(1) classify action → (2) agglomerate actions

16 frame input to C3D and extracted features in conv5b (last convolution)
“Advanced” temporal action localization

(1) **R-C3D** End-to-end model with combined activity proposal and classification stages

(2) **CMS-RC3D** Contextual information is fused from multiple time scales
RC3D

TASK REVIEW

NOVELTY

EXPERIMENTS

DISCUSSION I

CMS-RC3D

MOTIVATING PROBLEMS

NOVELTY

EXPERIMENTS

DISCUSSION II
R-C3D uses features at **any** granularity

![Diagram](image-url)

- Input Video
- Fully Convolutional 3-D Filters
- Feature Maps
- “Blown-up” C3D
R-C3D uses features at any granularity
Model walkthrough
3D CNN feature extractor (C3D)

Goal: Extract spatio-temporal features

L: number of frames (limited by memory)
H = W = 112

X 512 (feature dim)
Proposal subnet

Goal: Predict which anchor segments contain actions

Input video

3D ConvNet

512 x L/8 x 1 x 1
Proposal subnet

Goal: Predict which anchor segments contain actions
Proposal subnet

Goal: Predict which anchor segments contain actions

number of multiscale anchor segments = \( L / 8 \times K \)

K: number of scales ("dataset dependent")
Proposal subnet

Goal: Predict which anchor segments contain actions

1. Classify $L / 8 \times K$ segments as background vs action
2. Infer (offset, length difference) from anchor segments
Classification subnet

Goal: Select and classify proposals

condensing the proposals
Goal: Select and classify proposals

Problem: arbitrarily long regions
Classification subnet

Goal: Select and classify proposals

fixed sized 3D RoI pooling

L/8

7 x 512
Classification subnet

Goal: Select and classify proposals

fixed sized 3D RoI pooling

8192 features
Classification subnet

Goal: Select and classify proposals
Training the two subnets **jointly**

Regression on time-window + Classification on action / background

Regression on time-window + Classification on action
Loss function

Classification loss

Proposal net: single class
Classification net: multiclass

\[
\text{Loss} = \frac{1}{N_{\text{cls}}} \sum_i L_{\text{cls}}(a_i, a_i^*) + \lambda \frac{1}{N_{\text{reg}}} \sum_i a_i^* L_{\text{reg}}(t_i, t_i^*)
\]

Regression loss on time window

Time window:

\[
t_i = \{\delta \hat{c}_i, \delta \hat{l}_i\}
\]

\[
\begin{align*}
\delta c_i &= (c_i^* - c_i)/l_i \\
\delta l_i &= \log(l_i^*/l_i)
\end{align*}
\]
RC3D

TASK REVIEW

NOVELTY

EXPERIMENTS

DISCUSSION I

CMS-RC3D

MOTIVATING PROBLEMS

NOVELTY

EXPERIMENTS

DISCUSSION II
Qualitative evaluation on ActivityNet

Overlapping actions

GT ———
R-C3D ———
Qualitative evaluation on Charades
# Results on THUMOS’ 14

<table>
<thead>
<tr>
<th></th>
<th>IoU</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.1</td>
</tr>
<tr>
<td>Karaman et al. [13]</td>
<td>4.6</td>
</tr>
<tr>
<td>Wang et al. [37]</td>
<td>18.2</td>
</tr>
<tr>
<td>Oneata et al. [20]</td>
<td>36.6</td>
</tr>
<tr>
<td>Heilbron et al. [10]</td>
<td>-</td>
</tr>
<tr>
<td>Escorcia et al. [4]</td>
<td>-</td>
</tr>
<tr>
<td>Richard et al. [22]</td>
<td>39.7</td>
</tr>
<tr>
<td>Yeung et al. [39]</td>
<td>48.9</td>
</tr>
<tr>
<td>Yuan et al. [41]</td>
<td>51.4</td>
</tr>
<tr>
<td>Shou et al. [24]</td>
<td>47.7</td>
</tr>
<tr>
<td>Shou et al. [23]</td>
<td>-</td>
</tr>
<tr>
<td>R-C3D (our one-way buffer)</td>
<td>51.6</td>
</tr>
<tr>
<td>R-C3D (our two-way buffer)</td>
<td><strong>54.5</strong></td>
</tr>
</tbody>
</table>

Proposal classification: 85% precision, 83% recall
### Results on ActivityNet

**mAP@0.5**

<table>
<thead>
<tr>
<th>Method</th>
<th>Train Data</th>
<th>Validation</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>G. Singh et. al. [30]</td>
<td>train</td>
<td>34.5</td>
<td>36.4</td>
</tr>
<tr>
<td>B. Singh et. al. [29]</td>
<td>train+val</td>
<td>-</td>
<td>28.8</td>
</tr>
<tr>
<td>UPC [18]</td>
<td>train</td>
<td>22.5</td>
<td>22.3</td>
</tr>
<tr>
<td>R-C3D (ours)</td>
<td>train</td>
<td>26.8</td>
<td>26.8</td>
</tr>
<tr>
<td>R-C3D (ours)</td>
<td>train+val</td>
<td>-</td>
<td>28.4</td>
</tr>
</tbody>
</table>
RC3D is faster than existing methods

Inference speeds:

<table>
<thead>
<tr>
<th>Model</th>
<th>FPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>S-CNN [24]</td>
<td>60</td>
</tr>
<tr>
<td>DAP [4]</td>
<td>134.1</td>
</tr>
<tr>
<td>R-C3D (ours on Titan X Maxwell)</td>
<td>569</td>
</tr>
<tr>
<td>R-C3D (ours on Titan X Pascal)</td>
<td>1030</td>
</tr>
</tbody>
</table>
R-C3D key takeaways

(1) An End-to-end solution allows for arbitrary time granularity → can handle overlapping activity → improvements in performance

(2) Performance of the proposal net might / should allow for better activity prediction

(3) Newer graphics cards lead to large speed-ups
RC3D

TASK REVIEW

NOVELTY

EXPERIMENTS

DISCUSSION I

CMS-RC3D

MOTIVATING PROBLEMS

NOVELTY

EXPERIMENTS

DISCUSSION II
RC3D

TASK REVIEW

NOVELTY

EXPERIMENTS

DISCUSSION I

CMS-RC3D

MOTIVATING PROBLEMS
Multiple Timescales
Context

NOVELTY

EXPERIMENTS

DISCUSSION II
Activities take place over very different timescales

...perhaps representing multiple timescales will aid in activity detection
Other approaches use context outside of the “activity window” itself to assist prediction
RC3D

TASK REVIEW

NOVELTY

EXPERIMENTS

DISCUSSION I

CMS-RC3D

MOTIVATING PROBLEMS

NOVELTY

Multiple Timescales

Context

EXPERIMENTS

DISCUSSION II
Representing multiple time scales
Representing multiple time scales

This is slightly misleading
Adding context
Do they add half the window to each side?
Or just double the length?
This is slightly misleading
How do we pick the scale at which to pool a given proposal?
How do we pick the scale at which to pool a given proposal?

Strategy 1
"S1"

Strategy 2
"S2"

Strategy 3
"S3"
RC3D

TASK REVIEW

NOVELTY

EXPERIMENTS

DISCUSSION I

CMS-RC3D

MOTIVATING PROBLEMS

NOVELTY

Multiple Timescales

Context

EXPERIMENTS

DISCUSSION II
RC3D

TASK REVIEW

NOVELTY

EXPERIMENTS

DISCUSSION I

CMS-RC3D

MOTIVATING PROBLEMS

NOVELTY

EXPERIMENTS

Ablation Studies

Evaluations

DISCUSSION II
### ActivityNet Evaluation

<table>
<thead>
<tr>
<th>Method</th>
<th>0.5</th>
<th>0.75</th>
<th>0.95</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>RC3D [33]</td>
<td>26.33</td>
<td>10.46</td>
<td>1.25</td>
<td>12.71</td>
</tr>
</tbody>
</table>

### THUMOS ‘14 Evaluation

<table>
<thead>
<tr>
<th>Method</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>RC3D [33]</td>
<td>54.5</td>
<td>51.5</td>
<td>44.8</td>
<td>35.6</td>
<td>28.9</td>
</tr>
</tbody>
</table>
Are multi-scale proposals useful?

<table>
<thead>
<tr>
<th>Method</th>
<th>0.5</th>
<th>0.75</th>
<th>0.95</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>RC3D [33]</td>
<td>26.33</td>
<td>10.46</td>
<td>1.25</td>
<td>12.71</td>
</tr>
<tr>
<td>MS(MAX)(S1)</td>
<td>27.65</td>
<td><strong>13.93</strong></td>
<td>1.12</td>
<td>14.91</td>
</tr>
<tr>
<td>MS(CONV)(S1)</td>
<td>28.01</td>
<td><strong>13.80</strong></td>
<td>1.20</td>
<td>15.12</td>
</tr>
</tbody>
</table>
How do we pick the scale at which to classify a given proposal?

<table>
<thead>
<tr>
<th>Method</th>
<th>0.5</th>
<th>0.75</th>
<th>0.95</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>MS(CONV)(S1)(CTX)</td>
<td>32.57</td>
<td>16.92</td>
<td>1.07</td>
<td>17.89</td>
</tr>
<tr>
<td>MS(CONV)(S2)(CTX)</td>
<td>31.89</td>
<td>17.23</td>
<td>1.16</td>
<td>17.72</td>
</tr>
<tr>
<td>MS(CONV)(S3)(CTX)</td>
<td>32.92</td>
<td>18.36</td>
<td>1.13</td>
<td>18.46</td>
</tr>
<tr>
<td>Method</td>
<td>0.5</td>
<td>0.75</td>
<td>0.95</td>
<td>Average</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>----------</td>
</tr>
<tr>
<td>RC3D [33]</td>
<td>26.33</td>
<td>10.46</td>
<td>1.25</td>
<td>12.71</td>
</tr>
<tr>
<td>MS(MAX)(S1)</td>
<td>27.65</td>
<td>13.93</td>
<td>1.12</td>
<td>14.91</td>
</tr>
<tr>
<td>MS(CONV)(S1)</td>
<td>28.01</td>
<td>13.80</td>
<td>1.20</td>
<td>15.12</td>
</tr>
<tr>
<td>MS(MAX)(S1)(CTX)</td>
<td>31.81</td>
<td>17.05</td>
<td>1.06</td>
<td>17.58</td>
</tr>
<tr>
<td>MS(CONV)(S1)(CTX)</td>
<td>32.57</td>
<td>16.92</td>
<td>1.07</td>
<td>17.89</td>
</tr>
<tr>
<td>MS(CONV)(S2)(CTX)</td>
<td>31.89</td>
<td>17.23</td>
<td>1.16</td>
<td>17.72</td>
</tr>
<tr>
<td>MS(CONV)(S3)(CTX)</td>
<td>32.92</td>
<td>18.36</td>
<td>1.13</td>
<td>18.46</td>
</tr>
</tbody>
</table>
Both results together

<table>
<thead>
<tr>
<th>ABSOLUTE</th>
<th>No Multi-Scale</th>
<th>Multi-Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Context</td>
<td>12.71</td>
<td>15.01</td>
</tr>
<tr>
<td>Context</td>
<td>??</td>
<td>17.91</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RELATIVE</th>
<th>No Multi-Scale</th>
<th>Multi-Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Context</td>
<td>0.0</td>
<td>2.3</td>
</tr>
<tr>
<td>Context</td>
<td>??</td>
<td>5.2 (2.3+2.9?)</td>
</tr>
</tbody>
</table>
RC3D

TASK REVIEW

NOVELTY

EXPERIMENTS

DISCUSSION I

CMS-RC3D

MOTIVATING PROBLEMS

NOVELTY

EXPERIMENTS

Ablation Studies Evaluations

DISCUSSION II
## THUMOS 2014

<table>
<thead>
<tr>
<th>Method</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Karaman <em>et al.</em> [16]</td>
<td>4.6</td>
<td>3.4</td>
<td>2.1</td>
<td>1.4</td>
<td>0.9</td>
</tr>
<tr>
<td>Wang <em>et al.</em> [31]</td>
<td>18.2</td>
<td>17.0</td>
<td>14.0</td>
<td>11.7</td>
<td>8.3</td>
</tr>
<tr>
<td>Oneata <em>et al.</em> [20]</td>
<td>36.6</td>
<td>33.6</td>
<td>27.0</td>
<td>20.8</td>
<td>14.4</td>
</tr>
<tr>
<td>SparseProp [4]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DAPs [9]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SLM [23]</td>
<td>39.7</td>
<td>35.7</td>
<td>30.0</td>
<td>23.2</td>
<td>15.2</td>
</tr>
<tr>
<td>FG [35]</td>
<td>48.9</td>
<td>44.0</td>
<td>36.0</td>
<td>26.4</td>
<td>17.1</td>
</tr>
<tr>
<td>PSDF [36]</td>
<td>51.4</td>
<td>42.6</td>
<td>33.6</td>
<td>26.1</td>
<td>18.8</td>
</tr>
<tr>
<td>S-CNN [25]</td>
<td>47.7</td>
<td>43.5</td>
<td>36.3</td>
<td>28.7</td>
<td>19.0</td>
</tr>
<tr>
<td>CDC [24]</td>
<td>-</td>
<td>-</td>
<td>40.1</td>
<td>29.4</td>
<td>23.3</td>
</tr>
<tr>
<td>TCN [8]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>33.3</td>
<td>25.6</td>
</tr>
<tr>
<td>RC3D [33]</td>
<td>54.5</td>
<td>51.5</td>
<td>44.8</td>
<td>35.6</td>
<td>28.9</td>
</tr>
<tr>
<td>SS-TAD [1]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>45.7</td>
<td>29.2</td>
</tr>
<tr>
<td>SSN [37]</td>
<td>66.0</td>
<td>59.4</td>
<td>51.9</td>
<td>41.0</td>
<td>29.8</td>
</tr>
<tr>
<td>Our RC3D</td>
<td>57.4</td>
<td>54.9</td>
<td>51.1</td>
<td>43.1</td>
<td>35.8</td>
</tr>
<tr>
<td>CMS-RC3D</td>
<td>61.6</td>
<td>59.3</td>
<td>54.7</td>
<td>48.2</td>
<td>40.0</td>
</tr>
<tr>
<td>Method</td>
<td>0.1</td>
<td>0.2</td>
<td>0.3</td>
<td>0.4</td>
<td>0.5</td>
</tr>
<tr>
<td>--------------</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>PSDF [36]</td>
<td>51.4</td>
<td>42.6</td>
<td>33.6</td>
<td>26.1</td>
<td>18.8</td>
</tr>
<tr>
<td>TCN [8]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>33.3</td>
<td>25.6</td>
</tr>
<tr>
<td>RC3D [33]</td>
<td>54.5</td>
<td>51.5</td>
<td>44.8</td>
<td>35.6</td>
<td>28.9</td>
</tr>
<tr>
<td>SSN [37]</td>
<td>66.0</td>
<td>59.4</td>
<td>51.9</td>
<td>41.0</td>
<td>29.8</td>
</tr>
<tr>
<td>Our RC3D</td>
<td>57.4</td>
<td>54.9</td>
<td>51.1</td>
<td>43.1</td>
<td>35.8</td>
</tr>
<tr>
<td>CMS-RC3D</td>
<td>61.6</td>
<td>59.3</td>
<td>54.7</td>
<td>48.2</td>
<td>40.0</td>
</tr>
</tbody>
</table>
### Activity Net (version 1.3)

<table>
<thead>
<tr>
<th>Method</th>
<th>0.5</th>
<th>0.75</th>
<th>0.95</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>RC3D [33]</td>
<td>26.45</td>
<td>11.47</td>
<td>1.69</td>
<td>13.33</td>
</tr>
<tr>
<td>MSN [28]</td>
<td>28.67</td>
<td>17.78</td>
<td>2.88</td>
<td>17.68</td>
</tr>
<tr>
<td>TCN [8]</td>
<td>37.49</td>
<td>23.47</td>
<td>4.47</td>
<td></td>
</tr>
<tr>
<td>SSN [37]</td>
<td>43.26</td>
<td>28.70</td>
<td>5.63</td>
<td>28.28</td>
</tr>
<tr>
<td>CMS-RC3D</td>
<td>32.79</td>
<td>18.39</td>
<td>1.24</td>
<td>18.68</td>
</tr>
</tbody>
</table>
Shallower Feature Extractor?

C3D

Two-Stream Network
Shallower Feature Extractor?

From the **ORIGINAL** RC3D Paper

<table>
<thead>
<tr>
<th></th>
<th>mAP</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>standard</td>
<td>post-process</td>
<td></td>
</tr>
<tr>
<td>Random [25]</td>
<td>4.2</td>
<td>4.2</td>
<td></td>
</tr>
<tr>
<td>RGB [25]</td>
<td>7.7</td>
<td>8.8</td>
<td></td>
</tr>
<tr>
<td>Two-Stream [25]</td>
<td>7.7</td>
<td>10.0</td>
<td></td>
</tr>
<tr>
<td>Two-Stream+LSTM[25]</td>
<td>8.3</td>
<td>8.8</td>
<td></td>
</tr>
<tr>
<td>Sigurdsson et al. [25]</td>
<td>9.6</td>
<td>12.1</td>
<td></td>
</tr>
<tr>
<td>R-C3D (ours)</td>
<td><strong>12.4</strong></td>
<td><strong>12.7</strong></td>
<td></td>
</tr>
</tbody>
</table>
RC3D

TASK REVIEW

NOVELTY

EXPERIMENTS

DISCUSSION I

CMS-RC3D

MOTIVATING PROBLEMS

NOVELTY

EXPERIMENTS

Ablation Studies

Evaluations

DISCUSSION II
RC3D

Task Review

Novelty

Experiments

Discussion I

CMS-RC3D

Motivating Problems

Novelty

Experiments

Discussion II
Lingering Thoughts

It doesn’t seem like the feature extractor is the core reason why TCN and SSN might outperform this system. Perhaps something dataset-specific is at work here?

Do windows with “context” include extra information both before and after? Or just after?
**RC3D**

**Summary**

**CMS-RC3D**

- Same time windows
  - + extra context
  - + multiple resolutions
Thank You!

(this is in THUMOS2014)
Action recognition
in the spirit of object detection

Nick Turner, Sven Dorkenwald

COS 598 - 04/23/18
PROPOSED OUTLINE:

OVERVIEW

TASK DEFINITION / REVIEW (ask about this)
-What are we trying to do?
-What prior methods have we seen so far?
--C3D Architecture

NOVELTY:
Review R-CNN / Faster R-CNN
-Region proposals -> refined classifications

-R-C3D

EXPERIMENTS
-Training Procedure
-Representing ground truth activities
-Forming the loss function
-Performance Experiments
-Activity Detection Speed

DISCUSSION I (?)
-Lots of references to hand designed features. What's the true issue there?

CMS-RC3D

PROPOSED PROBLEMS WITH R-C3D
-Multiple time scales - show an example video
-Use of “contextual information”

NOVELTY
-Multiple time scales
-Contextual information

EXPERIMENTS
-Training Procedure
-Representing ground truth activities
-Forming the loss function
-Ablation Studies (Do they analyze R-C3D with CTX but without MS anywhere?)
Which variables are most important? Reformat the results table?

DISCUSSION (?)
-Are there other experiment we wish they would do? What’s really most important?

END