Understanding real-world scenes for human-like machine perception

Armin Mustafa
Royal Academy of Engineering Research Fellow
CVSSP, University of Surrey
Difference between humans and machines?

<table>
<thead>
<tr>
<th>Humans:</th>
<th>Machines:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crowd</td>
<td>?</td>
</tr>
<tr>
<td>Station</td>
<td>?</td>
</tr>
<tr>
<td>Indoor</td>
<td>?</td>
</tr>
<tr>
<td>Dance</td>
<td>?</td>
</tr>
<tr>
<td><strong>Location of objects</strong></td>
<td>?</td>
</tr>
</tbody>
</table>
How humans process data?

- We “Humans” perceive world in 3D?
  - No that’s not right
  - We live dynamic world not static world

Static scene – does not change with time

Dynamic scene elements change with time like Humans, Cars etc.
How humans process data?

- We perceive the world in 4D – which is 3D in time.
  - That’s how we detect actions
  - That’s how we interpret gestures

Creating machine interpretable 4D data from videos is called **4D vision**
4D Vision

Spatio – Temporally Coherent Models from Video

- No prior
- Moving cameras
- 3D Reconstruction
- Registration
- 4D scene reconstruction and segmentation
Why 4D Vision?

- Real ‘dynamic world’ is inherently 4D – 3D in time
- Modelling & understanding the real world

Analysis of human motion
Realistic interactive media production
Robust human-computer interaction
<table>
<thead>
<tr>
<th>Existing technology</th>
<th>4D Vision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Large setups with multiple sensors</td>
<td>Minimal setup with RGB cameras</td>
</tr>
<tr>
<td>Large amount of data</td>
<td>Small amount of data</td>
</tr>
<tr>
<td>Constrained environment</td>
<td>Challenging scenes</td>
</tr>
<tr>
<td>1-2, static or rigid objects</td>
<td>Multiple moving objects</td>
</tr>
<tr>
<td>Static cameras</td>
<td>Moving cameras</td>
</tr>
<tr>
<td>Manual user interactions</td>
<td>Automatic</td>
</tr>
<tr>
<td>High cost</td>
<td>Low cost</td>
</tr>
</tbody>
</table>
Results

Original videos

Semantic co-segmentation

Semantic reconstruction

4D dense points

Frame 1
Results

**Input videos**

**Semantic reconstruction**

**Semantic co-segmentation**

Monday, September 2, 2019