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

Least

Most

Expensive

Visual Sensing – the landscape

Expensive
Wearable?

Dima Damen

5 July 2018
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

- Detect to:
  - Use?
  - Remove?
Hyperlapse

- https://youtu.be/sA4Za3Hv6ng

The Unique Problems
2. Mapping and Localisation

Mapping and Localisation
- https://youtu.be/uf8Lu1VUQ-E
The Unique Problems
3. Attention and Task Relevance

Attention and Task Relevance

• What is attention?
  • Non-Egocentric Attention Models (→ Saliency)

Figure: Zoya Bylinskii & Tilke Judd, ECCV 2016 Tutorial
http://saliency.mit.edu/ECCVTutorial/newDirectionsInSaliency.pdf
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
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
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

- Suggested Problem Formulation...
  - Given a sequence of egocentric images \( \{I_1, ..., I_T\} \)
  - 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

Gaze
Task-Relevant

SLAM
Hot
Spot
RGB features

instances
categories

Discovering becomes a clustering task...
• Considers attention, position and appearance
• Unknown number of objects
Discovering Task-Relevant Objects

Discovering Modes of Interaction
**Definition**

**Modes of Interaction (MOI)**

*the different ways in which TROs are used*

---

**Discovering Modes of Interaction**

- **Attention**
- **Position**
- **Hot Spots**
- **Appearance**
- **MOIs**
- **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

Open & get sugar
Put
Pick
Open door

Back to.... the goal...
You Do, I Learn - Demonstration

More info...

The Unique Problems
4. Object Interactions
Action Recognition – an Introduction

• CNNs for Action Recognition
  1. Dual-Stream Neural Networks

Egocentric Action Recognition

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

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

Action Recognition – an Introduction

- CNNs for Action Recognition
  1. Dual-Stream Neural Networks

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

- pull
- open
- push


Object Interactions – the Dilemma

• Verbs cannot be separated into classes with hard boundaries.

• Rather the boundaries are more nuanced – what is correct in one video is incorrect for another.

• Singular classes are not enough.
Towards an Unequivocal Representation of Actions

Top 3 retrieved classes across all datasets.

- Turn On/Off Press Rotate
- Turn On/Off Press Rotate

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

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++
- Predicted class: talk

University of BRISTOL
Dima Damen

Trespassing the Boundaries

with: Davide Moltisanti
Michael Wray

Walterio Mayol-Cuevas

Action Labelling approach proposal for temporally consistent annotations

 Decomposes an object interaction into two phases:
  - pre-actional phase
  - actional phase
The Rubicon Boundaries

Dima Damen

5 July 2018

with:
Davide Moltisanti
Michael Wray
Walterio Mayol-Cuevas

Action


Cut pepper (UTA Base++)

Conventional annotations vs. BB annotations

Merk

Actions
The Rubicon Boundaries

Actions

More info...

Visualising Learnt Models

- BEOID EBP videos:
  - http://youtu.be/4cZS39c7IL0
The Unique Problems
5. Multi-View Action Recognition

FPV with SPV

FPV with TPV (top-view)
FPV with TPV (top-view)

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

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 vs Video 2

1. Surgery
2. Dough-Rolling
3. Drawing
4. Chopstick-Using

\[ L_{\text{rank1}} = \sum_{(p_n, p_m) \in A} \max(0, m - f(p_n) + f(p_m)) \]  \hspace{1cm} (3)
\[ L_{\text{rank2}} = \sum_{(p_n, p_m) \in A} \max(0, m - f_n(p_n) + f_m(p_m)) \]  \hspace{1cm} (5)
\[ L_{\text{max}} = \sum_{(p_n, p_m) \in A} \max(0, [f_n(p_n) - f_m(p_m)] - m) \]  \hspace{1cm} (7)
\[ L_{\text{max}} = \beta L_{\text{rank2}} + (1 - \beta) L_{\text{max}} \]  \hspace{1cm} (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>S</td>
<td>T</td>
</tr>
<tr>
<td>Score T(N) with stategy loss</td>
<td>64.4</td>
<td>72.3</td>
<td>70.0</td>
<td>79.1</td>
</tr>
<tr>
<td></td>
<td>S</td>
<td>T</td>
<td>S</td>
<td>T</td>
</tr>
<tr>
<td>+ strategy loss</td>
<td>64.4</td>
<td>72.3</td>
<td>70.0</td>
<td>79.1</td>
</tr>
</tbody>
</table>

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

Accuracy (%)

- Surgery
- Dough-Rolling
- Drawing
- Chopstick-Using

Example Rankings

Lowest

Highest

Sonic-Drawing task - part of new skill dataset

More info...

The Unique Applications

3. Real-time Solutions
Wearable (Systems)!

- On-the-cloud processing
- On-the-mobile processing
- Onboard processing!

Connecting-to-the-cloud

![Image of a system overview.](image)

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.
Arxiv. https://epic-kitchens.github.io

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 dataset, action classes with ≥ 3 samples**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>KITTI-360</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>KITTI-360-200</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Epic-Kitchens</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Splits</th>
<th># Videos</th>
<th># Scenes</th>
<th># Frames</th>
<th>Segmented Scenes</th>
<th>Segmented Video</th>
<th>Recognition Basis</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>109</td>
<td>2,347</td>
<td>45,582</td>
<td>1,375</td>
<td>1,375</td>
<td>67.965</td>
</tr>
<tr>
<td>S2</td>
<td>57</td>
<td>1,247</td>
<td>27,069</td>
<td>1,102</td>
<td>1,102</td>
<td>67.785</td>
</tr>
</tbody>
</table>

**TABLE 6: Baseline results for the action recognition challenge**

<table>
<thead>
<tr>
<th>Model</th>
<th>Retrieval 1</th>
<th>Retrieval 2</th>
<th>Retrieval 3</th>
<th>Retrieval 4</th>
<th>Retrieval 5</th>
<th>Retrieval 6</th>
<th>Retrieval 7</th>
<th>Retrieval 8</th>
<th>Retrieval 9</th>
</tr>
</thead>
<tbody>
<tr>
<td>KITTI-360</td>
<td>65.2</td>
<td>65.0</td>
<td>64.9</td>
<td>65.0</td>
<td>65.0</td>
<td>64.9</td>
<td>65.0</td>
<td>65.0</td>
<td>65.0</td>
</tr>
<tr>
<td>Epic-Kitchens</td>
<td>66.0</td>
<td>65.8</td>
<td>65.7</td>
<td>65.8</td>
<td>65.8</td>
<td>65.7</td>
<td>65.8</td>
<td>65.8</td>
<td>65.8</td>
</tr>
</tbody>
</table>

**TABLE 7: Sample-based action recognition per-class metrics using (3x3) kernels**

<table>
<thead>
<tr>
<th>Model</th>
<th>Positive IOU</th>
<th>True POSITIVE</th>
<th>False POSITIVE</th>
<th>True NEGATIVE</th>
<th>False NEGATIVE</th>
</tr>
</thead>
<tbody>
<tr>
<td>KITTI-360</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
</tr>
<tr>
<td>Epic-Kitchens</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
</tr>
</tbody>
</table>
Interactive Conclusions

• Fill in the blanks:
  • Egocentric vision is ------------------
  • Pick up an action (e.g. open door). Draw a sketch of how it looks like from FPV and TPV
  • The biggest challenge (in your opinion) in egocentric vision is ------------------
  • The most interesting problem (to you) in egocentric vision is ------------------

Interested in More?

• Egocentric Perception, Interaction and Computing (EPIC) Workshop Series
  • ECCV 2016 (Amsterdam)
  • ICCV 2017 (Venice)
  • ECCV 2018 (Munich)
    • Paper deadline: Tomorrow!
    • Abstract submission till 23rd of July (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
Thank you…

For further info, datasets, code, publications…

http://www.cs.bris.ac.uk/~damen
@dimadamen
http://www.linkedin.com/in/dimadamen