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- Assistant Prof in Computer Vision
Egocentric Vision?

• Research interests: action and activity recognition
• Particularly centred around the viewpoint or the perspective
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
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?

[Images: various wearable devices and VR setups]
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 (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 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

3. Attention and Task Relevance
Attention and Task Relevance

• What is attention?

• Non-Egocentric Attention Models (→ Saliency)

DUT-OMRON dataset
http://saliencydetection.net/dut-omron
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 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

• [Video Link](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

Figure: Lei Ba et al (2015). Multiple Object Recognition with Visual Attention. ICLR
Quick introduction to human gaze

Gaze Information to prime Object Detection

Dima Damen, Osian Haines, Andrew Calway and Walterio Mayol-Cuevas.

Object detection is based on the paper:
Dima Damen, Pished Bunnun, Andrew Calway and Walterio Mayol-Cuevas.
*Real-time Learning and Detection of 3D Texture-less Objects: A Scalable Approach.*
British Machine Vision Conference (BMVC), 2012. [Best Poster Paper]

Jan. 2013
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
BEOID Dataset

- Released July 2014
- Wearable gaze tracker (ASL Mobile Eye XG)
- 6 locations: kitchen, workspace, printer, corridor with locked door, cardiac gym and weight-lifting machine
- 5 operators (2 sequences each)
BEOID Dataset

• Q. How to ‘ground-truth’ objects that have been used?
• Q. How to ‘ground-truth’ how these objects have been used?
Try it yourself
• 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
Task-Relevant Object (TRO)

an object, or part of an object, with which a person interacts during task performance
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
- HOG + BOW

Hot Spot

instances

categories
Discovering TROs

Discovering becomes a clustering task...

- Considers attention, position and appearance
- Unknown number of objects
- Davies-Bouldin (DB) index
- K-Means vs Spectral
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

with: Walterio Mayol-Cuevas
Teesid Leelasawassuk

6 July 2017
Discovering Task-Relevant Objects
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

Hot Spots

instances

MOIs

Interactions

Motion

Appearance

categories
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

You Do, I Learn – Google Glass Prototype
Activity

Task Monitoring - 2017

with: Longfei Chen
Kazuaki Kondo
Yuichi Nakamura
Walterio Mayol-Cuevas

Project You-Do, I-Learn


The Unique Problems

4. Object Interactions
Action Recognition – an Introduction

• CNNs for Action Recognition

1. Dual-Stream Neural Networks
Action Recognition – an Introduction

- CNNs for Action Recognition

1. Dual-Stream Neural Networks

Figure by: Will Price, BSc Project, University of Bristol
Action Recognition – an Introduction

- CNNs for Action Recognition

1. Dual-Stream Neural Networks

Figure by: Will Price, BSc Project, University of Bristol
# Action Recognition – an Introduction

<table>
<thead>
<tr>
<th>Dataset</th>
<th>N. of <em>gt</em> segments</th>
<th>N. of <em>gen</em> segments</th>
<th>Classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>BEOID [1]</td>
<td>742</td>
<td>16691</td>
<td>34</td>
</tr>
<tr>
<td>GTEA Gaze+ [6]</td>
<td>1141</td>
<td>22221</td>
<td>42</td>
</tr>
<tr>
<td>CMU [2]</td>
<td>450</td>
<td>26160</td>
<td>31</td>
</tr>
</tbody>
</table>

[1] Reference 1
[2] Reference 2
Action Recognition – an Introduction

Length distribution

- BEOID
- GTEA Gaze+
- CMU

Percentage of videos

Length (seconds)

0 5 10 15 20 25

0.5 1 1.5 2.5 3 4 5 6 9 15 20 30 40 60
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

(a) GTEA 71 classes  (b) Gaze 40 classes  (c) Gaze+ 44 classes
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
Object Interactions – the Dilemma

- pull
- open
- push
Object Interactions – the Dilemma

with: Michael Wray
Davide Moltisanti
Walterio Mayol-Cuevas

Dima Damen
6 July 2017
Object Interactions – the Dilemma
Object Interactions – the Dilemma

with: Michael Wray
Davide Moltisanti
Walterio Mayol-Cuevas
Object Interactions – the Dilemma

Open

Cut

with: Michael Wray
Davide Moltisanti
Walterio Mayol-Cuevas
Object Interactions – the Dilemma

with: Michael Wray
Davide Moltisanti
Walterio Mayol-Cuevas
Object Interactions – the Dilemma

Open

Cut

with: Michael Wray
Davide Moltisanti
Walterio Mayol-Cuevas

Dima Damen
6 July 2017
Object Interactions – the Dilemma

Open  Take  Cut

with: Michael Wray
Davide Moltisanti
Walterio Mayol-Cuevas
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.
Visualising Learnt Models

• BEOID EBP videos:
  https://youtu.be/Fu7Db7Pau_A
  https://youtu.be/4aDwQ-a3M68
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

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

Multiple POV features of B

Displacement

Time

$f_B$: First-person POV feature of B

$f_{B\leftarrow A}$: Second-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)
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

- Object-Driven

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

- Fixation-Driven with Constraints
Egocentric Video Summarisation

• Fixations from IMUs
The Unique Applications

2. Skill Determination
Who’s Better?

with: Hazel Doughty
Walterio Mayol-Cuevas

Dima Damen
6 July 2017
Who’s Best?

Novice

Expert

Dima Damen

Behaviours

Spatial TSN

Dima Damen

With: Hazel Doughty
Walterio Mayol-Cuevas

Spatial TSN

f(p_i)

Spatial TSN

Ranking loss layer

\[
\min \sum_{(p_i, p_j) \in \Psi} \max(0, 1 - f(p_i) + f(p_j))
\]

Dima Damen

with: Hazel Doughty
Walterio Mayol-Cuevas

6 July 2017

Dima Damen

University of BRISTOL

104


<table>
<thead>
<tr>
<th>Task</th>
<th>Videos</th>
<th>Max Pairs</th>
<th>#Consistent Pairs</th>
<th>%Consistent Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surgery (Knot Tying)</td>
<td>36</td>
<td>630</td>
<td>596</td>
<td>95%</td>
</tr>
<tr>
<td>Surgery (Needle Passing)</td>
<td>28</td>
<td>378</td>
<td>362</td>
<td>96%</td>
</tr>
<tr>
<td>Surgery (Suturing)</td>
<td>39</td>
<td>741</td>
<td>701</td>
<td>95%</td>
</tr>
<tr>
<td>Dough-Rolling</td>
<td>33</td>
<td>528</td>
<td>181</td>
<td>34%</td>
</tr>
<tr>
<td>Drawing (Sonic)</td>
<td>20</td>
<td>190</td>
<td>118</td>
<td>62%</td>
</tr>
<tr>
<td>Drawing (Hand)</td>
<td>20</td>
<td>190</td>
<td>129</td>
<td>68%</td>
</tr>
<tr>
<td>Chopstick-Using</td>
<td>40</td>
<td>780</td>
<td>536</td>
<td>69%</td>
</tr>
</tbody>
</table>


<table>
<thead>
<tr>
<th>Task</th>
<th>Siamese TSN</th>
<th>Siamese TSN with data augmentation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Spatial</td>
<td>Temporal</td>
</tr>
<tr>
<td>Surgery</td>
<td>66.5%</td>
<td>74.4%</td>
</tr>
<tr>
<td>Dough-Rolling</td>
<td>73.9%</td>
<td>76.7%</td>
</tr>
<tr>
<td>Drawing</td>
<td>75.6%</td>
<td>76.5%</td>
</tr>
<tr>
<td>Chopstick-Using</td>
<td>67.7%</td>
<td>67.4%</td>
</tr>
</tbody>
</table>

![Graphs showing accuracy for different tasks]

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

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

![Graph showing the accuracy as a function of \(\alpha\) for different activities: Surgery, Dough-Rolling, Drawing, and Chopstick-Using.](image)

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

Surgery

Dough-Rolling

Drawing

Chopstick-Using

- Spatial
- Temporal
- Two-Stream

Consensus Function

Consensus Function

Consensus Function

Consensus Function

Accuracy (%)

Example Rankings

Lowest

Newly recorded Sonic-Drawing task

Highest
More info...

Project Who's Better, Who's Best: Skill Determination in Video

Video
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-Cuevas
University of Bristol

October 2014
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 – this October)
    • Paper deadline expired
    • Abstract submission still open till Sep
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
Thank you…

For further info, datasets, code, publications…

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

@dimadamen

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