visuAAL

Privacy-Aware and Acceptable Video-Based Technologies and Services for Active and Assisted Living

Understanding Human Behaviour With Wearable Cameras Based on Information From the Human Hand

ESR 10 – Final Project Meeting

Alicante, Spain
17 June, 2024

Wiktor Mucha
Computer Vision Lab
TU Wien

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 861091".

Universitat d'Alacant
Universidad de Alicante
Project Coordinator

RWTH AACHEN UNIVERSITY

Stockholm University

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

TU WIEN

Introduction

Egocentric vision and hand pose estimation

Introduction

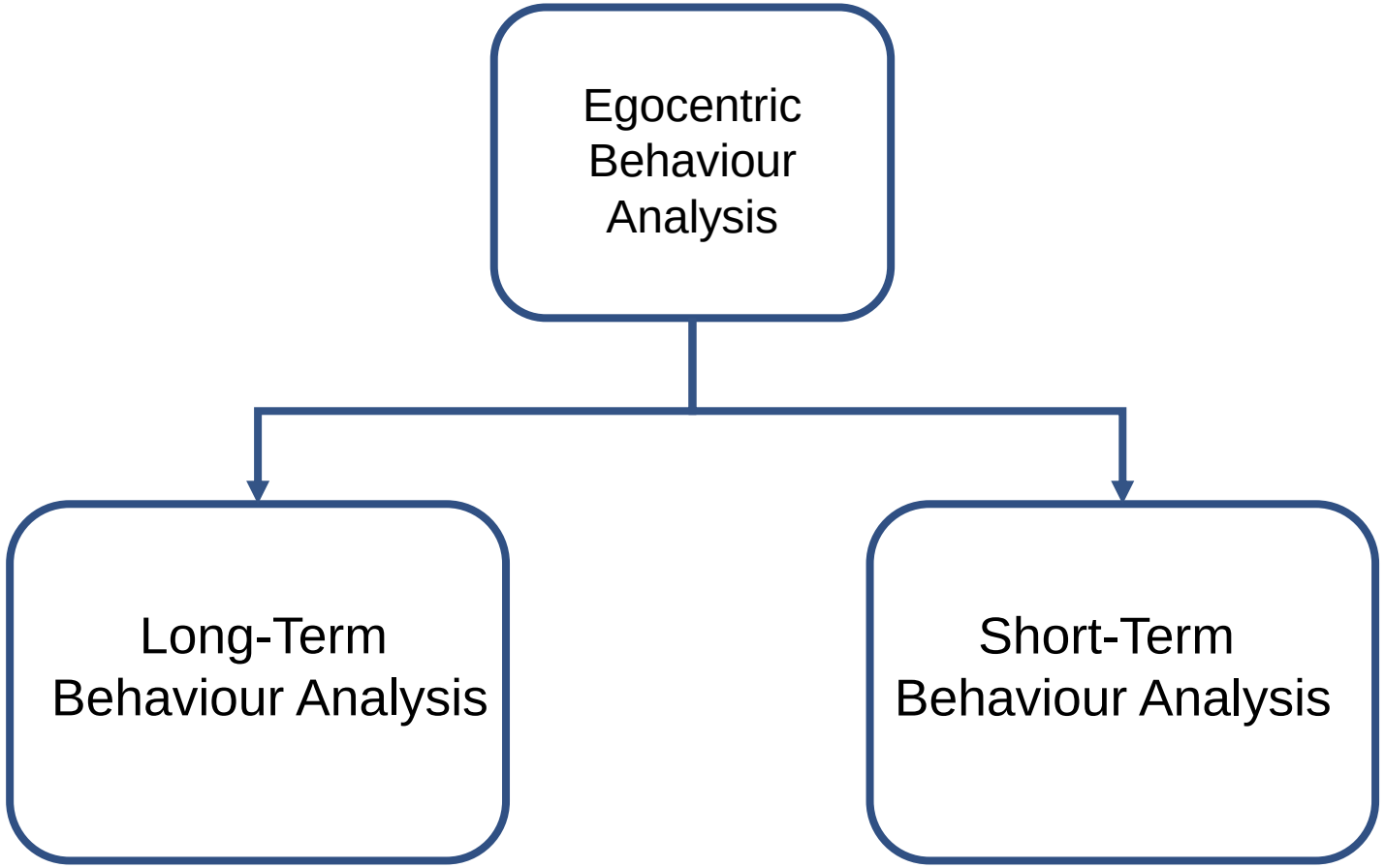
- Using a wearable camera can **illustrate** in detail which **activities** the person wearing the camera has done during the **day**
- **Egocentric** → placing a camera on a **human body** giving a **view** from this **person's perspective**
- **2D Hand Pose Estimation** → Reconstruction of **21 points** describing the finger joints, palm and wrist **positions** in **2D** space

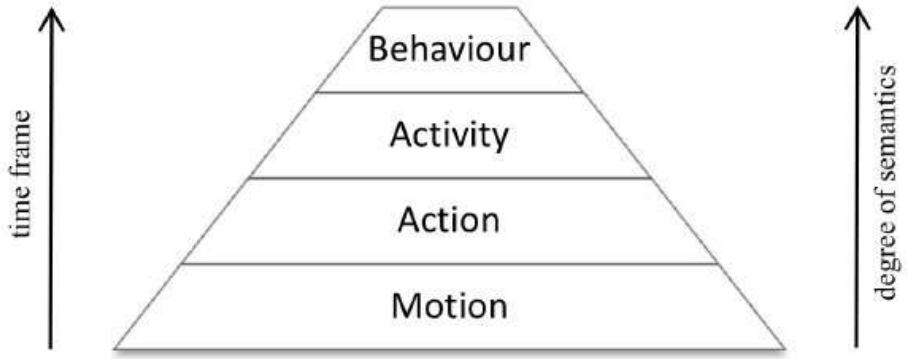
RayBan Stories[3]

User wearing a lifelogging device[1]

[1] https://newatlas.com/narrative-clip-2/35422/ visited on 20.01.2022

Division of egocentric behaviour analysis





The difference between action and activity visualized [7]

[7] Nguyenet al., Recognition of Activities of Daily Living with Egocentric Vision: A Review. *Sensors (Basel, Switzerland)* vol. 16,1 72. 7 Jan. 2016

Division of egocentric behaviour analysis

Long-Term Behaviour Analysis ***vs.*** Short-Term Behaviour Analysis

Routines *vs.* real-time
Registration of **activities *vs.*** detailed understanding of **actions**
Identifying **where** people **eat *vs.*** identifying food and **estimating micro/macro ingredients**

Egocentric Actions

Examples of actions in EPIC-KITCHEN dataset[5]

[5] https://epic-kitchens.github.io/2021 visited on: 24.09.2022

Observations from Egocentric Recordings

- **Hands play the main** role in the scene next to the manipulated object
- Most of the actions are **different hand movements** related to the **object**, e.g. rotating, picking up
- Several **ADLs** based on **hands:**
  - Eating
  - Drinking
  - Taking medication

Estimation of Hand Pose

In the Egocentric View

Egocentric 2D Hand Pose

Single Hand Approach → *EffHandNet*:

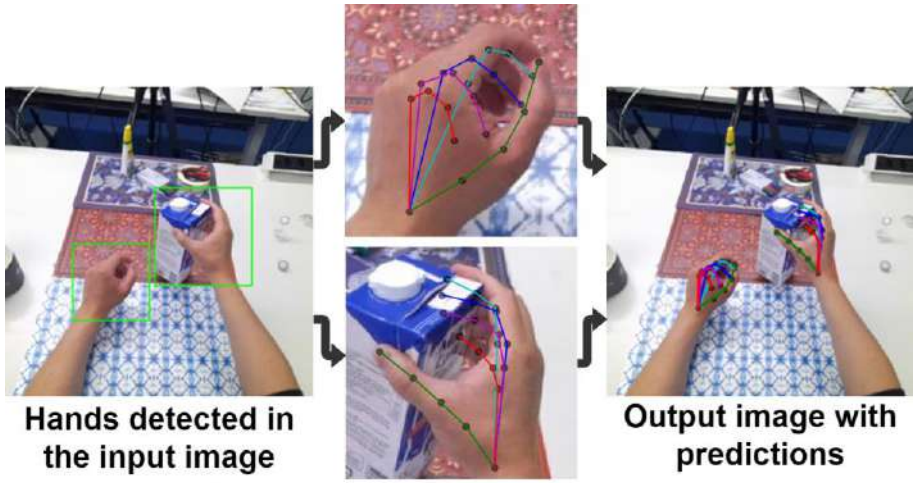
- Pre-trained **hand detector** in the egocentric input image
- Hand pose prediction in segmented regions $\boldsymbol{R_1}, \boldsymbol{R_2}$
- Feature extractor:
  → *EfficientNetV2-S* [11]
- Prediction head:
  → Sequence of transposed convolution resulting in heatmaps

Pros and Cons:

+More datasets available

-Multiplied forward pass through backbone network

-Considers region of a single hand only



[11] Mingxing Tan and Quoc Le, “Efficientnetv2: Smaller models and faster training,” in International Conference on Machine Learning. PMLR, 2021, pp. 10096–10106.

Egocentric 2D Hand Pose

Egocentric Approach → *EffHandEgoNet*

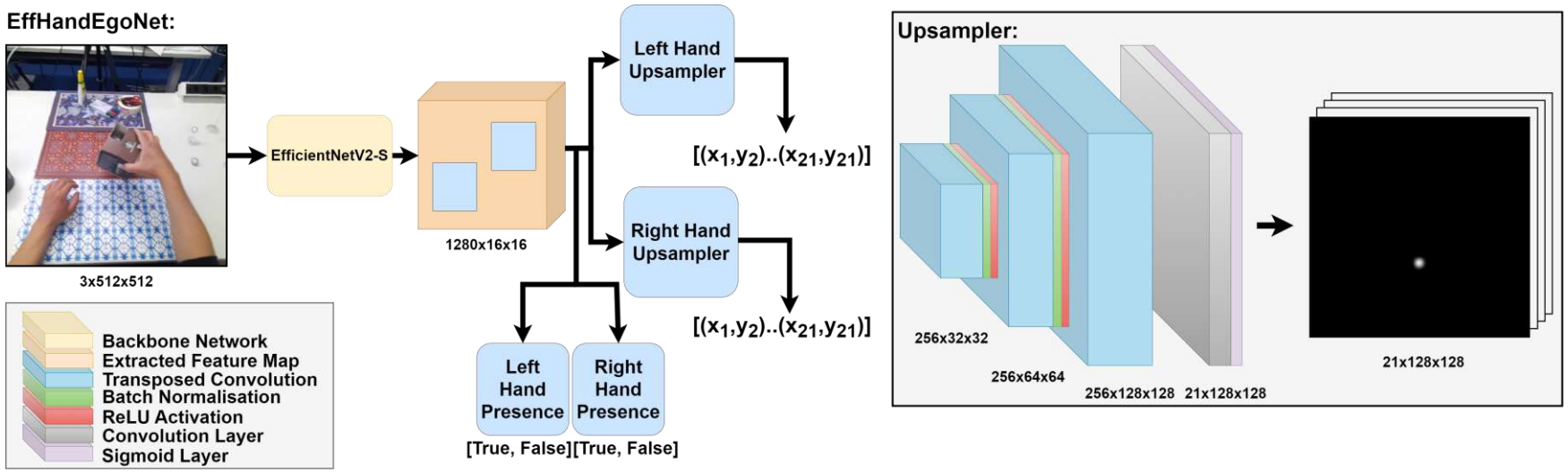
- Handness prediction module H_L, H_R
- Two up-sampling heads
- Improves modelling of hand-object interaction
- Output hand pose: $Ph_i^i = (x, y)$

Pros and Cons:

-Less public data

+Single forward pass in the backbone stage

+Learning hand-hand interactions



Evaluation

EffHandNet:

- **Bets** result in **non-egocentric FreiHand** dataset

TABLE I: Results of 2D single-hand models on *FreiHAND dataset*. Referenced results are reported by the authors of the methods, while unreferenced results are computed by us using open-source implementations.

Method	Year	PCK0.2↑	EPE↓	AUC↑
test subset from random data split 80/10/10				
PoseResNet50 [7]	2020	99.20%	3.27	86.8
MediaPipe	2020	71.77%	7.45	79.7
Santavas et al. [29]	2020	-	4.00	87.0
EffHandNet	2024	98.70%	2.24	92.1
EffHandNet+P	2024	**99.32%**	**1.59**	**93.5**
final test subset				
MediPipe	2020	81.73%	5.29	83.9
PoseResNet50	2020	87.48%	4.32	86.0
EffHandNet	2024	88.76%	4.19	86.5
EffHandNet+P	2024	**91.08%**	**3.67**	**87.9**

Evaluation – Egocentric Hand Pose

EffHandNet:

- **Poor** performance in **hand detection** stage
- **High** End-Point Error (**EPE**) (pixels)

EffHandEgoNet:

- Best performance on both datasets

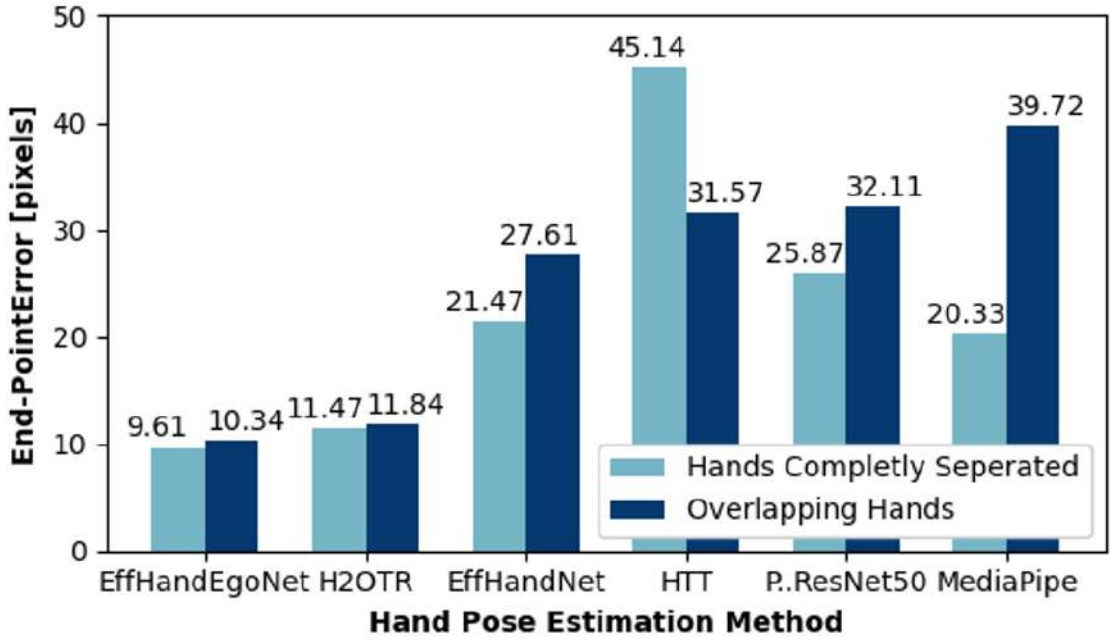
TABLE II: Results for 2D hand pose estimation in egocentric *H2O Dataset*. The table includes hand detection accuracy, hand pose estimation PCK0.2, EPE and AUC metrics in pixels for an image size of 1280x720. Results are calculated using open-source implementations and authors' model weights.

Method:	Year	Acc.↑	PCK0.2↑	EPE↓	AUC↑
H2O Dataset					
PoseResNet50 [7]	2020	99.47	74.42%	26.69	81.4
MediaPipe [40]	2020	96.93	86.22%	21.22	85.1
HTT [37]	2023	-	84.75	19.94	84.8
H2OTR [6]	2023	-	95.55	12.46	89.4
EffHandNet	2024	99.47	76.27%	22.52	82.0
EffHandEgoNet	2024	**99.91**	**97.38%**	**9.80**	**90.7**
FPHA Dataset					
H2OTR [6]	2023	-	94.67	17.50	89.3
HTT [37]	2023	-	92.07	18.07	88.7
Ours	2024	-	**96.37**	**15.20**	**88.5**

Pose Error in Edge Scenarios

Pose error for different methods in edge scenarios for **overlapping** and **fully separated hands:**

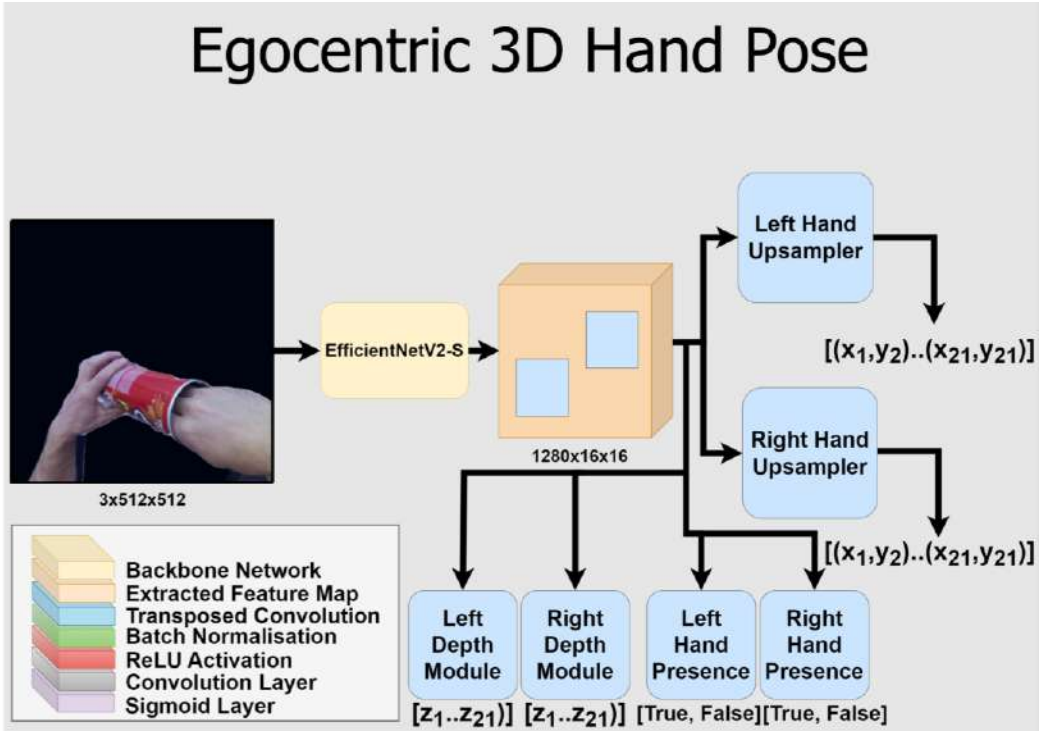
- Bottom-up **the smallest error**
- Bottom-up performs with **minimal difference** between **scenarios**
- **Detection-based** methods result in **higher difference** between **scenarios**



Egocentric 3D Hand Pose

I. Extending *EffHandEgoNet* to 3D

- Architecture with regressions module for estimation of **z** coordinate representing depth
- Regression head + upsampler = 2.5D coordinates (image space)
- Pinhole camera model transformation to 3D



Egocentric 3D Hand Pose with SHARP

SHARP: Segmentation of Hands and Arms by Range using Pseudo-Depth

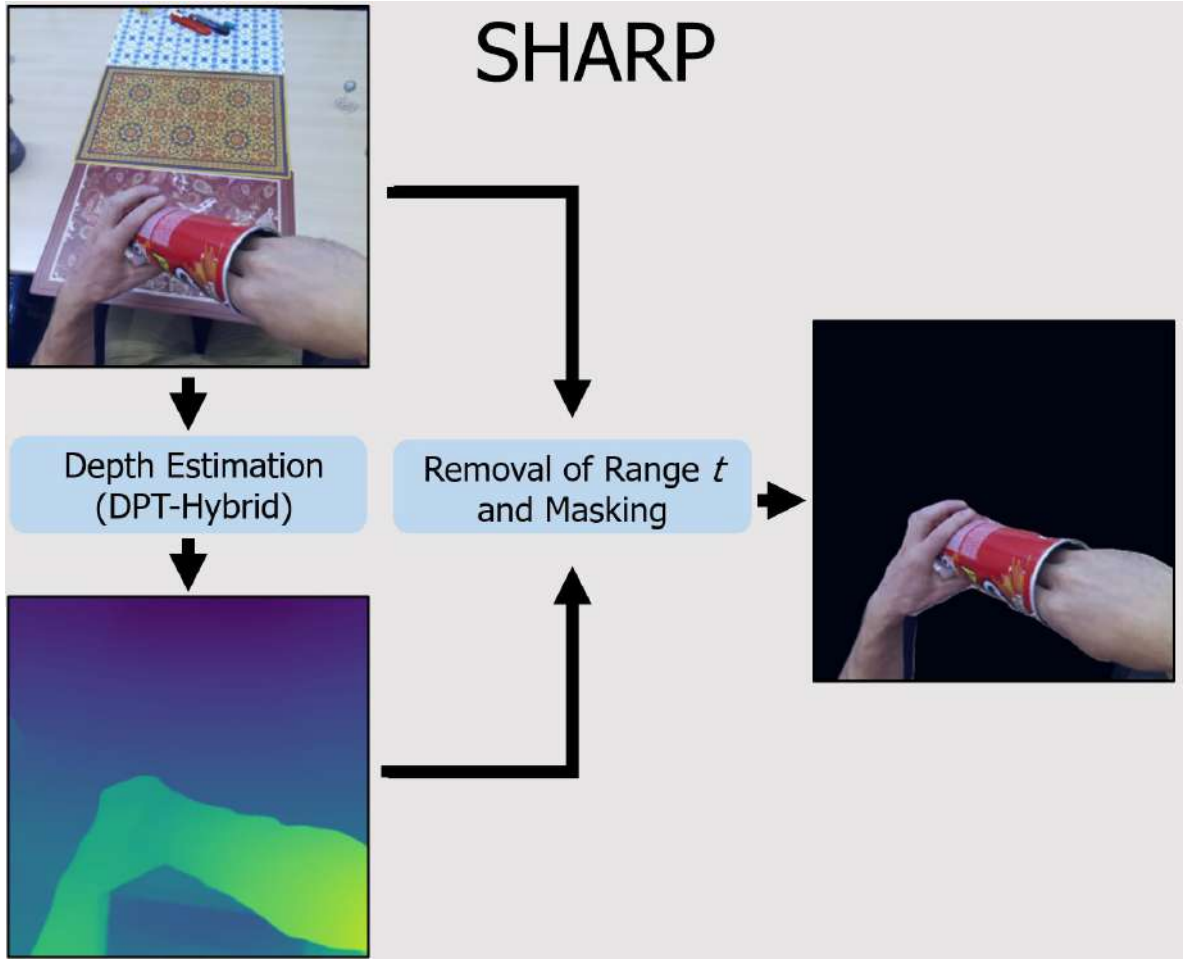
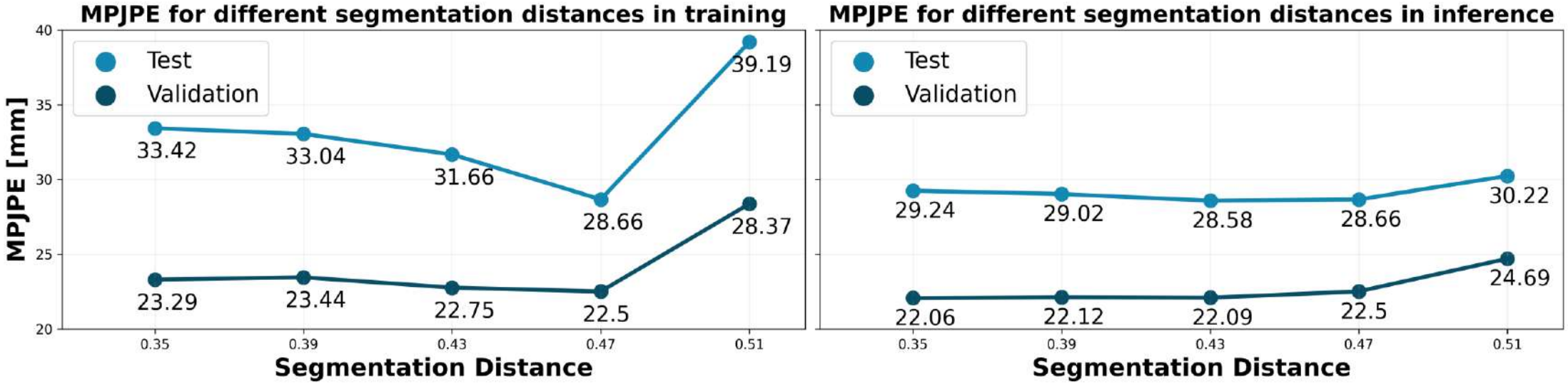


Table 3. Results of ablations studies with different depth image types used in *SHARP*. All results provided in *mm* in camera space for left, right and both hands.

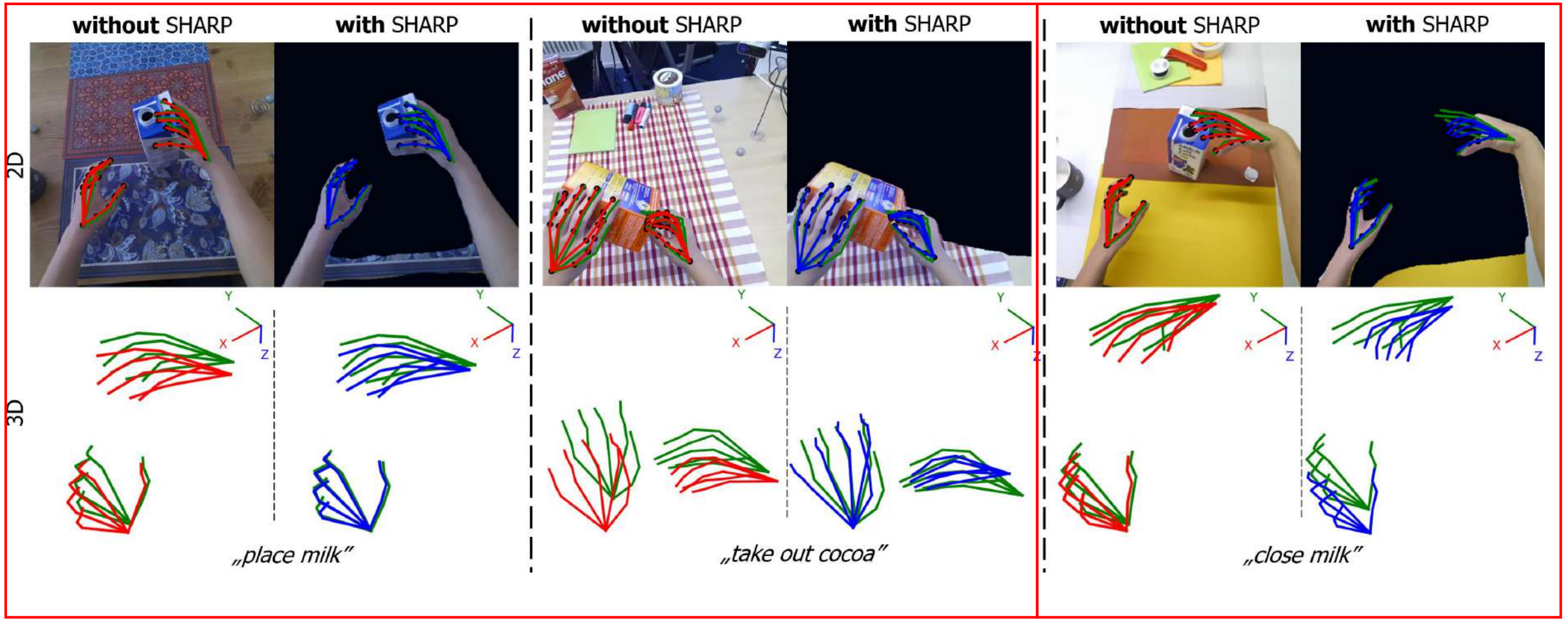
	Depth	MPJPE Left ↓	MPJPE Right ↓	MPJPE Both ↓
Ours	Estimated	30.31	27.02	28.66
Ablation I	✗	32.95	38.01	35.48
Ablation II	Ground Truth	21.31	28.86	25.09
Ablation III	Est.+De-sharpen	39.49	35.01	37.25

SHARP and Segmentation Distance

Fig. 7. On the left, frame processed with *SHARP* and different values of t. On the right, the same frame processed with *SHARP*, $t = 0.47$ and with de-sharpening applied.



Qualitative Results of 3D Hand Pose Estimation with *SHARP*

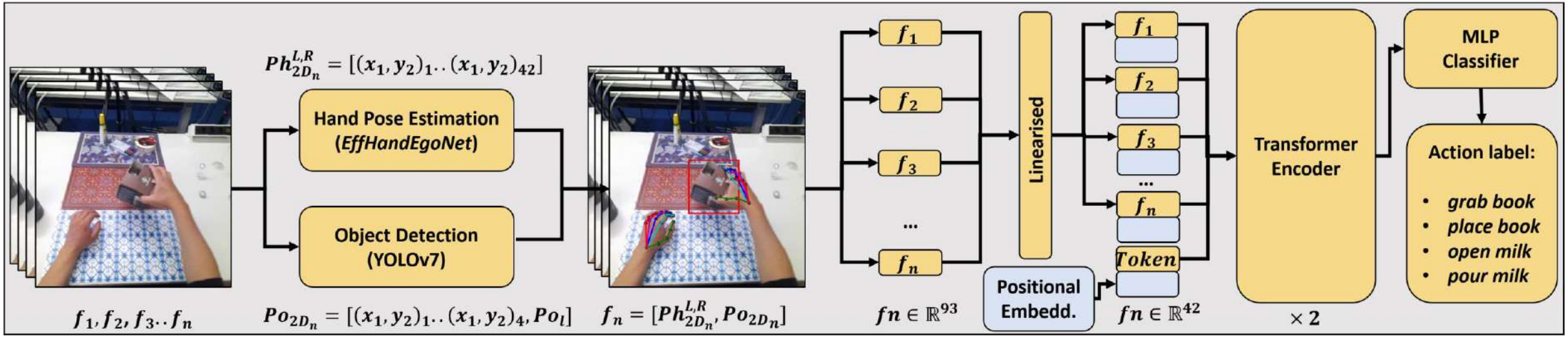


Applications of Hand Pose Estimation

In the Egocentric View

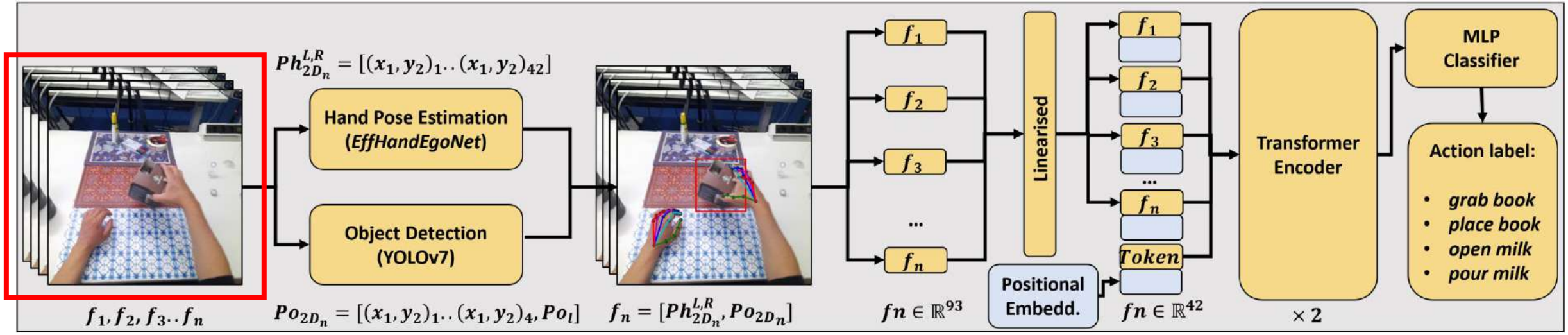
Action Recognition

Hand-based Action Recognition - Overview



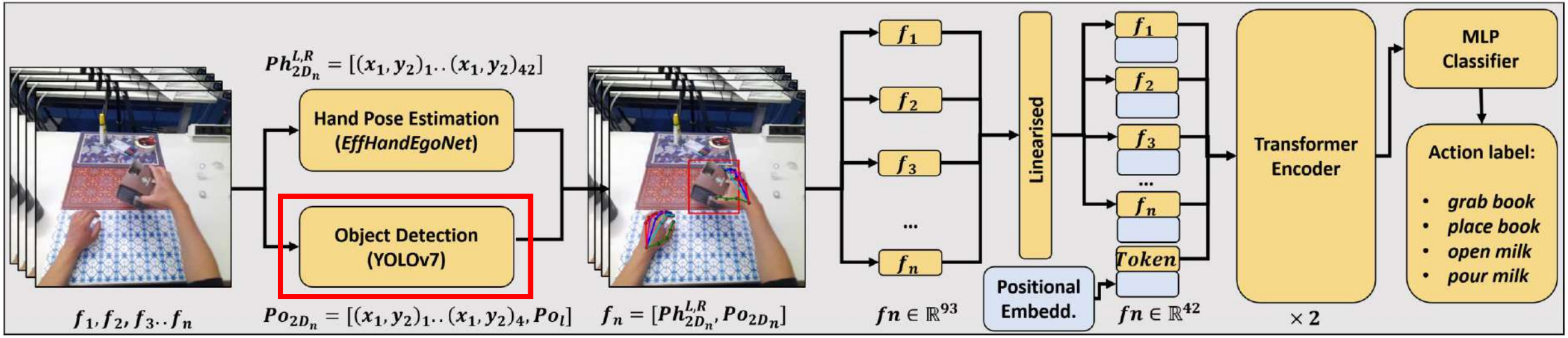
- Usage of **hands** and **objects** as **input** for supervised **sequence** model
- Allows to use of **pre-trained** models reducing the learning costs

Hand-based Action Recognition - Overview



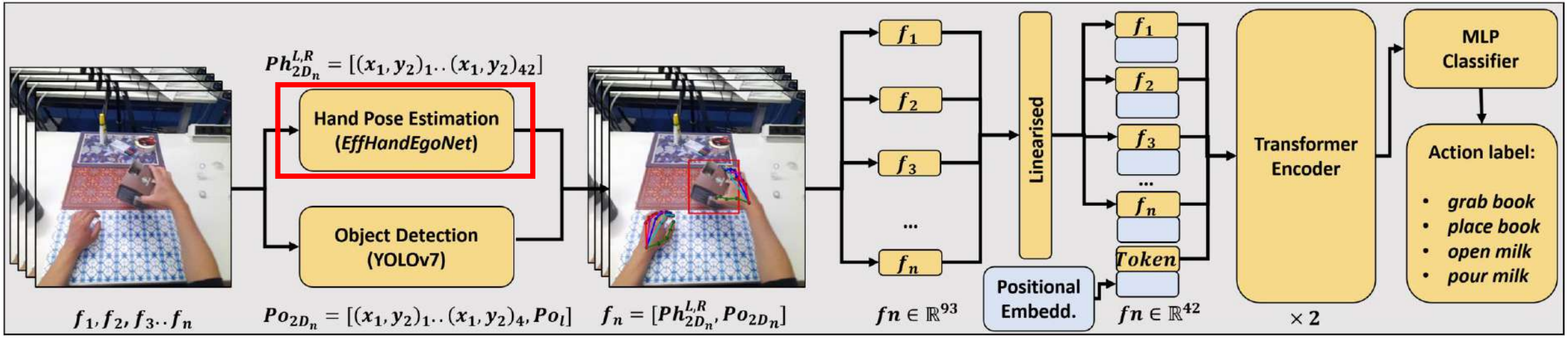
- Input sequence of frames $f_1, f_2 .. f_n$ where $n \in [1, 2..20]$
- Actions **shorter** than n frames → **zero padding**
- Actions **longer** than n frames → **uniform subsampling**

Hand-based Action Recognition - Overview



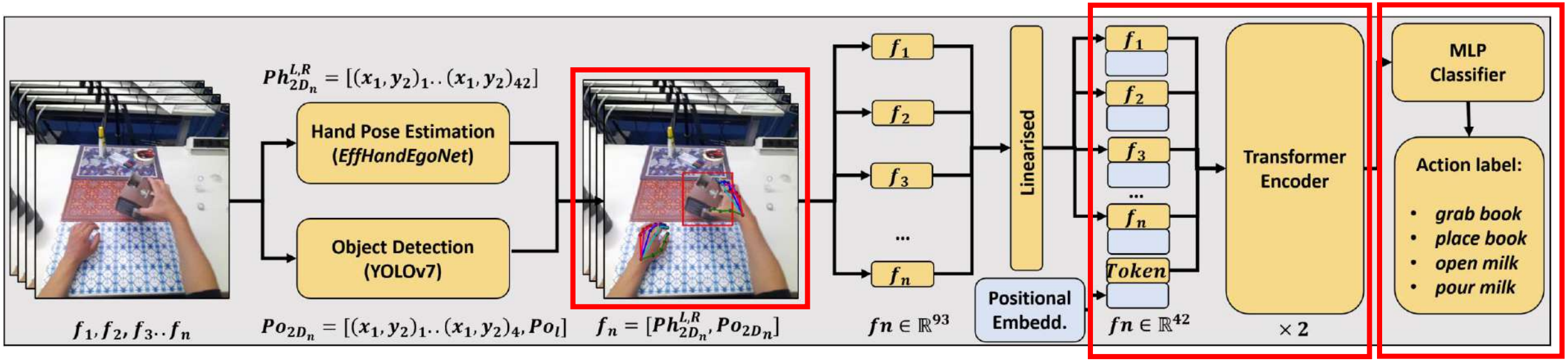
- Implemented using state-of-the-art **YOLOv7** model
- Object described as $\boldsymbol{Po_{bb}^{i}(x, y)}$ where $\boldsymbol{i} \in [\mathbf{1..4}]$ corresponds to the **bounding box corners**
- $\boldsymbol{P_{o_l}}$ describes object label

Hand-based Action Recognition - Overview



- We implement two different approaches for hand pose estimation:
  → **top-down** and **bottom-up**

Hand-based Action Recognition - Overview

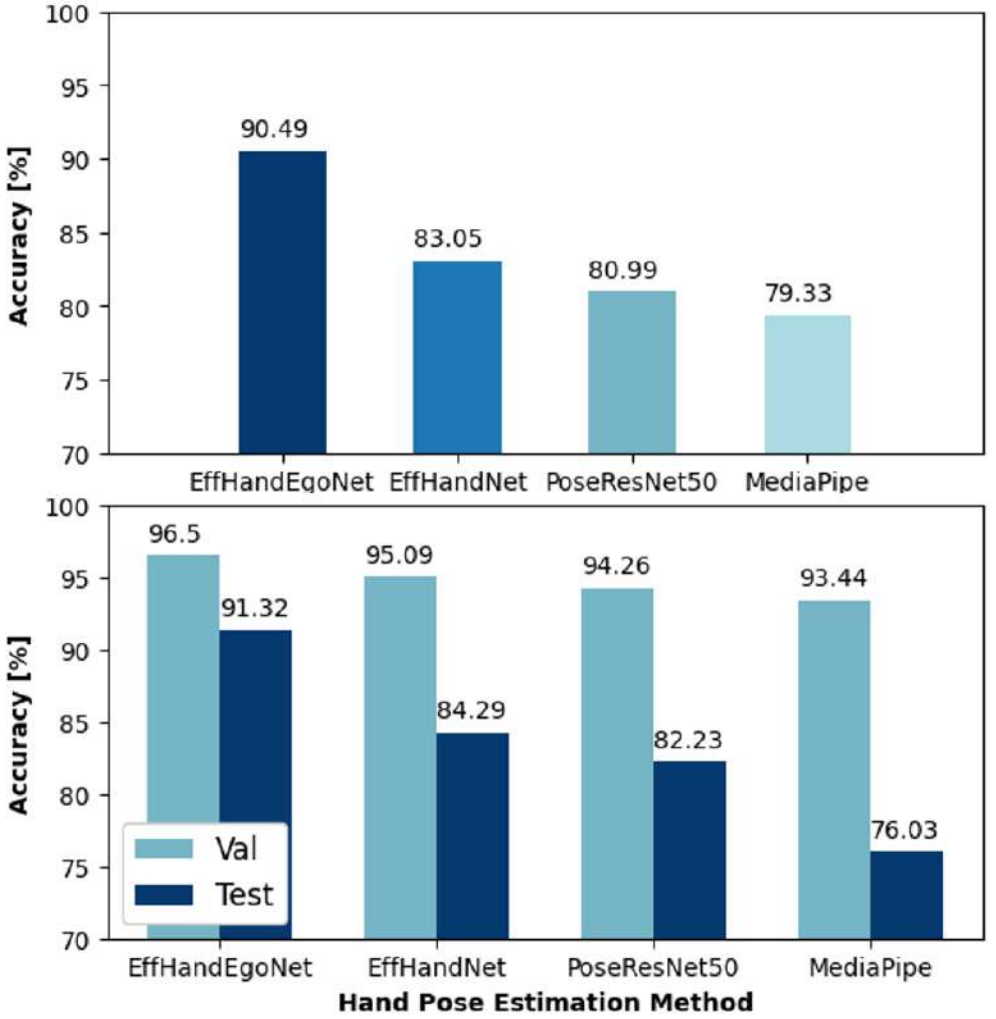


- Each frames describes: $\boldsymbol{f_n = Ph_l^i(x,y)\ Ph_r^i(x,y) Po_{bb}^i(x,y) P_{o_l}}$
- Sequence of frames: $\boldsymbol{V_{seq} = [f_1, f_2 .. f_n]}$

Action Recognition and Hand Pose Models

Action recognition with different hand pose estimation methods:

- **Strong correlation** in **action recognition** accuracy of predicted **hand poses** when we **train** with **ground truth** skeletons and **test with prediction**
- **Similar** observation for **training and testing** with **estimates**



Evaluation - Action Recognition

State-of-the-art results **despite** using **only 2D** information

- **Best results** in *H2O Dataset*
- **Competitive** result in *FPHA Dataset*

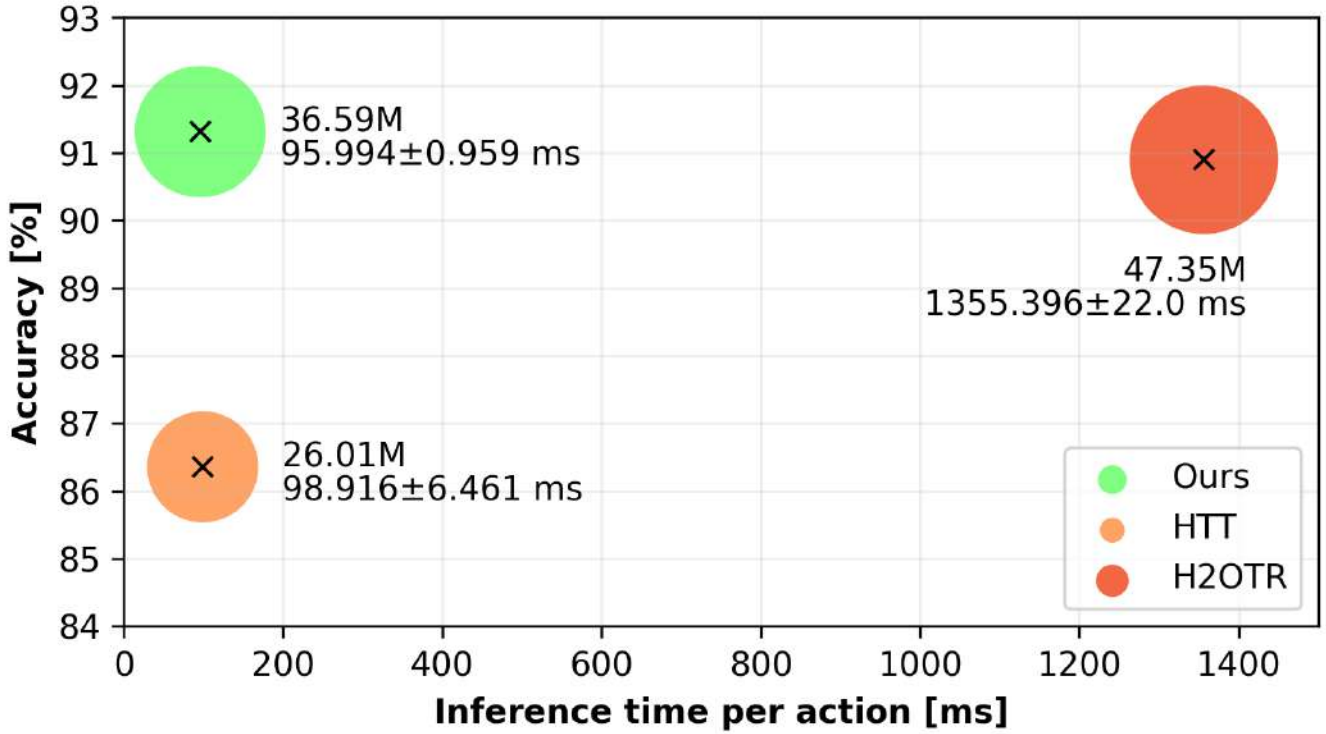
TABLE III: Results in accuracy of action recognition methods on *H2O* and *FPHA* datasets. Inputs of methods are: *Img* stands for semantic features extracted from an image using CNN network, *Hand P.* and *Obj P.* stand for pose information type for hands and objects, and *Obj L.* stands for object label. Results are from referenced papers.

H2O Dataset						
Method:	Year	Img	H. P.	Obj P.	Obj L.	Acc.↑
C2D [35]	2018	✓	✗	✗	✗	70.66
I3D [4]	2017	✓	✗	✗	✗	75.21
SlowFast [12]	2019	✓	✗	✗	✗	77.69
H+O [32]	2019	✗	3D	6D	✓	68.88
ST-GCN [38]	2018	✗	3D	6D	✓	73.86
TA-GCN [17]	2021	✗	3D	6D	✓	79.25
HTT [37]	2023	✓	3D	✗	✓	86.36
H2OTR [6]	2023	✗	3D	6D	✓	90.90
Ours	2024	✗	**2D**	**2D**	✓	**91.32**
FPHA Dataset						
Method:	Year	Img	H. P.	Obj P.	Obj L.	Acc.↑
FPHA [13]	2018	✗	3D	-	✓	78.73
H+O [32]	2019	✗	3D	-	✓	82.43
Coll. [39]	2020	✓	3D	-	✓	85.22
HTT [37]	2023	✓	3D	-	✓	94.09
VPA [28]	2021	✗	3D	-	✓	**95.93**
Ours	2024	✗	**2D**	-	✓	94.43

Inference time with 2D Hand Pose

Inference test performed over **1000 runs** on NVIDIA RTX 3060:

- Circles in the figure represent the **number** of trainable **parameters**
- Our method has the **fastest inference** and highest **action recognition**

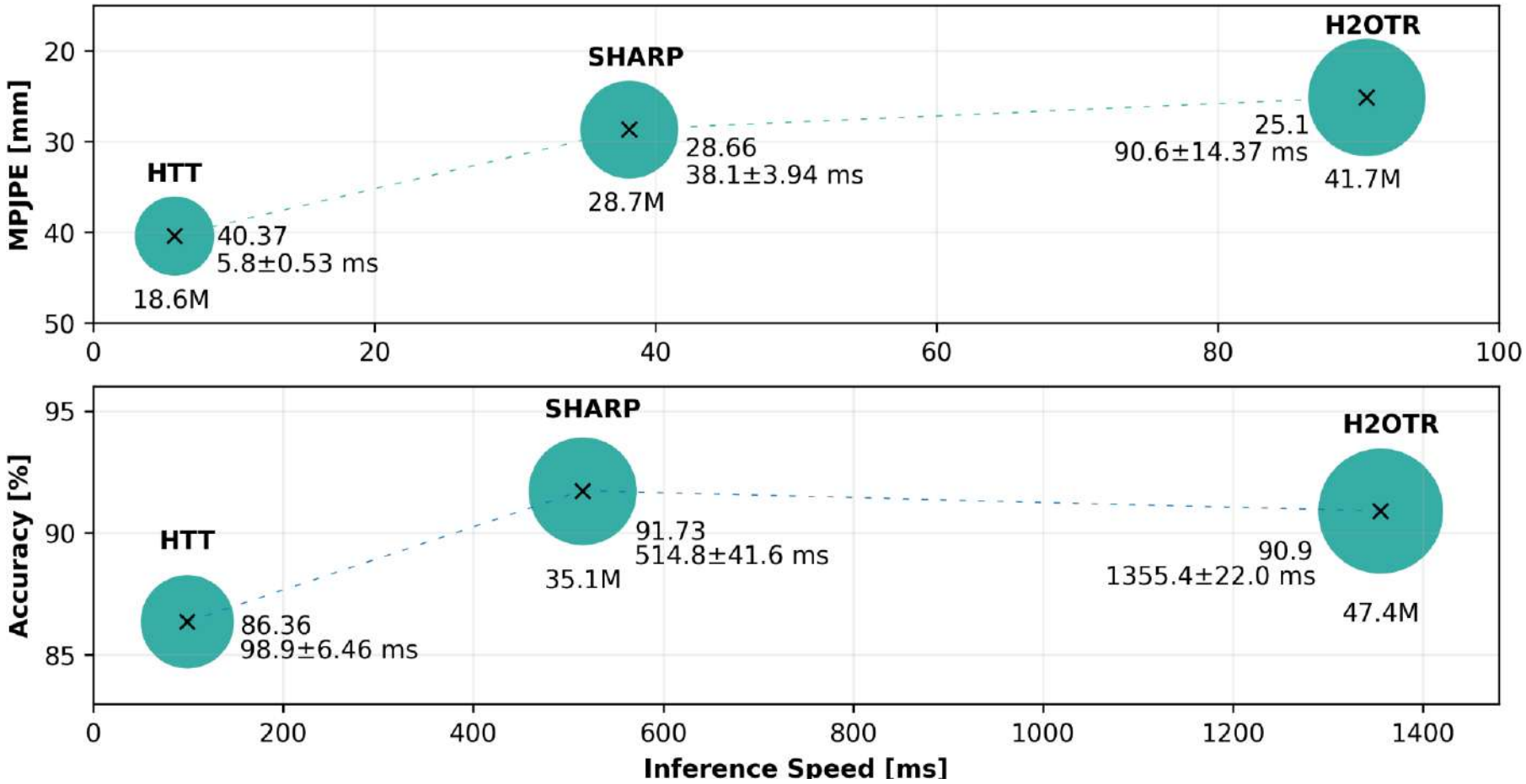


Action Recognition Results with SHARP

Table 2. Results in accuracy of action recognition methods on *H2O Dataset*. Inputs of methods are: *Img* stands for semantic features extracted from an image using CNN network, *Hand Pose* and *Obj Ppose* stand for pose information type for hands and objects, and *Obj Label* stands for object label. Results origin from referenced studies.

Method:	Year	Img	Hand Pose	Obj Pose	Obj Label	Acc. ↑
C2D [27]	2018	✓	✗	✗	✗	70.66
I3D [2]	2017	✓	✗	✗	✗	75.21
SlowFast [10]	2019	✓	✗	✗	✗	77.69
H+O [25]	2019	✗	3D	6D	✓	68.88
ST-GCN [31]	2018	✗	3D	6D	✓	73.86
TA-GCN [16]	2021	✗	3D	6D	✓	79.25
HTT [29]	2023	✓	3D	✗	✓	86.36
H2OTR [5]	2023	✗	3D	6D	✓	90.90
EffHandEgoNet [17]	2024	✗	2D	2D	✓	91.32
Ours	Now	✗	3D	2D	✓	**91.73**

Inference Time with SHARP



Struggle Determination

Struggle Determination

Determination of struggle level in **three** different task → **binary** and **4-way**

Motivation:

→ Correct struggle recognition leads to robust assistance for individuals

Current results and outcomes:

- **Binary** determination with **89%** of accuracy
- Best when **merging hand pose** information and **semantic** features from image

Tower of Hanoi

Tent Assembly

Pipes Assembly

Hand Rehabilitation

Examples of common hand rehabilitation exercises for stroke patients

Wrist Curls

Ball Grip

Palm Up and Down

Motivation:

- Stroke remains the **third** leading cause of **mortality** and **disability** worldwide
- Approximately **85%** of stroke patients worldwide **experience hand dysfunction**
- No egocentric studies available

Challenges:

Exercise recognition
Repetition counting
Exercise detection
Form evaluation

Preliminary stage

- 9 users of different age (25-88 years old)
- 13 common exercises for each hand → 25 exercises
- 97% recognition of exercises in validation subset with SlowFast Network

Conclusion

- **Bottom-up** methods **tend** to **fail** in **egocentric perspective** for **hand pose** estimation
- In certain scenarios, i.e. **reducing inference time**, **2D** pose information is a **promising alternative** to estimated **3D** pose for **egocentric action recognition**
- Using **pseudo-depth** information to **remove irrelevant scene** parts from the egocentric view **improves hand pose estimation**

- **Accurate pose** description is **essential** for **correct action understanding**, struggle determination, and **more** in **egocentric** image processing

Thank you!

Wiktor Mucha

TU Wien

wiktor.mucha@tuwien.ac.at