A Fine(r)-Grained Perspective onto Object Interactions
Natural Interactions
Fine-Grained Object Interactions

- Skill Determination
- Action Completion

How 'well'

Dual- 

Audio-visual

Vision+
Language

- DDLSTM
- Multi-modal UDA
- Retro-Actions

- Multi-Verb Labels
- Part-of-Speech
- Adverbs

- Temporal Binding
How 'well'

- Skill Determination
  - Action Completion

Dual-

- DDLSTM
  - Multi-modal UDA
  - Retro-Actions

Audio-

- Multi-Verb Labels
  - Part-of-Speech
  - Adverbs

visual

Vision+

Temporal Binding

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
Skill Determination in Video

The Pros and Cons: Rank-Aware Temporal Attention

$$p_i > p_j$$

Disparity Loss
$$\sum_{(p_i,p_j) \in \Phi} \max(0, m_2 - (s^+(p_i) - s^+(p_j)) + (u(p_i) - u(p_j))$$

Ranking Loss
$$\sum_{(p_i,p_j) \in \Phi} \max(0, m - s^+(p_i) + s^+(p_j))$$

Ranking Loss
$$\sum_{(p_i,p_j) \in \Phi} \max(0, m - s^-(p_i) + s^-(p_j))$$

Ranking Loss
$$\sum_{(p_i,p_j) \in \Phi} \max(0, m - u(p_i) + u(p_j))$$

Disparity Loss
$$\sum_{(p_i,p_j) \in \Phi} \max(0, m_2 - (s^-(p_i) - s^-(p_j)) + (u(p_i) - u(p_j))$$

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

Low-skill Attention Module

Surgery

Apply Eyeliner

Origami

The Pros and Cons: Rank-Aware Temporal Attention

High-skill Attention Module

Dough Rolling

Origami

Drawing

The Pros and Cons: Rank-Aware Temporal Attention

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]
Fine-Grained Object Interactions

- Skill Determination
- Action Completion
- Dual-Vision+
- Language
- Audio-visual
- Temporal Binding
- How 'well'
- DDLSTM
- Multi-modal UDA
- Retro-Actions
- Multi-Verb Labels
- Part-of-Speech
- Adverbs
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

2. Regression-Based Voting
Action Completion Detection

2. Regression-Based Voting

Action Completion Detection

Action Completion Detection

Pre-V
V\textsuperscript{T}_R
C-C
R-R
R-C
C-R
GT

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
- How 'well'
- Dual-
- Audio-visual
- Vision+
- Language
- Temporal Binding
- Multi-Verb Labels
- Part-of-Speech
- Adverbs
- DDLSTM
- Multi-modal UDA
- Retro-Actions
Fine-Grained Object Interactions

- Skill Determination
- Action Completion
- DDLSTM
- Multi-modal UDA
- Retro-Actions

How 'well'

Audio-visual

Vision+ Language

- Temporal Binding
- Multi-Verb Labels
- Part-of-Speech
- Adverbs
Multi-Modal Domain Adaptation for Fine-Grained Action Recognition

with: Jonathan Munro

Multi-Modal Domain Adaptation for Fine-Grained Action Recognition

with: Jonathan Munro

Multi-Modal Domain Adaptation for Fine-Grained Action Recognition

Source-Only | Self-Supervision | Our Proposal

RGB

Flow

Dima Damen

## Multi-Modal Domain Adaptation for Fine-Grained Action Recognition

<table>
<thead>
<tr>
<th>Method</th>
<th>D2→D1</th>
<th>D3→D1</th>
<th>D1→D2</th>
<th>D3→D2</th>
<th>D1→D3</th>
<th>D2→D3</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>MM Source-only</td>
<td>42.5</td>
<td>44.3</td>
<td>42.0</td>
<td><strong>56.3</strong></td>
<td>41.2</td>
<td>46.5</td>
<td>45.5</td>
</tr>
<tr>
<td>AdaBN [29]</td>
<td>44.6</td>
<td>47.8</td>
<td>47.0</td>
<td>54.7</td>
<td>40.3</td>
<td>48.8</td>
<td>47.2</td>
</tr>
<tr>
<td>MMD [32]</td>
<td>43.1</td>
<td>48.3</td>
<td>46.6</td>
<td>55.2</td>
<td>39.2</td>
<td>48.5</td>
<td>46.8</td>
</tr>
<tr>
<td>MCD [45]</td>
<td>42.1</td>
<td>47.9</td>
<td>46.5</td>
<td>52.7</td>
<td>43.5</td>
<td>51.0</td>
<td>47.3</td>
</tr>
<tr>
<td>MM-SADA</td>
<td><strong>48.2</strong> ▶+5.7</td>
<td><strong>50.9</strong> ▶+6.6</td>
<td><strong>49.5</strong> ▶+7.5</td>
<td>56.1 ▼-0.2</td>
<td><strong>44.1</strong> ▶+2.9</td>
<td><strong>52.7</strong> ▶+6.3</td>
<td><strong>50.3</strong> ▶+4.8</td>
</tr>
<tr>
<td>Supervised target</td>
<td>62.8</td>
<td>62.8</td>
<td>71.7</td>
<td>71.7</td>
<td>74.0</td>
<td>74.0</td>
<td>69.5</td>
</tr>
</tbody>
</table>
Fine-Grained Object Interactions

- Skill Determination
- Action Completion
- Dual-
  - How 'well'
- Audio-visual
  - Temporal Binding
- Vision+
  - DDLSTM
  - Multi-modal UDA
  - Retro-Actions
- Language
  - Multi-Verb Labels
  - Part-of-Speech
  - Adverbs
Retro-actions

ICCV MDALC Workshop
Retro-actions

- **INVARIANT**
  - Moving [part] of [something]
  - Moving [part] of [something]

- **EQUIVARIANT**
  - Removing [something], revealing [something] behind
  - Putting [something] in front of [something]

- **IRREVERSIBLE**
  - Poking a stack of [something] so the stack collapses
  - Irreversible

Retro-actions – Zero-Shot Learning

ICCV MDALC Workshop

28 April 2020
Fine-Grained Object Interactions

- Skill Determination
- Action Completion

- Multi-Verb Labels
- Part-of-Speech
- Adverbs

- DDLSTM
- Multi-modal UDA
- Retro-Actions

How 'well'

Dual-

Audio-

Vision+

Language

• Temporal Binding
The Verbs Dilemma

with: Michael Wray

BMVC
The Verbs Dilemma

Open

The *Verbs* Dilemma

with: Michael Wray

The Verbs Dilemma

Open

Cut

The Verbs Dilemma

The Verbs Dilemma

Open

Cut

The Verbs Dilemma

• Action representations using a single verb is highly-ambiguous
  • Solution1: pre-selected non-overlapping verbs (SL)
    • run, walk, open, close
  • Solution2: Using nouns to disambiguate actions (V-N)
    • open-drawer, open-bottle, open-fridge
    • actions constrained to known nouns
  • Solution3: Multi-verb labels (ML, SAML)
    • open, hold, pull

The Verbs Dilemma

The Verbs Dilemma

Top 3 retrieved classes across all datasets.

Labelling Method can differentiate turn On/Off tap by pressing and by rotating.

Turn On/Off
Press
Rotate

Turn On/Off
Press
Rotate
Fine-Grained Object Interactions

- Skill Determination
- Action Completion

- Dual-
  - DDLSTM
  - Multi-modal UDA
  - Retro-Actions

- How 'well'
  - Audio-visual
  - Temporal Binding

- Vision+
  - Language
  - Multi-Verb Labels
  - Part-of-Speech
  - Adverbs
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

[pød]

[meat, ball, dough]

Verb Embedding

Noun Embedding

take

open

put

carrot
door

meat

ball
dough
Fine-Grained Action Retrieval

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

Individual Modality Feature Extraction

Video $x_i$: Feature Extraction $\nu_i$

Caption $y_i$: Feature Extraction $\nu_i$

I put the cup down on the counter

Verb: $\langle$Put$\rangle$  Noun: $\langle$Cup$\rangle$, $\langle$Counter$\rangle$
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

- Dual-Vision
- Audio-visual
- How 'well'

- Vision+
- Language
- Multi-Verb Labels
- Part-of-Speech
- Adverbs

- DDLSTM
- Multi-modal UDA
- Retro-Actions

- Temporal Binding
... if you turn the bowl upside down slowly they won't come out ...

... mix it well until it is completely dissolved ...

... you want to make sure you fill it up partially ...

... you want to dice it finely...
Action Modifiers: Learning from Adverbs in Instructional Videos

..start by **quickly rolling** our lemons...

Action Modifiers: Learning from Adverbs in Instructional Videos

... we're going to mix these up real quick...
Fine-Grained Object Interactions

- Skill Determination
- Action Completion
- Temporal Binding
- DDLSTM
- Multi-modal UDA
- Retro-Actions
- Multi-Verb Labels
- Part-of-Speech
- Adverbs
Audio-Visual Temporal Binding for Egocentric Action Recognition

Audio-Visual Temporal Binding for Egocentric Action Recognition

Multimodal Fusion

BN-Inception
avg pool 8x8
concat

BN-Inception
avg pool 7x7

BN-Inception
avg pool 7x7

x and y flow components

Audio-Visual Temporal Binding for Egocentric Action Recognition

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 Kazakos¹, Arsha Nagrani², Andrew Zisserman² and Dima Damen¹

¹University of Bristol, VIL, ²University 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]

Dima Damen

28 April 2020
Thank you…

For further info, datasets, code, publications…

http://dimadamen.github.io

@dimadamen

http://www.linkedin.com/in/dimadamen
Scaling Egocentric Vision: The EPIC-KITCHENS Dataset
peel potato
put lid on milk
pour contents into tupperware
tie the bag again
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 sink)
(dishwasher) (towel)

(kitchen) (oven) (towel)

(kitchen) (towel) (tofu)
454 200
OBJECT ANNOTATIONS
Annotations (4) – Verb and Noun Classes

- 123 verb classes
- 331 noun classes

| take, grab, pick, get, fetch, pick-up, ...

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.

\[
\hat{y}_{\text{verb}} = \text{open} \\
\hat{y}_{\text{noun}} = \text{oven} \\
\hat{y}_{\text{action}} = (\text{open, oven})
\]
## Action Recognition Challenge

### Seen Kitchens (S1)

<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 ▲ Noun ▲ Action ▲</td>
<td>Verb ▲ Noun ▲ Action ▲</td>
<td>Verb ▲ Noun ▲ Action ▲</td>
<td>Verb ▲ Noun ▲ Action ▲</td>
</tr>
<tr>
<td>1</td>
<td>action_banks</td>
<td>11</td>
<td>03/02/20</td>
<td></td>
<td>63.22 (7) 46.49 (5) 41.34 (1)</td>
<td>87.34 (9) 69.98 (5) 63.50 (1)</td>
<td>53.75 (5) 43.18 (4) 24.28 (1)</td>
<td>40.27 (8) 42.82 (5) 25.67 (1)</td>
</tr>
<tr>
<td>2</td>
<td>wasun</td>
<td>11</td>
<td>02/22/20</td>
<td>UTS_BAIDU</td>
<td>69.85 (1) 51.14 (1) 41.18 (2)</td>
<td>90.67 (2) 74.44 (1) 62.13 (2)</td>
<td>58.85 (4) 45.06 (1) 23.45 (4)</td>
<td>45.23 (4) 48.38 (1) 25.35 (1)</td>
</tr>
<tr>
<td>3</td>
<td>aptx4869lm</td>
<td>12</td>
<td>01/30/20</td>
<td>GT-WISC-MPI</td>
<td>68.51 (2) 49.96 (2) 38.75 (3)</td>
<td>89.33 (4) 72.30 (3) 58.99 (3)</td>
<td>51.04 (10) 44.00 (3) 23.70 (5)</td>
<td>43.70 (2) 47.32 (3) 23.92 (3)</td>
</tr>
<tr>
<td>4</td>
<td>weiyaowang</td>
<td>13</td>
<td>11/14/19</td>
<td></td>
<td>65.91 (4) 48.48 (3) 36.76 (4)</td>
<td>89.51 (3) 71.36 (3) 56.17 (6)</td>
<td>51.76 (8) 41.26 (5) 20.84 (2)</td>
<td>46.73 (4) 49.42 (2) 21.98 (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) 47.88 (4) 36.66 (5)</td>
<td>91.28 (1) 72.80 (2) 58.62 (4)</td>
<td>60.73 (3) 44.89 (2) 24.01 (2)</td>
<td>46.81 (1) 43.88 (5) 22.92 (4)</td>
</tr>
<tr>
<td>6</td>
<td>Sudhakaran</td>
<td>30</td>
<td>08/10/19</td>
<td>FBK_HuPBA</td>
<td>63.34 (6) 44.75 (6) 35.54 (6)</td>
<td>89.01 (5) 69.88 (6) 57.18 (5)</td>
<td>63.21 (5) 42.26 (5) 19.76 (7)</td>
<td>37.77 (12) 41.28 (7) 21.19 (6)</td>
</tr>
<tr>
<td>7</td>
<td>tnet</td>
<td>16</td>
<td>03/06/20</td>
<td></td>
<td>64.74 (6) 44.75 (6) 34.81 (6)</td>
<td>88.82 (7) 68.98 (6) 55.49 (5)</td>
<td>52.68 (7) 42.18 (6) 20.76 (6)</td>
<td>45.29 (3) 41.97 (6) 20.06 (7)</td>
</tr>
<tr>
<td>8</td>
<td>antoninofernari</td>
<td>1</td>
<td>07/19/19</td>
<td></td>
<td>56.93 (11) 43.05 (7) 33.06 (8)</td>
<td>85.68 (15) 67.12 (8) 55.32 (8)</td>
<td>50.42 (12) 39.84 (8) 18.91 (8)</td>
<td>37.82 (10) 38.11 (8) 19.12 (8)</td>
</tr>
<tr>
<td>9</td>
<td>cvg_uni_bonn</td>
<td>7</td>
<td>01/07/20</td>
<td></td>
<td>58.63 (9) 41.44 (8) 29.81 (7)</td>
<td>88.73 (9) 66.57 (11) 48.64 (9)</td>
<td>50.32 (13) 37.67 (9) 18.30 (9)</td>
<td>41.37 (6) 37.61 (9) 18.07 (10)</td>
</tr>
</tbody>
</table>
Evaluating Action Recognition Models

## Evaluating Action Recognition Models

<table>
<thead>
<tr>
<th>Model</th>
<th>GFLOP/s</th>
<th></th>
<th>Params (M)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RGB</td>
<td>Flow</td>
<td>RGB</td>
<td>Flow</td>
</tr>
<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: Uith, The Spoon, El Sol, 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 (egocentric) 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
- Head-mounted camera
- 96 hours of recording - Full HD, 60fps
- 11.5M frames
- Multi-language narrations
- 36,974 action segments
- 424,008 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 epicpa@eipa.bristol.ac.uk with the subject subscribe epic-community and a blank message body.