Egocentric Vision

Dr Dima Damen
Department of Computer Science
Short Bio

• 1998-2002 BSC in Computer Science
• 2002-2003 MSc in Distributed Multimedia Sys.
• 2006-2009 PhD in Computer Vision
Short Bio

- 2010-2012 Postdoc on EU-FP7 project
Short Bio

- 2013-2017 Assistant Prof in Computer Vision
- 2017- Associate Prof in Computer Vision
Egocentric Vision?

• Research interests: action and activity recognition

• Particularly centred around the perception of object interactions
Ego…

Ego… a person’s sense of self-esteem or self-importance

Egocentric vision… the wearer serves as the central reference point in the study of interesting entities: objects, actions, interactions and intentions
Ego…
Visual Sensing – the landscape
Visual Sensing – the landscape

Least Expensive

Most Expensive

Expensive
Visual Sensing – the landscape

Most Mobile!

Moveable

Least Static
Visual Sensing – the landscape

Most Wearable!  Hand-Held Wireless  Hand-Held Wired  Least Static

Wearable
Wearable?
Wearable?
Wearable?
Wearable?
Wearable?

- Hat-Mounted
- Head-Mounted ★
- Glass-Mounted ★
- Shoulder-Mounted
- Chest-Mounted ★
- Wrist-Mounted
- Belt-Mounted
- Ankle-Mounted
But why do we care about… hardware???

- OPV (Ordinal-Person Views)
  - FPV (First-Person View)
  - SPV (Second-Person View)
  - TPV (Third-Person View)
See for yourself!

- **Videos...**
Conclusions?

• Just another camera?
• Just a shaking camera?
Egocentric Vision

• The Unique Problems
  1. Camera Motion
  2. Mapping and Localisation
  3. Attention and Task-Relevance
  4. Object Interactions
  5. Multi-view Solutions

• The Unique Applications
  1. Video Summarisation
  2. Skill Determination
  3. Real-time solutions
The Unique Problems

1. Camera Motion
1. Camera Motion

• Two types of motion
  • Egomotion
  • Foreground motion
Ego-motion

rotating  approaching  receding
Ego-motion

- Detect to:
  - Use?
  - Remove?
Hyperlapse

- https://youtu.be/sA4Za3Hv6ng
The Unique Problems

2. Mapping and Localisation
Mapping and Localisation

- [https://youtu.be/ufBLu1VUQ-E](https://youtu.be/ufBLu1VUQ-E)
The Unique Problems

3. Attention and Task Relevance
Attention and Task Relevance
Attention and Task Relevance

• Attention in egocentric vision
  • Foreground segmentation
  • Hand-region segmentation
  • Gaze tracking
Quick introduction to human gaze

• Humans iterate between “fixations” and “saccades”
  • Fixation: short stops
  • Saccade: quick movements between fixations

• [https://youtu.be/pknohrsz4Qs](https://youtu.be/pknohrsz4Qs)
Quick introduction to human gaze
Quick introduction to human gaze
Quick introduction to human gaze

Land and Hayhoe (2001) In what ways do eye movements contribute to everyday activities? Vision Research
Quick introduction to human gaze

- The notion of fixation/saccade has recently inspired attention models in vision
The Unique Problems

3. Attention and Task Relevance

Case Study: You-Do, I-Learn
You-Do, I-Learn

• First-person view
• Offers a unique insight into ‘used’ or ‘attended-to’ objects
• How these objects have been used
Try it yourself
You-Do, I-Learn

• Q. How to ‘ground-truth’ objects that have been used?
• Q. How to ‘ground-truth’ how these objects have been used?

**BEOID**

- Ground-truth by written narration
- Released with dataset

```
pick the charger and plug it into the socket. Check that the screwdriver is powered by looking at the button. Pick the tape and place it in the box. Walk to the printer. Open the drawer to check the paper and press keys on the printer pad. Use the card to unlock the door
```
You Do, I Learn

• Discover used objects
• Discover how objects have been used
• Extract guidance videos
• Fully unsupervised
  • No prior knowledge of objects (number, size)
  • Static and moveable objects
Definition

Task-Relevant Object (TRO)

*an object, or part of an object, with which a person interacts during task performance*
Which Objects?

Discovering Task-Relevant Objects

Discovering Task-Relevant Objects

**Suggested Problem Formulation...**

- Given a sequence of egocentric images \( \{I_1, \ldots, I_T\} \)
- Collected from multiple operators around a common environment
- Automatically discover all task-relevant objects

\[
\{O_k; 1 \leq k \leq K\}
\]

\[
O_k = \{\Omega(I_t); 1 \leq t \leq T\}
\]

- **Assumption:** at most one task-relevant image part is present within each image
Discovering Task-Relevant Objects

Attention

Task-Relevant

Position

Hot Spot

Appearance

instances
categories
Discovering Task-Relevant Objects

- **Gaze**
- **SLAM**
- **RGB features**

**Task-Relevant**

**Hot Spot**

**instances**

**categories**

Discovering TROs

Discovering becomes a clustering task...

- Considers attention, position and appearance
- Unknown number of objects
Discovering Task-Relevant Objects
Discovering Task-Relevant Objects

Unsupervised Discovery of Objects and their Modes of Interaction Towards Video-Based Guidance. *Computer Vision and Image Understanding*
Discovering Task-Relevant Objects
Discovering Task-Relevant Objects

Action

with: Walterio Mayol-Cuevas
Teesid Leelasawassuk
Discovering Task-Relevant Objects

Discovering Task-Relevant Objects

Discovering Modes of Interaction

Modes of Interaction (MOI)

the different ways in which TROs are used
Discovering Modes of Interaction

Attention

Position

Appearance

Hot Spots

MOIs

Instances

Categories

Interactions

Motion

Discovering Modes of Interaction

• Motion
  • Video snippets for each discovered object
  • Descriptor per snippet
  • Clustering using DB-index
Discovering Modes of Interaction

Discovering Modes of Interaction

Open & get sugar

Put

Pick

Open door

Back to…. the goal…

You Do, I Learn - Demonstration

More info…

Project You-Do, I-Learn


The Unique Problems

4. Object Interactions
Action Recognition – an Introduction

• CNNs for Action Recognition

Dual-Stream Neural Networks
Action Recognition – an Introduction

- CNNs for Action Recognition
- Dual-Stream Neural Networks

Figure by: Will Price, BSc Project, University of Bristol
Egocentric Action Recognition

Figure from: Ma et al. Going Deeper into First-Person Activity Recognition. CVPR 2016
Egocentric Action Recognition

Figure from: Ma et al. Going Deeper into First-Person Activity Recognition. CVPR 2016
Egocentric Action Recognition

Figure from: Ma et al. Going Deeper into First-Person Activity Recognition. CVPR 2016
Visualising Learnt Models

- [http://youtu.be/4cZS39c7IL0](http://youtu.be/4cZS39c7IL0)
Action Recognition – an Introduction

- CNNs for Action Recognition
  Dual-Stream Neural Networks

![Diagram of Dual-Stream Neural Networks]

Figure by: Will Price, BSc Project, University of Bristol
Object Interactions – the Dilemma

pull

open

push

Object Interactions – the Dilemma

Object Interactions – the Dilemma
Object Interactions – the Dilemma

Object Interactions – the Dilemma

Open

Cut

Object Interactions – the Dilemma

Object Interactions – the Dilemma

Object Interactions – the Dilemma

• Verbs cannot be separated into classes with hard boundaries.

• Singular classes are not enough.
Learning Visual Actions Using Multiple Verb-Only Labels

- 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
    - How many verbs would be enough?
Learning Visual Actions Using Multiple Verb-Only Labels

- Soft-Assigned Multi-Label
  - Multi-label using verbs only
  - Each verb assigned a value between 0 and 1
  - Object agnostic
  - Trained with Sigmoid Binary Cross Entropy

Learning Visual Actions Using Multiple Verb-Only Labels

• Collected from AMT

• Annotators agree:
  • Relevant Verb -> Main action
  • Irrelevant Verb -> unrelated motion

• Annotators disagree:
  • Relevant motion but not the main action

Learning Visual Actions Using Multiple Verb-Only Labels

• Collected from AMT

SL
  • Majority Vote.
  • One-hot vector.

ML
  • Threshold of 0.5.
  • Binary Vector

SAML
  • Full Annotation.
  • Continuous Vector.

Learning Visual Actions Using Multiple Verb-Only Labels

Top 3 retrieved classes across all datasets.

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

Action Recognition – an Introduction

- CNNs for Action Recognition

Dual-Stream Neural Networks

Figure by: Will Price, BSc Project, University of Bristol
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

predicted class: take knife

Trespassing the Boundaries

**Trespassing the Boundaries**


Dima Damen

11 July 2019
The Rubicon Boundaries

- Labelling approach proposal for temporally consistent annotations
- Decomposes an object interaction into two phases:
  - *pre-actional* phase
  - *actional* phase

The Rubicon Boundaries

The Rubicon Boundaries

Cut pepper (GTEA Gaze+)

The Rubicon Boundaries

Dima Damen


with: Davide Moltisanti
Michael Wray
Walterio Mayol-Cuevas

Action

Conventional annotations — RB annotations

![Graph showing IoU comparison between conventional annotations and RB annotations for various actions](image)

**Actions**

- pick-up cup
- turn tap
- put cup
- press button
- take cup
- pick-up jar
- put jar
- open jar
- take spoon
- scoop jar
- stir cup
- wash cup
- scan card

MTurk
The Rubicon Boundaries

with: Davide Moltisanti
Michael Wray
Walterio Mayol-Cuevas

Action Recognition from a Single Timestamp

• Learning from Single timestamps
Action Recognition from a Single Timestamp

• Learning from Single timestamps
Action Recognition from a Single Timestamp

The Unique Problems

5. Multi-View Action Recognition
FPV with SPV

Input: paired egocentric videos

Egocentric video of person A

Egocentric video of person B

Multiple POV features of A

Displacement

Time

$f_A$: First-person POV feature of A

Multiple POV features of B

Displacement

Time

$f_{B \leftarrow A}$: Second-person POV feature of B

$f_{A \leftarrow B}$: Second-person POV feature of A

$f_B$: First-person POV feature of B

Figure from: Yonetani et al (2016). Recognizing Micro-Actions and Reactions From Paired Egocentric Videos. CVPR
FPV with TPV (top-view)

Figure from: Ardeshir and Borji (2016). Egocentric Meets Top-view CVPR
FPV with TPV (top-view)

Figure from: Ardeshir and Borji (2016). Ego2Top: Matching Viewers in Egocentric and Top-view Videos. ECCV
Egocentric Vision

• The Unique Problems
  1. Camera Motion
  2. Mapping and Localisation (ref tomorrow’s talk)
  3. Attention and Task-Relevance
  4. Object Interactions
  5. Multi-view Solutions

• The Unique Applications
  1. Video Summarisation
  2. Skill Determination
  3. Real-time solutions
The Unique Applications

1. Video Summarisation
Video Summarisation

- Fixations
- Highlight Detection
Egocentric Video Summarisation

- Object-Driven
Egocentric Video Summarisation

• Object-Driven

Figure from: Lu and Grauman (2013). Story-Driven Summarization for Egocentric Video. CVPR
Egocentric Video Summarisation

- Fixation-Driven with Constraints

Figure from: Xu et al (2015). Gaze-enabled Egocentric Video Summarization via Constrained Submodular Maximization. CVPR
Egocentric Video Summarisation

- Fixations from IMUs
The Unique Applications

2. Skill Determination

Assess relative skill for a collection of video sequences, applicable to a variety of tasks.

**Input:** Pairwise annotations of videos, indicating higher skill or no skill preference


Video 1 ~ Video 2


Surgery


Drawing


Dough-Rolling

Chopstick Using

\[
L_{rank1} = \sum_{(p_i, p_j) \in \Psi} \max(0, m - f(p_i) + f(p_j)) \quad (3)
\]

\[
L_{rank2} = \sum_{(p_i, p_j) \in \Psi} \sum_{k=1}^{N} \max(0, m - f_k(p_i) + f_k(p_j)) \quad (5)
\]

\[
L_{sim} = \sum_{(p_i, p_j) \in \Phi} \sum_{k=1}^{N} \max(0, |f(p_i) - f(p_j)| - m) \quad (7)
\]

\[
L_{rank3} = \beta L_{rank2} + (1 - \beta) L_{sim} \quad (8)
\]

<table>
<thead>
<tr>
<th>Method</th>
<th>Surgery</th>
<th>Dough-Rolling</th>
<th>Drawing</th>
<th>Chopstick-Using</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>S</td>
<td>T</td>
<td>TS</td>
<td>S</td>
</tr>
<tr>
<td>Siamese TSN with margin loss</td>
<td>64.7</td>
<td>72.8</td>
<td>69.1</td>
<td>77.6</td>
</tr>
<tr>
<td>+ splits</td>
<td>64.4</td>
<td>73.3</td>
<td>69.0</td>
<td>79.1</td>
</tr>
<tr>
<td>+ similarity loss</td>
<td>66.4</td>
<td>72.5</td>
<td>70.2</td>
<td>79.5</td>
</tr>
</tbody>
</table>

\[
\frac{1}{\sigma} \sum_{j=1}^{\sigma} \alpha f_s(p_{ij}) + (1 - \alpha) f_t(p_{ij})
\]


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

Best ← Worst

Dima Damen


Example Rankings

Sonic-Drawing task - part of new skill dataset

The Pros and Cons: Rank-Aware Temporal Attention

Best

Worst

The Pros and Cons: Rank-Aware Temporal Attention

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
\[ \sum_{(p_i, p_j) \in \Phi} \max(0, m_3 - (s^+(p_i) - s^-(p_j)) + (u(p_i) - u(p_j)) \]
 Novel Rank-Aware Loss

$$L^+_{rank} = \sum_{(p_i, p_j) \in P} \max(0, m - s^+(p_i) + s^+(p_j)) \quad (8)$$

$$L^-_{rank} = \sum_{(p_i, p_j) \in P} \max(0, m - s^-(p_i) + s^-(p_j)) \quad (9)$$

$$L^u_{rank} = \sum_{(p_i, p_j) \in P} \max(0, m - u(p_i) + u(p_j)) \quad (10)$$

$$L^+_{disp} = \sum_{(p_i, p_j) \in P} \max(0, m_2 - (s^+(p_i) - s^+(p_j)) + (u(p_i) - u(p_j))) \quad (11)$$

$$L_{rAware} = \sum_{(p_i, p_j) \in P} \max(0, m_3 - (s^+(p_i) - s^-(p_j)) + (u(p_i) - u(p_j))) \quad (12)$$

$$L_R = \sum_{i=\{+,-,u\}} L^i_{rank} + \sum_{i=\{+,-\}} L^i_{disp} + L_{rAware} \quad (13)$$
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

<table>
<thead>
<tr>
<th>Method</th>
<th>EPIC Skills</th>
<th>YouTube Skill</th>
</tr>
</thead>
<tbody>
<tr>
<td>Who’s Better [4]</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 [18]</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 Unique Applications

3. Real-time Solutions
Wearable (Systems)!

• On-the-cloud processing
• On-the-mobile processing
• Onboard processing!
Figure 1. System overview. The user asks the device to inform her about her current view of Arc de Triomphe, and the system responds with the most relevant description in its database.
You Do, I Learn – Google Glass Prototype

GlaciAR
Final Demo

Teesid Leelasawassuk, Dima Damen and Walterio Mayol
University of Bristol

October 2014
The need for large-scaled datasets…

with: Hazel Doughty
Giovanni Maria Farinella
Sanja Fidler
Antonino Furnari
Evangelos Kazakos
Davide Moltisanti
Jonathan Munro
Toby Perrett
Will Price
Michael Wray
TABLE 1: Comparative overview of relevant datasets. *action classes with > 50 samples

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Ego?</th>
<th>Non-Scripted?</th>
<th>Native Env?</th>
<th>Year</th>
<th>Frames</th>
<th>Sequences</th>
<th>Action Segments</th>
<th>Action Classes</th>
<th>Object BBs</th>
<th>Object Classes</th>
<th>Participants</th>
<th>No. Env.s</th>
</tr>
</thead>
<tbody>
<tr>
<td>EPIC-KITCHENS</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>2018</td>
<td>11.5M</td>
<td>432</td>
<td>39,596</td>
<td>149*</td>
<td>454,158</td>
<td>323</td>
<td>32</td>
<td>32</td>
</tr>
<tr>
<td>EGTEA Gaze+ [19]</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>2018</td>
<td>2.4M</td>
<td>86</td>
<td>10,325</td>
<td>106</td>
<td>0</td>
<td>0</td>
<td>32</td>
<td>1</td>
</tr>
<tr>
<td>BEOID [21]</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>2014</td>
<td>0.1M</td>
<td>58</td>
<td>1,488</td>
<td>34</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>GTEA Gaze+ [20]</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>2012</td>
<td>0.4M</td>
<td>35</td>
<td>3,371</td>
<td>42</td>
<td>0</td>
<td>0</td>
<td>13</td>
<td>1</td>
</tr>
<tr>
<td>ADL [23]</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>2012</td>
<td>1.0M</td>
<td>20</td>
<td>436</td>
<td>32</td>
<td>137,780</td>
<td>42</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>CMU [22]</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>2009</td>
<td>0.2M</td>
<td>16</td>
<td>516</td>
<td>31</td>
<td>0</td>
<td>0</td>
<td>16</td>
<td>1</td>
</tr>
<tr>
<td>VLOG [15]</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>2017</td>
<td>37.2M</td>
<td>114K</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10.7K</td>
<td>N/A</td>
</tr>
<tr>
<td>Charades [16]</td>
<td>×</td>
<td>×</td>
<td>✓</td>
<td>2016</td>
<td>7.4M</td>
<td>9,848</td>
<td>67,000</td>
<td>157</td>
<td>0</td>
<td>0</td>
<td>N/A</td>
<td>267</td>
</tr>
<tr>
<td>Breakfast [24]</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>2014</td>
<td>3.0M</td>
<td>433</td>
<td>3078</td>
<td>50</td>
<td>0</td>
<td>0</td>
<td>52</td>
<td>18</td>
</tr>
<tr>
<td>50 Salads [25]</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>2013</td>
<td>0.6M</td>
<td>50</td>
<td>2967</td>
<td>52</td>
<td>0</td>
<td>0</td>
<td>25</td>
<td>1</td>
</tr>
<tr>
<td>MPII Cooking 2  [26]</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>2012</td>
<td>2.9M</td>
<td>273</td>
<td>14,105</td>
<td>88</td>
<td>0</td>
<td>0</td>
<td>30</td>
<td>1</td>
</tr>
</tbody>
</table>
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>28,588</td>
<td>28,561</td>
<td>326,298</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>97,865</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>

**15 Most Frequent Object Classes**

<table>
<thead>
<tr>
<th></th>
<th>pan</th>
<th>plate</th>
<th>bowl</th>
<th>onion</th>
<th>tap</th>
<th>pot</th>
<th>knife</th>
<th>spoon</th>
<th>meat</th>
<th>food</th>
<th>potato</th>
<th>cup</th>
<th>pasta</th>
<th>cupboard</th>
<th>lid</th>
<th>few-shot</th>
<th>many-shot</th>
<th>all</th>
</tr>
</thead>
<tbody>
<tr>
<td>IoU &gt; 0.05</td>
<td>74.00</td>
<td>72.61</td>
<td>71.50</td>
<td>60.72</td>
<td>84.44</td>
<td>69.97</td>
<td>44.03</td>
<td>40.93</td>
<td>29.65</td>
<td>58.52</td>
<td>62.82</td>
<td>53.30</td>
<td>78.39</td>
<td>51.95</td>
<td>62.77</td>
<td>9.71</td>
<td>49.80</td>
<td>38.23</td>
</tr>
<tr>
<td>IoU &gt; 0.5</td>
<td>67.60</td>
<td>66.21</td>
<td>65.98</td>
<td>39.96</td>
<td>73.80</td>
<td>64.71</td>
<td>28.80</td>
<td>23.89</td>
<td>20.75</td>
<td>49.85</td>
<td>55.48</td>
<td>42.99</td>
<td>69.75</td>
<td>29.20</td>
<td>58.48</td>
<td>6.98</td>
<td>36.50</td>
<td>28.06</td>
</tr>
<tr>
<td>IoU &gt; 0.75</td>
<td>21.94</td>
<td>44.60</td>
<td>39.48</td>
<td>3.52</td>
<td>25.83</td>
<td>19.67</td>
<td>3.42</td>
<td>2.59</td>
<td>5.27</td>
<td>15.78</td>
<td>13.18</td>
<td>8.00</td>
<td>24.53</td>
<td>4.05</td>
<td>26.51</td>
<td>0.36</td>
<td>8.73</td>
<td>6.50</td>
</tr>
<tr>
<td>IoU &gt; 0.05</td>
<td>75.94</td>
<td>87.36</td>
<td>72.72</td>
<td>47.61</td>
<td>78.14</td>
<td>75.92</td>
<td>55.51</td>
<td>41.28</td>
<td>71.59</td>
<td>38.61</td>
<td>N/A</td>
<td>80.58</td>
<td>53.88</td>
<td>58.40</td>
<td></td>
<td>6.00</td>
<td>51.71</td>
<td>40.61</td>
</tr>
<tr>
<td>IoU &gt; 0.5</td>
<td>62.88</td>
<td>84.86</td>
<td>68.61</td>
<td>32.18</td>
<td>59.75</td>
<td>62.86</td>
<td>39.60</td>
<td>27.52</td>
<td>53.54</td>
<td>35.47</td>
<td>N/A</td>
<td>39.19</td>
<td>76.27</td>
<td>32.54</td>
<td>49.36</td>
<td>5.32</td>
<td>36.27</td>
<td>28.57</td>
</tr>
<tr>
<td>IoU &gt; 0.75</td>
<td>14.56</td>
<td>62.82</td>
<td>38.44</td>
<td>2.25</td>
<td>4.89</td>
<td>14.91</td>
<td>3.85</td>
<td>1.51</td>
<td>9.36</td>
<td>8.10</td>
<td>N/A</td>
<td>7.60</td>
<td>43.30</td>
<td>5.61</td>
<td>25.48</td>
<td>0.18</td>
<td>9.05</td>
<td>7.04</td>
</tr>
</tbody>
</table>
### TABLE 6: Baseline results for the action recognition challenge

<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>S1</td>
<td>RGB</td>
<td>45.25</td>
<td>35.78</td>
<td>18.91</td>
</tr>
<tr>
<td></td>
<td>FLOW</td>
<td>43.27</td>
<td>17.92</td>
<td>09.10</td>
</tr>
<tr>
<td></td>
<td>FUSION</td>
<td>47.36</td>
<td>36.05</td>
<td>19.44</td>
</tr>
<tr>
<td>S2</td>
<td>RGB</td>
<td>35.96</td>
<td>21.74</td>
<td>09.96</td>
</tr>
<tr>
<td></td>
<td>FLOW</td>
<td>40.56</td>
<td>14.91</td>
<td>07.28</td>
</tr>
<tr>
<td></td>
<td>FUSION</td>
<td>39.67</td>
<td>22.33</td>
<td>10.84</td>
</tr>
</tbody>
</table>

### TABLE 7: Sample baseline action recognition per-class metrics (using fusion)

<table>
<thead>
<tr>
<th></th>
<th>put</th>
<th>take</th>
<th>wash</th>
<th>open</th>
<th>close</th>
<th>cut</th>
<th>mix</th>
<th>pour</th>
<th>move</th>
<th>turn-on</th>
<th>remove</th>
<th>turn-off</th>
<th>throw</th>
<th>dry</th>
<th>peel</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>RECALL</td>
<td>65.32</td>
<td>51.01</td>
<td>80.45</td>
<td>60.98</td>
<td>27.13</td>
<td>74.27</td>
<td>52.63</td>
<td>24.87</td>
<td>00.00</td>
<td>05.63</td>
<td>01.58</td>
<td>03.67</td>
<td>10.11</td>
<td>29.73</td>
</tr>
<tr>
<td></td>
<td>PRECISION</td>
<td>35.62</td>
<td>41.24</td>
<td>63.17</td>
<td>72.67</td>
<td>72.46</td>
<td>69.38</td>
<td>69.52</td>
<td>66.20</td>
<td>-</td>
<td>53.33</td>
<td>66.67</td>
<td>50.00</td>
<td>56.25</td>
<td>88.00</td>
</tr>
<tr>
<td>S2</td>
<td>RECALL</td>
<td>64.16</td>
<td>48.03</td>
<td>87.76</td>
<td>42.06</td>
<td>15.10</td>
<td>45.69</td>
<td>35.85</td>
<td>06.06</td>
<td>00.00</td>
<td>00.00</td>
<td>00.81</td>
<td>00.00</td>
<td>00.00</td>
<td>00.00</td>
</tr>
<tr>
<td></td>
<td>PRECISION</td>
<td>30.19</td>
<td>30.46</td>
<td>67.79</td>
<td>57.31</td>
<td>61.54</td>
<td>85.48</td>
<td>65.52</td>
<td>40.00</td>
<td>-</td>
<td>00.00</td>
<td>100.00</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

- Challenges open on CodaLab – 9 Sep
- First Challenge Results in CVPR 2019
- EPIC@CVPR2019
- ActivityNet@CVPR2019
Given a trimmed action segment: \((t_{\text{start}}, t_{\text{stop}})\)
classify the action within.

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

\[
\hat{y}_{\text{action}} = (\text{open, oven})
\]

Dima Damen
University of BRISTOL
11 July 2019
142

with: Hazel Doughty
Giovanni Maria Farinella
Sanja Fidler
Antonino Furnari
Evangelos Kazakos
Davide Moltisanti
Jonathan Munro
Toby Perrett
Will Price
Michael Wray

Dima Damen
11 July 2019
Interested in More?

• Egocentric Perception, Interaction and Computing (EPIC) Workshop Series
  • ECCV 2016 (Amsterdam)
  • ICCV 2017 (Venice)
  • ECCV 2018 (Munich)
  • CVPR 2019 (Long Beach)
  • ICCV 2019 (Seoul) – 5th edition!
    • Paper deadline: 26 July 2019
    • Abstract submission: 5 August 2019 (ongoing work)
Interested in More?

• Subscribe to the newly introduced mailing list: epic-community@bristol.ac.uk

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

Davide
Mike
Will
Toby
Hazel
Jonny
Alessandro

Young
Evangelos
Thank you…

For further info, datasets, code, publications…

http://www.cs.bris.ac.uk/~damen

@dimadamen

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