A Fine-Grained Perspective onto Object Interactions
Natural interactions
Scaling Egocentric Vision:
The EPIC-KITCHENS Dataset
Scaling...
... Egocentric Vision
Scaling Egocentric Vision

CMU (2009)

ADL (2012)

Charades-Ego (2018)

BEOID (2014)

GTEA ... (2011-)

EGTEA+ (2018)
Native & Multiple Environments?

CMU (2009)

ADL (2012)

Charades-Ego (2018)

BEOID (2014)

GTEA … (2011-)

EGTEA+ (2018)
Non-scripted?

CMU (2009)

ADL (2012)

Charades-Ego (2018)

BEOID (2014)

GTEA … (2011-)

EGTEA+ (2018)
Scaling Egocentric Vision

Data Collection

Benchmark and Challenges

Native Environment, Natural Interactions

Active Object Bounding Boxes

Live Narrations

Dense Action Segments

EPIC-Kitchens Object Detection

Secret URL: https://competitions.codalab.org

Organized by hazeldoughty - Current server time: S

Learn the Details Phases Participate Results

EPIC-Kitchens Object Detection

DCCV 2018 Object Recognition Challenge

June 30, 2018, midnight UTC
Data Collection

- 32 kitchens
- Single-person environments
- 4 cities
- May – Nov 2017 – 55 hours
- 10 nationalities
- 3 days - all kitchen activities
39 000
ACTION SEGMENTS
Annotations (3) – Object Bounding Boxes

Action segments

(kitchen | towel | tofu)
(dishwasher | kitchen | towel)
(kitchen | towel | tofu)
454 200
OBJECT ANNOTATIONS
|take, grab, pick, get, fetch, pick-up, ...

- 125 verb classes
- 331 noun classes
Train/Test Splits

● 20% - Seen Test Set
  ○ 28 Kitchens

● 7% - Unseen Test Set
  ○ 4 Kitchens

Table 4: Statistics of test splits: seen (S1) and unseen (S2) kitchens

<table>
<thead>
<tr>
<th></th>
<th>#Subjects</th>
<th>#Sequences</th>
<th>Duration (s)</th>
<th>%</th>
<th>Narrated Segments</th>
<th>Action Segments</th>
<th>Bounding Boxes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train/Val</td>
<td>28</td>
<td>272</td>
<td>141731</td>
<td></td>
<td>28,587</td>
<td>28,561</td>
<td>326,388</td>
</tr>
<tr>
<td>S1 Test</td>
<td>28</td>
<td>106</td>
<td>39084</td>
<td>20%</td>
<td>8,069</td>
<td>8,064</td>
<td>97,872</td>
</tr>
<tr>
<td>S2 Test</td>
<td>4</td>
<td>54</td>
<td>13231</td>
<td>7%</td>
<td>2,939</td>
<td>2,939</td>
<td>29,995</td>
</tr>
</tbody>
</table>
Three open challenges:

- Action Recognition
- Action Anticipation
- Object Detection
Action Recognition Challenge
Given a trimmed action segment:
\((t_{\text{start}}, t_{\text{stop}})\)
classify the action within.

\[
\hat{y}_{\text{verb}} = \text{open} \\
\hat{y}_{\text{noun}} = \text{oven}
\]

\[
\hat{y}_{\text{action}} = (\text{open, oven})
\]
<table>
<thead>
<tr>
<th>#</th>
<th>User</th>
<th>Entries</th>
<th>Date of Last Entry</th>
<th>Team Name</th>
<th>Top-1 Accuracy (%)</th>
<th>Top-5 Accuracy (%)</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Verb (1)</td>
<td>Noun (1)</td>
<td>Action (1)</td>
<td>Verb (1)</td>
</tr>
<tr>
<td>1</td>
<td>wasun</td>
<td>3</td>
<td>08/12/19</td>
<td>UTS_Baidu</td>
<td>69.80 (2)</td>
<td>52.27 (3)</td>
<td>41.37 (1)</td>
<td>90.95 (1)</td>
</tr>
<tr>
<td>2</td>
<td>action_banks</td>
<td>2</td>
<td>11/11/19</td>
<td></td>
<td>63.23 (6)</td>
<td>47.00 (2)</td>
<td>39.34 (2)</td>
<td>86.94 (5)</td>
</tr>
<tr>
<td>3</td>
<td>aptx4869lim</td>
<td>10</td>
<td>09/27/19</td>
<td>GT-WISC-MPI</td>
<td>68.14 (2)</td>
<td>49.62 (3)</td>
<td>38.18 (3)</td>
<td>88.93 (5)</td>
</tr>
<tr>
<td>4</td>
<td>weiyaowang</td>
<td>13</td>
<td>11/14/19</td>
<td></td>
<td>65.91 (4)</td>
<td>48.48 (4)</td>
<td>36.76 (4)</td>
<td>89.51 (3)</td>
</tr>
<tr>
<td>5</td>
<td>TBN_Ensemble</td>
<td>1</td>
<td>07/20/19</td>
<td>Bristol-Oxford</td>
<td>66.10 (3)</td>
<td>47.88 (4)</td>
<td>36.66 (4)</td>
<td>91.28 (3)</td>
</tr>
<tr>
<td>6</td>
<td>Sudhakaran</td>
<td>11</td>
<td>08/10/19</td>
<td>FBK_HuPBA</td>
<td>63.34 (5)</td>
<td>44.75 (6)</td>
<td>35.54 (6)</td>
<td>89.01 (4)</td>
</tr>
<tr>
<td>7</td>
<td>antoninofurnari</td>
<td>1</td>
<td>07/19/19</td>
<td></td>
<td>56.93 (9)</td>
<td>43.05 (7)</td>
<td>33.06 (7)</td>
<td>85.68 (11)</td>
</tr>
<tr>
<td>8</td>
<td>georkap</td>
<td>6</td>
<td>08/29/19</td>
<td>uuol</td>
<td>58.36 (7)</td>
<td>40.87 (8)</td>
<td>30.79 (8)</td>
<td>87.49 (7)</td>
</tr>
<tr>
<td>9</td>
<td>bduke</td>
<td>2</td>
<td>11/16/19</td>
<td></td>
<td>57.52 (8)</td>
<td>38.00 (9)</td>
<td>26.98 (9)</td>
<td>88.27 (6)</td>
</tr>
<tr>
<td>10</td>
<td>hepic</td>
<td>6</td>
<td>12/04/19</td>
<td></td>
<td>49.96 (11)</td>
<td>39.24 (10)</td>
<td>25.30 (10)</td>
<td>85.80 (10)</td>
</tr>
<tr>
<td>11</td>
<td>Saptarshi_sinha</td>
<td>8</td>
<td>12/19/19</td>
<td></td>
<td>52.78 (10)</td>
<td>39.72 (11)</td>
<td>23.69 (11)</td>
<td>81.66 (18)</td>
</tr>
<tr>
<td>12</td>
<td>gkallari</td>
<td>3</td>
<td>11/20/19</td>
<td></td>
<td>49.78 (9)</td>
<td>32.62 (12)</td>
<td>21.69 (12)</td>
<td>82.93 (15)</td>
</tr>
</tbody>
</table>
Evaluating Action Recognition Models

### Evaluating Action Recognition Models


<table>
<thead>
<tr>
<th>Model</th>
<th>GFLOP/s RGB</th>
<th>GFLOP/s Flow</th>
<th>Params (M) RGB</th>
<th>Params (M) Flow</th>
</tr>
</thead>
<tbody>
<tr>
<td>TSN</td>
<td>33.12</td>
<td>35.33</td>
<td>24.48</td>
<td>24.51</td>
</tr>
<tr>
<td>TRN</td>
<td>33.12</td>
<td>35.32</td>
<td>25.33</td>
<td>25.35</td>
</tr>
<tr>
<td>M-TRN</td>
<td>33.12</td>
<td>35.33</td>
<td>27.18</td>
<td>27.21</td>
</tr>
<tr>
<td>TSM</td>
<td>33.12</td>
<td>35.33</td>
<td>24.48</td>
<td>24.51</td>
</tr>
</tbody>
</table>

Table 3: Model parameter and FLOP/s count using a ResNet-50 backbone with 8 segments for a single video.
http://epic-kitchens.github.io

NEWS

- EPIC-KITCHENS accepted for oral presentation at ECCV 2018 in Munich this September
- News coverage: Unib, The Spoon, El Sole 24 Ore, La Sicilia, Elpais
- EPIC-Kitchens Released: 9th of April 2018!!!
- Watch YouTube Release Trailer here

What is EPIC-Kitchens?

The largest dataset in first-person (occupentics) vision; multi-faceted non-scripted recordings in native environments - i.e. the wearers' homes, capturing all daily activities in the kitchen over multiple days. Annotations are collected using a novel 'live' audio commentary approach.

Characteristics

- 30 kitchens - 4 cities
- 16 head-mounted camera
- 86 hours of recording - Full HD, 30fps
- 11,841 frames
- Multi-language narrations
- 35,914 action segments
- 454,009 object bounding boxes
- 125 verb classes, 362 noun classes

Updates

Stay tuned with updates on epic-kitchens2018, as well as EPIC workshop series by joining the epic-community mailing list send an email to epkpa@cs.bristol.ac.uk with the subject subscribe epic-community and a blank message body.
Fine-Grained Object Interactions

- Skill Determination
- Action Completion
- Single-timestamp

How 'well'

Multi-Modal

• Vision+Audio
• Vision+Language

When

Dual-

• DDLSTM
• Retro-Actions
Fine-Grained Object Interactions

- **Skill Determination**
- **Action Completion**

- **How 'well'**

- **When**
- **Single-timestamp**

- **Multi-Modal**
  - **Vision+Audio**
  - **Vision+Language**

- **Dual-**
  - **DDLSTM**
  - **Retro-Actions**

Assess relative skill for a collection of video sequences, applicable to a variety of tasks.
Skill Determination from Video

**Input:** Pairwise annotations of videos, indicating higher skill or no skill preference
The Pros and Cons: Rank-Aware Temporal Attention

\[ \sum_{(p_i, p_j) \in \Phi} \max(0, m_2 - (s^+(p_i) - s^+(p_j)) + (u(p_i) - u(p_j)) \]

\[ \sum_{(p_i, p_j) \in \Phi} \max(0, m - s^-(p_i) + s^+(p_j)) \]

\[ \sum_{(p_i, p_j) \in \Phi} \max(0, m - s^+(p_i) + u(p_j)) \]

\[ \sum_{(p_i, p_j) \in \Phi} \max(0, m_3 - (s^+(p_i) - s^-(p_j)) + (u(p_i) - u(p_j)) \]
### The Pros and Cons: Rank-Aware Temporal Attention

<table>
<thead>
<tr>
<th>Method</th>
<th>EPIC Skills</th>
<th>BEST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Who’s Better [7]</td>
<td>76.0</td>
<td>75.8</td>
</tr>
<tr>
<td>Last Segment</td>
<td>76.8</td>
<td>61.0</td>
</tr>
<tr>
<td>Uniform Weighting</td>
<td>78.8</td>
<td>73.6</td>
</tr>
<tr>
<td>Softmax Attention</td>
<td>74.5</td>
<td>72.3</td>
</tr>
<tr>
<td>STPN [22]</td>
<td>74.3</td>
<td>70.0</td>
</tr>
<tr>
<td>Ours (Rank Aware Attention)</td>
<td><strong>80.3</strong></td>
<td><strong>81.2</strong></td>
</tr>
</tbody>
</table>

Table 2. Results of our method in comparison to baseline. Our final method outperforms every baseline on both datasets.
The Pros and Cons: Rank-Aware Temporal Attention

Low-skill Attention Module

Surgery

Apply Eyeliner

Origami

The Pros and Cons: Rank-Aware Temporal Attention

High-skill Attention Module

Dough Rolling

Origami

Drawing

Skill Determination in Video

Holes in the dough
Curved or rolled edges
Tissue damage
Spoon
Abnormal needle pass
Loose Stitching

Best  Worst

The Pros and Cons: Rank-Aware Temporal Attention for Skill Determination in Long Videos

Hazel Doughty
Walterio Mayol-Cuevas
Dima Damen

University of Bristol

ABSTRACT VIDEO DOWNLOADS BIBTEX RELATED

Abstract

We present a new model to determine relative skill from long videos, through learnable temporal attention modules. Skill determination is formulated as a ranking problem, making it suitable for common and generic tasks. However, for long videos, parts of the video are irrelevant for assessing skill, and there may be variability in the skill exhibited throughout a video. We therefore propose a method which assesses the relative overall level of skill in a long video by attending to its skill-relevant parts.

Our approach trains temporal attention modules, learned with only video-level supervision, using a novel rank-aware loss function. In addition to attending to task-relevant video parts, our proposed loss jointly trains two attention modules to separately attend to video parts which are indicative of higher (pros) and lower (cons) skill. We evaluate our approach on the EPIC-Skills dataset and additionally annotate a larger dataset from YouTube videos for skill determination with five previously unexplored tasks. Our method outperforms previous approaches and classic softmax attention on both datasets by over 4% pairwise accuracy, and as much as 12% on individual tasks. We also demonstrate our model’s ability to attend to

Downloads

- Paper [PDF] [ArXiv]
- Supplementary [Video]
- Code and data [GitHub - Available Now]

Dima Damen
26 December 2019

Dima Damen
with: Hazel Doughty
Walterio Mayol-Cuevas

Computer Vision and Pattern Recognition (CVPR) 2019

Fine-Grained Object Interactions

- Skill Determination
- Action Completion

- Single-timestamp

When

Multi-Modal
- Vision+Audio
- Vision+Language

Dual-
- DDLSTM
- Retro-Actions

How 'well'
Action Completion Detection

Action Completion Detection

Action Completion Detection

- Each frame in the sequence contributes to the completion moment detection via ‘voting’
1. Classification-Based Voting
Action Completion Detection

1. Classification-Based Voting
Action Completion Detection

2. Regression-Based Voting
2. Regression-Based Voting
Action Completion Detection

Action Completion Detection

**Action Completion Detection**

**Frame-level labels:** annotations are expensive, subjective and noisy.

We detect completion using only **weak labels** during training.

**sequence-level complete and incomplete labels**
Action Completion Detection

\[
\begin{align*}
\text{Attention LSTM} & \quad \rightarrow \quad O_t^a \\
\text{Completion LSTM} & \quad \rightarrow \quad O_t^s \\
\text{Softmax} & \quad \rightarrow \quad a_t \\
\sigma(o_t^s \times a_t) & \quad \rightarrow \quad S_t
\end{align*}
\]
Action Completion Detection

Completion scores
Attention scores
WS-U
WS-Att
GT

Fine-Grained Object Interactions

- Skill Determination
- Action Completion

- Single-timestamp

- How 'well'
- When

- Multi-Modal
- Dual-

- Vision+Audio
- Vision+Language

- DDLSTM
- Retro-Actions
**Temporal Boundaries for Object Interactions**

- How robust are current state-of-the-art approaches to annotated boundaries in test segments?
- Modify test segment boundaries, maintaining significant overlap of segments IoU > 0.5
- **Correct in Green – Incorrect in Red**

Trespassing the Boundaries

GTEA Gaze+

*ground truth*

![Image](image.png)

*predicted class: take knife*

Trespassing the Boundaries

Action Recognition from a Single Timestamp

- Learning from Single timestamps
Action Recognition from a Single Timestamp

i) EPIC Kitchens (success)

epoch: 84
Fine-Grained Object Interactions

- Skill Determination
- Action Completion
- Single-timestamp
- Vision+Audio
- Vision+Language
- DDLSTM
- Retro-Actions
Dual-Domain LSTM for Cross-Dataset Action Recognition

Dual-Domain LSTM for Cross-Dataset Action Recognition

BNLSTM
1 dataset

DDLSTM
2 datasets

Dual-Domain LSTM for Cross-Dataset Action Recognition

\[ \mathcal{D}_1 \]

\[ \mathcal{D}_2 \]

\[ N - 1 \]

Batch index

Batch features

\[ \begin{array}{c}
\tau_2 \\
1 - \alpha_2 N \\
\alpha_1 N \\
0.0 \\
0.5 \\
1.0
\end{array} \]

Contribution
Dual-Domain LSTM for Cross-Dataset Action Recognition

$$\tau_1(\alpha_1, j) = \frac{1 - \tanh(j - \alpha_1 N)}{2}$$

$$\tau_2(\alpha_2, j) = \frac{1 + \tanh(j - \alpha_2 N)}{2}$$
Dual-Domain LSTM for Cross-Dataset Action Recognition

Dual-Domain LSTM for Cross-Dataset Action Recognition

(a) Breakfast
(b) 50 Salads
(c) MPII Cooking 2
## Dual-Domain LSTM for Cross-Dataset Action Recognition

<table>
<thead>
<tr>
<th>D1</th>
<th>D2</th>
<th>Training</th>
<th>LSTM Type</th>
<th>D1 Acc</th>
<th>D2 Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>ActivityNet</td>
<td>50 Salads</td>
<td>Pt/ft</td>
<td>LSTM</td>
<td>44.4</td>
<td>42.1</td>
</tr>
<tr>
<td>ActivityNet</td>
<td>50 Salads</td>
<td>Joint</td>
<td>DDLSTM</td>
<td>44.3</td>
<td>42.2</td>
</tr>
<tr>
<td>Thumos</td>
<td>50 Salads</td>
<td>Pt/ft</td>
<td>LSTM</td>
<td>65.9</td>
<td>42.0</td>
</tr>
<tr>
<td>Thumos</td>
<td>50 Salads</td>
<td>Joint</td>
<td>DDLSTM</td>
<td>66.1</td>
<td>42.3</td>
</tr>
<tr>
<td>EPIC</td>
<td>50 Salads</td>
<td>Pt/ft</td>
<td>LSTM</td>
<td>31.5</td>
<td><strong>44.9</strong></td>
</tr>
<tr>
<td>EPIC</td>
<td>50 Salads</td>
<td>Joint</td>
<td>DDLSTM</td>
<td>33.1</td>
<td><strong>48.9</strong></td>
</tr>
</tbody>
</table>
Fine-Grained Object Interactions

- Skill Determination
- Action Completion

- Single-timestamp

How 'well'

When

Multi-Modal

- Vision+Audio
- Vision+Language

Dual-

- DDLSTM
- Retro-Actions
Retro-actions
Retro-actions

moving [part] of [something]

removing [something], revealing [something] behind

putting [something] in front of [something]

poking a stack of [something] so the stack collapses

irreversible
# Retro-actions

<table>
<thead>
<tr>
<th>ORIG</th>
<th>Opening</th>
<th>Pulling Left to Right</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 2 3</td>
<td><img src="image1" alt="Opening" /></td>
<td><img src="image2" alt="Pulling" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>HF</th>
<th>Opening</th>
<th>Pulling Right to Left</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 2 1</td>
<td><img src="image3" alt="Opening" /></td>
<td><img src="image4" alt="Pulling" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TR</th>
<th>Closing</th>
<th>Pushing Right to Left</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 2 1</td>
<td><img src="image5" alt="Closing" /></td>
<td><img src="image6" alt="Pushing" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>HF+TR</th>
<th>Closing</th>
<th>Pushing Left to Right</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 3</td>
<td><img src="image7" alt="Closing" /></td>
<td><img src="image8" alt="Pushing" /></td>
</tr>
</tbody>
</table>

Retro-actions

<table>
<thead>
<tr>
<th>No Transform</th>
<th>HF No LT</th>
<th>HF LT</th>
<th>TR No LT</th>
<th>TR LT</th>
</tr>
</thead>
</table>

- Invariant
- Equivariant
- Other

ICCV MDALC Workshop

Dima Damen
26 December 2019
Retro-actions – Zero-Shot Learning

Fine-Grained Object Interactions

- Skill Determination
- Action Completion

- Single-timestamp

- Vision+Audio
- Vision+Language

- DDLSTM
- Retro-Actions
Audio-Visual Temporal Binding for Egocentric Action Recognition

with: Vangelis Kazakos, Arsha Nagrani, Andrew Zisserman

Audio-Visual Temporal Binding for Egocentric Action Recognition

Audio-Visual Temporal Binding for Egocentric Action Recognition

### Audio-Visual Temporal Binding for Egocentric Action Recognition

<table>
<thead>
<tr>
<th></th>
<th>Top-1 Accuracy</th>
<th>Top-5 Accuracy</th>
<th>Avg Class Precision</th>
<th>Avg Class Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>VERB</td>
<td>NOUN</td>
<td>ACTION</td>
<td>VERB</td>
</tr>
<tr>
<td>RGB</td>
<td>45.68</td>
<td>36.80</td>
<td>19.86</td>
<td>85.56</td>
</tr>
<tr>
<td>Flow</td>
<td>55.65</td>
<td>31.17</td>
<td>20.10</td>
<td>85.99</td>
</tr>
<tr>
<td>Audio</td>
<td>43.56</td>
<td>22.35</td>
<td>14.21</td>
<td>79.66</td>
</tr>
<tr>
<td>TBN (RGB+Flow)</td>
<td>60.87</td>
<td>42.93</td>
<td>30.31</td>
<td>89.68</td>
</tr>
<tr>
<td>TBN (All)</td>
<td>64.75</td>
<td>46.03</td>
<td>34.80</td>
<td>90.70</td>
</tr>
<tr>
<td>RGB</td>
<td>34.89</td>
<td>21.82</td>
<td>10.11</td>
<td>74.56</td>
</tr>
<tr>
<td>Flow</td>
<td>48.21</td>
<td>22.98</td>
<td>14.48</td>
<td>77.85</td>
</tr>
<tr>
<td>Audio</td>
<td>35.43</td>
<td>11.98</td>
<td>06.45</td>
<td>69.20</td>
</tr>
<tr>
<td>TBN (RGB+Flow)</td>
<td>49.61</td>
<td>25.68</td>
<td>16.80</td>
<td>78.36</td>
</tr>
<tr>
<td>TBN (All)</td>
<td>52.69</td>
<td>27.86</td>
<td>19.06</td>
<td>79.93</td>
</tr>
</tbody>
</table>
Audio-Visual Temporal Binding for Egocentric Action Recognition

EPIC-Fusion - Qualitative Results

E. Kazakos, A. Nagrani, A. Zisserman, D. Damen, EPIC-Fusion: Audio-Visual Temporal Binding for Egocentric Action Recognition, ICCV 2019
Audio-Visual Temporal Binding for Egocentric Action Recognition

EPIC-Fusion: Audio-Visual Temporal Binding for Egocentric Action Recognition

Evangelos Kazakos\(^1\), Arsha Nagrani\(^2\), Andrew Zisserman\(^2\) and Dima Damen\(^1\)

\(^1\)University of Bristol, VIL, \(^2\)University of Oxford, VGG

Abstract

We focus on multi-modal fusion for egocentric action recognition, and propose a novel architecture for multi-modal temporal-binding, i.e. the combination of modalities within a range of temporal offsets. We train the architecture with three modalities – RGB, Flow and Audio – and combine them with mid-level fusion alongside sparse temporal sampling of fused representations. In contrast with previous works, modalities are fused before temporal aggregation, with shared modality and fusion weights over time. Our proposed architecture is trained end-to-end, outperforming individual modalities as well as late-fusion of modalities.

We demonstrate the importance of audio in egocentric vision, on per-class basis, for identifying actions as well as interacting objects. Our method achieves state of the art results on both the seen and unseen test sets of the largest egocentric dataset: EPIC-Kitchens, on all metrics using the public leaderboard.

Downloads

- Paper [ArXiv]
- Code and models [GitHub]
Fine-Grained Object Interactions

- Skill Determination
- Action Completion

How 'well'

When

Multi-Modal

- Vision+Audio
- Vision+Language

Dual-

- Single-timestamp
- DDLSTM
- Retro-Actions
The *Verbs* Dilemma
The *Verbs* Dilemma

Open
The Verbs Dilemma
The Verbs Dilemma

Open

Cut

The Verbs Dilemma

The *Verbs* Dilemma

The Verbs Dilemma

• Action representations using a single verb is highly-ambiguous
  • Solution 1: pre-selected non-overlapping verbs (SL)
    • run, walk, open, close
  • Solution 2: Using nouns to disambiguate actions (V-N)
    • open-drawer, open-bottle, open-fridge
    • actions constrained to known nouns
  • Solution 3: Multi-verb labels (ML, SAML)
    • open, hold, pull
The Verbs Dilemma

**Single Verb**

- Pour
- Fill
- Move
- Hold
- Grasp
- Push
- Take
- Open
- Close

**Multi Verb**

- Pour
- Fill
- Move
- Hold
- Grasp
- Push
- Take
- Open
- Close

**Soft Assigned Multi Verb**

- Pour
- Fill
- Move
- Hold
- Grasp
- Push
- Take
- Open
- Close
Top 3 retrieved classes across all datasets.

Turn On/Off 
Press 
Rotate

Labelling Method can differentiate turn On/Off tap by pressing and by rotating.
In this work we focus on Fine-Grained Action Retrieval

I put meat on a ball of dough
Fine-Grained Action Retrieval

We embed the video and representations

[put]

[meat, ball, dough]

Verb Embedding

take

open

put

Noun Embedding

carrot

meat

ball

doork

dough

Dima Damen

Fine-Grained Action Retrieval

Finally, we combine the outputs and embed these into an action space
Fine-Grained Action Retrieval

Fine-Grained Action Retrieval

<table>
<thead>
<tr>
<th>EPIC</th>
<th>SEEN</th>
<th>UNSEEN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>vt</td>
<td>tv</td>
</tr>
<tr>
<td>Random Baseline</td>
<td>0.6</td>
<td>0.6</td>
</tr>
<tr>
<td>CCA Baseline</td>
<td>20.6</td>
<td>7.3</td>
</tr>
<tr>
<td>MMEN (Verb)</td>
<td>3.6</td>
<td>4.0</td>
</tr>
<tr>
<td>MMEN (Noun)</td>
<td>9.9</td>
<td>9.2</td>
</tr>
<tr>
<td>MMEN (Caption)</td>
<td>14.0</td>
<td>11.2</td>
</tr>
<tr>
<td>MMEN ([Verb, Noun])</td>
<td>18.7</td>
<td>13.6</td>
</tr>
<tr>
<td>JPoSE (Verb, Noun)</td>
<td>23.2</td>
<td>15.8</td>
</tr>
</tbody>
</table>

Table 2. Cross-modal action retrieval on EPIC.
Maximum activation examples for a neuron in a noun PoS Embedding (Cutting Board) - Figure 4
Fine-Grained Object Interactions

- Skill Determination
- Action Completion
- Single-timestamp
- DDLSTM
- Retro-Actions

How 'well'

When

Multi-Modal

Dual-Modal

- Vision+Audio
- Vision+Language

26 December 2019
Now on Arxiv…

...just **finely slice** around an inch of ginger...

...**spread it out finely** and leave it to set…
Now on Arxiv…

Method

Video $x$

..start by quickly rolling our lemons...

Dima Damen

University of BRISTOL

with: Hazel Doughty
Ivan Laptev
Walterio Mayol-Cuevas
April/May 2020....

Bigger
Better
Denser

100 hours
The Team
Thank you…

For further info, datasets, code, publications…

http://dimadamen.github.io

@dimadamen

http://www.linkedin.com/in/dimadamen

Join epic-community@bristol.ac.uk

• send an email to: sympa@sympa.bristol.ac.uk
• with the subject: subscribe epic-community
• and blank message content