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?

Humans:	Machines:
Crowd	?
Station	?
Indoor	?
Dance	?
Location of objects	?

How humans process data?

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

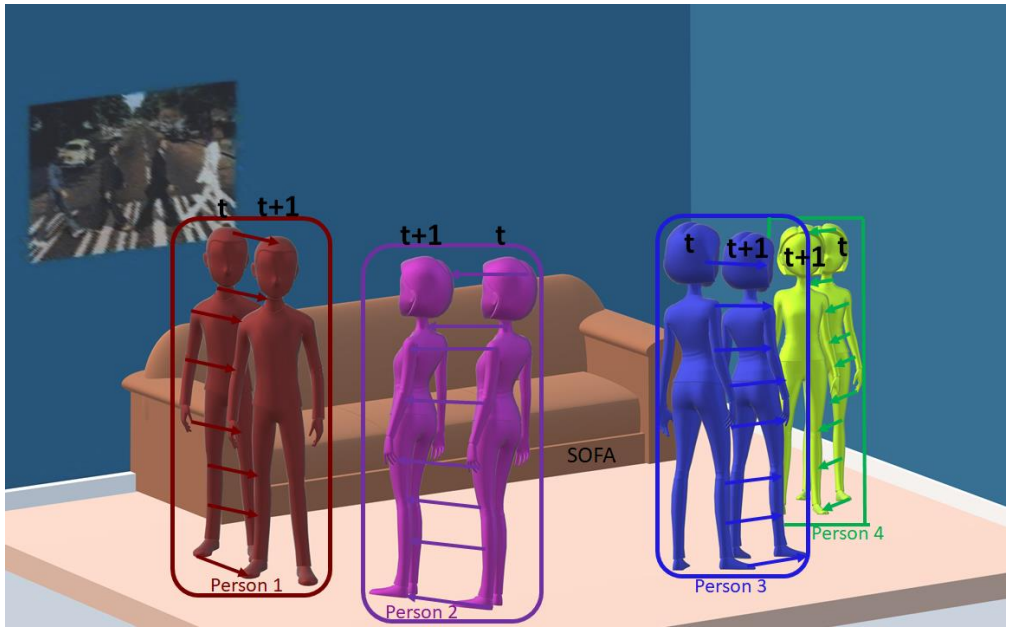
Static scene – does not change with time

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

Single-view Video

Multi-view Videos

Framework

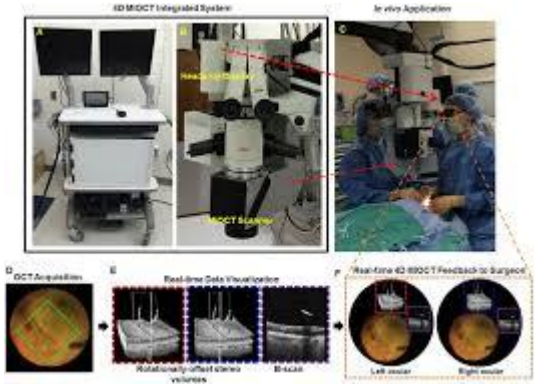
- No prior
- Moving cameras
- 3D Reconstruction
- Registration
- 4D scene reconstruction and segmentation

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

Existing systems vs 4D Vision

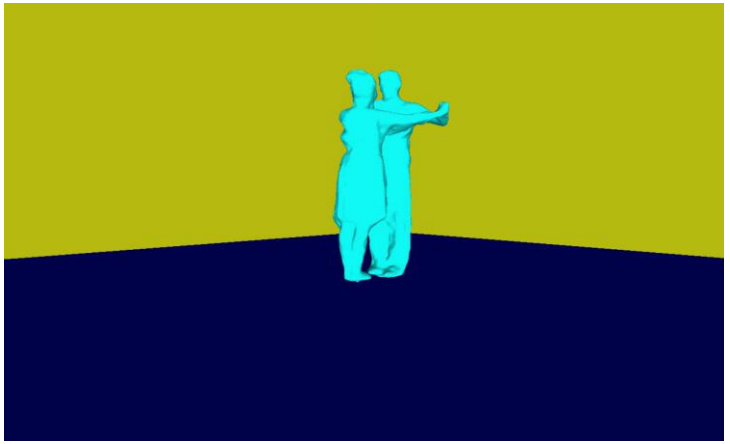
UNIVERSITY OF SURREY

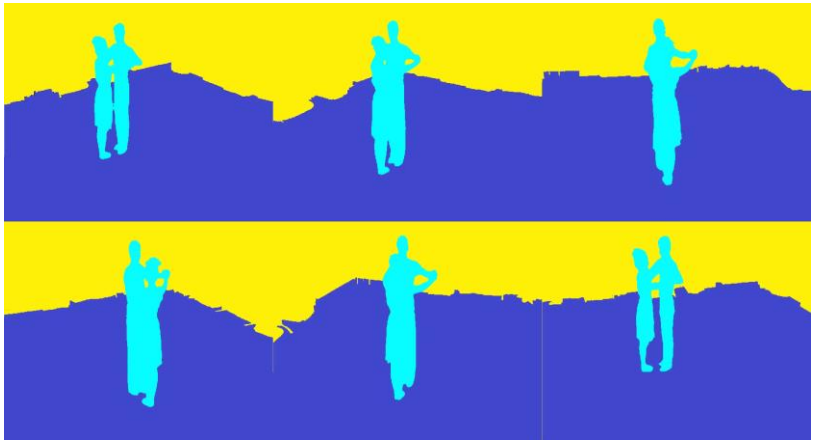
✗ Existing technology	✓ 4D Vision
Large setups with multiple sensors	Minimal setup with RGB cameras
Large amount of data	Small amount of data
Constrained environment	Challenging scenes
1-2, static or rigid objects	Multiple moving objects
Static cameras	Moving cameras
Manual user interactions	Automatic
High cost	Low cost

Results

Original videos

Semantic reconstruction





Semantic co-segmentation

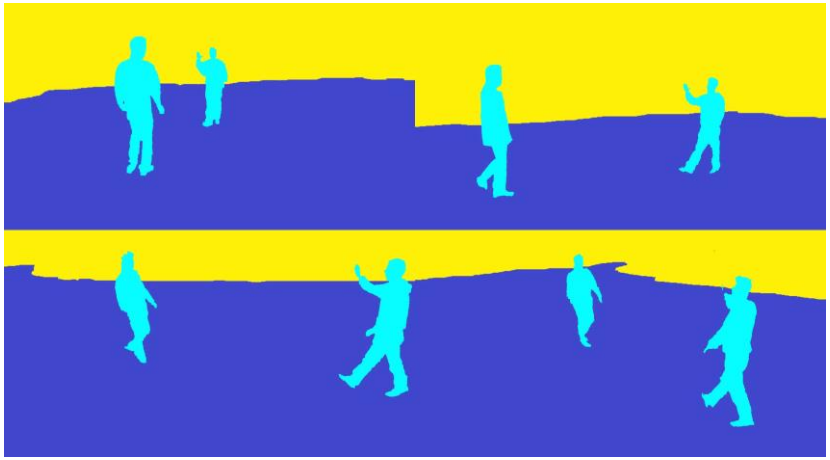
4D dense points

Frame 1

Results

Input videos

Semantic reconstruction



Semantic co-segmentation