3D-HOG EMBEDDING FRAMEWORKS FOR SINGLE AND MULTI-VIEWPOINTS ACTION RECOGNITION BASED ON HUMAN SILHOUETTES

Federico Angelini*, Zeyu Fu*, Sergio A. Velastin‡, Jonathon A. Chambers*, Syed Mohsen Naqvi*

*Intelligent Sensing and Communications Research Group, Newcastle University, UK

‡Department of Computer Science and Engineering, University Carlos III de Madrid, Spain

Intelligent Sensing · Newcastle University · THALES · EPSRC Engineering and Physical Sciences Research Council

1 Overview

Given the high demand for automated systems for human action recognition, great efforts have been undertaken in recent decades to progress the field [1]. In this paper, we present frameworks for single and multi-viewpoints action recognition based on:

- Space-Time Volume (STV) of human silhouettes
- 3D-Histogram of Oriented Gradient (3D-HOG) Embedding [2]

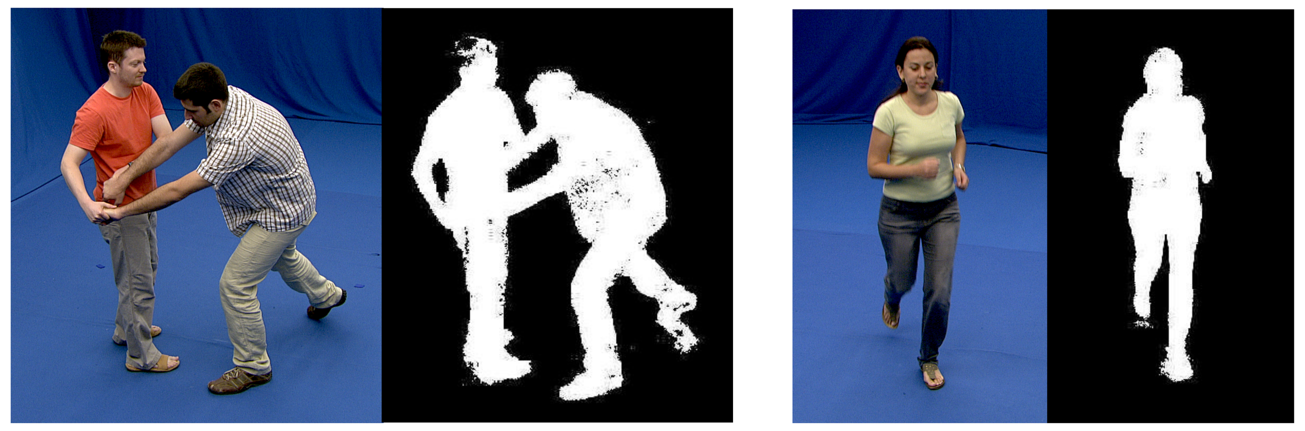


Fig 1. Examples of RGB and Silhouettes data.

Our contributions

- 3D-HOG Embedding [2] based frameworks exploiting local gestures analysis
- single and multi-viewpoints cases
- accuracy and robustness to appearance changes
- **outperforming** results on Weizmann and i3DPost datasets

2 Baseline Method

- Baseline method: 3D-HOG Embedding [2]
- It defines the basic data processing structure (Fig 2), also used in the Proposed Frameworks

Key drawbacks

- **Attention problem** (Fig 3): it has not been addressed;
- **Performance stability**: affected by *randomly selected library* in the Embedding phase;
- **Action-labels-based local classifiers**, without considering cross-location local gestures relationships;

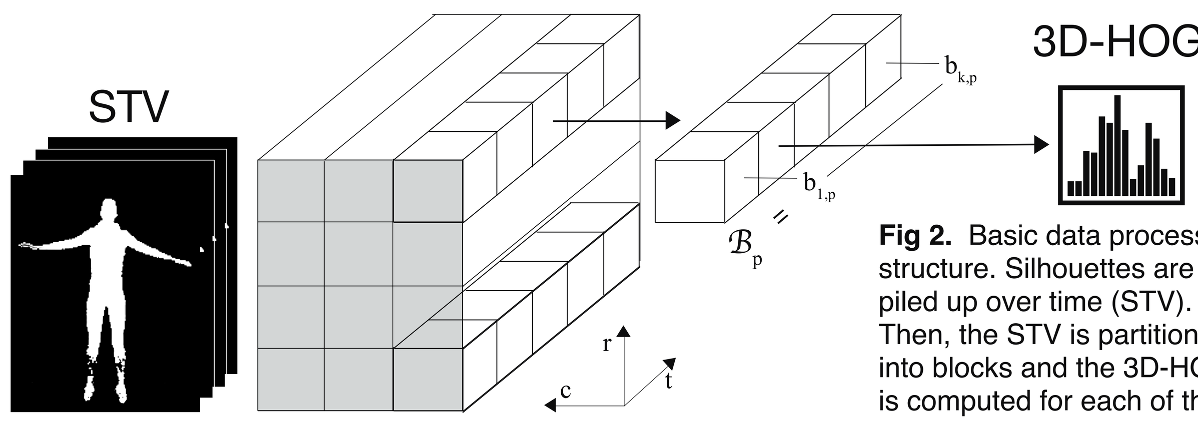
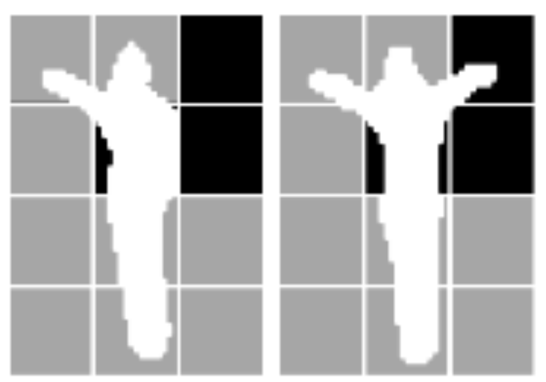


Fig 2. Basic data processing structure. Silhouettes are piled up over time (STV). Then, the STV is partitioned into blocks and the 3D-HOG is computed for each of them.

Fig 3. Example of Attention Problem.



- ■ locations where Action 1 and Action 2 look different
- ■ locations where Action 1 and Action 2 look similar

3 Proposed Frameworks

Main ideas

- Locally, actions look like simpler gestures
- Globally, an action can be seen as a particular combination of local gestures (Fig 4)

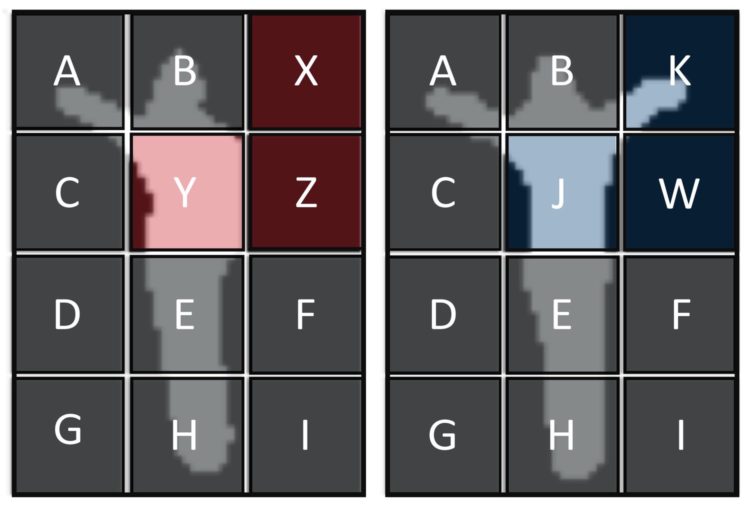
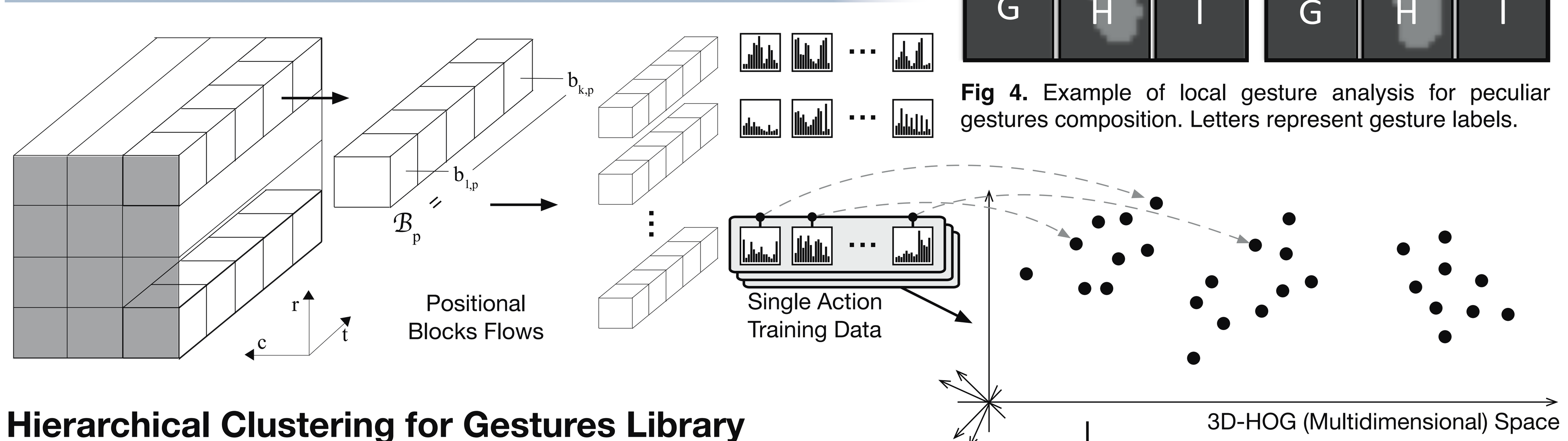


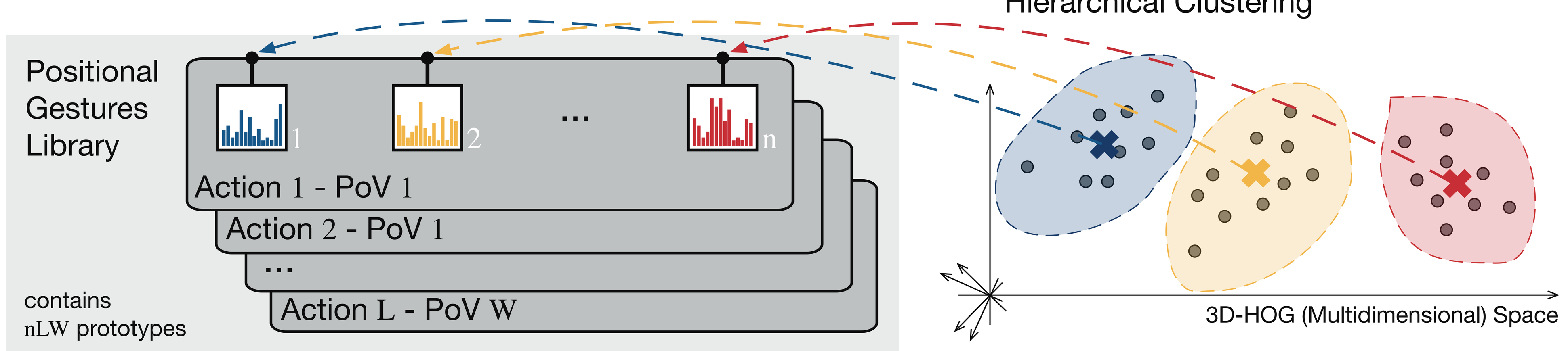
Fig 4. Example of local gesture analysis for peculiar gestures composition. Letters represent gesture labels.

Multi-viewpoints framework

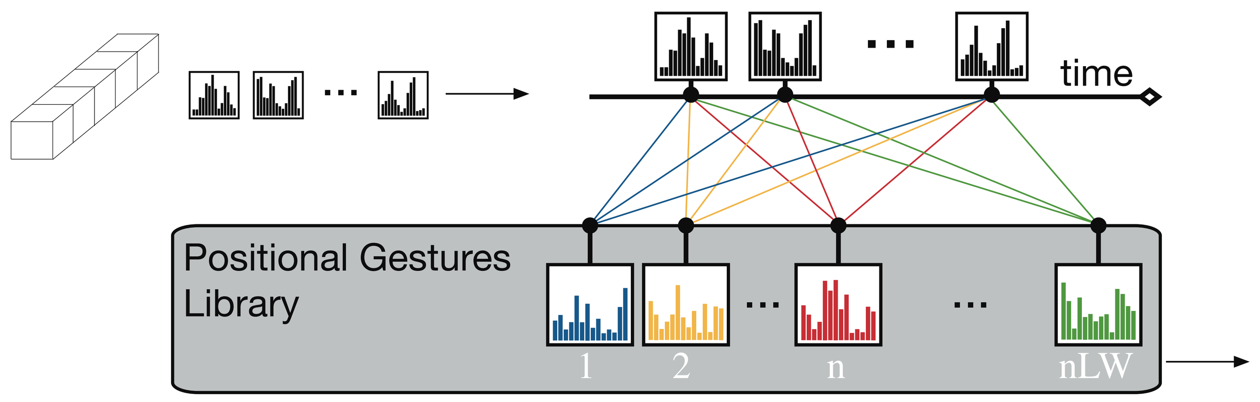


Hierarchical Clustering for Gestures Library

$$\{\mathbf{b}_{j,p}\}_{j=1}^{H} = \{\mathbf{b}_{j,p}\}_{j=1}^{S_1} \cup \{\mathbf{b}_{j,p}\}_{j=S_1}^{S_2} \cup \cdots \cup \{\mathbf{b}_{j,p}\}_{j=S_{n-1}}^{S_n}$$



3D-HOG Embedding of block flows



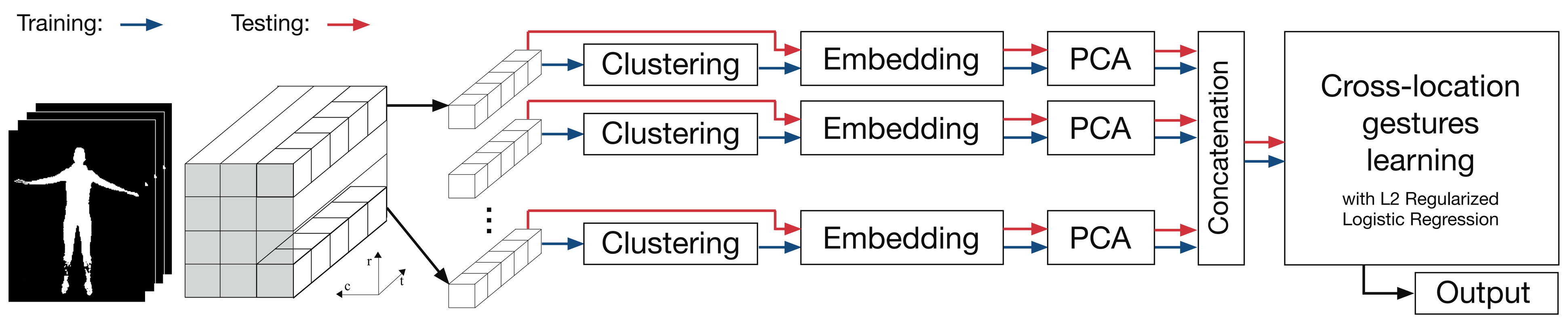
Embedding Rule

$$D_i(\mathcal{B}_p) = \min_{j=1,\ldots,K} d(\mathbf{b}_{j,p}, \mathbf{v}_i)$$

$$i = 1, \ldots, nLW$$

$$\mathcal{D}_p = [D_1(\mathcal{B}_p), \ldots, D_{nLW}(\mathcal{B}_p)]$$

General Pipeline



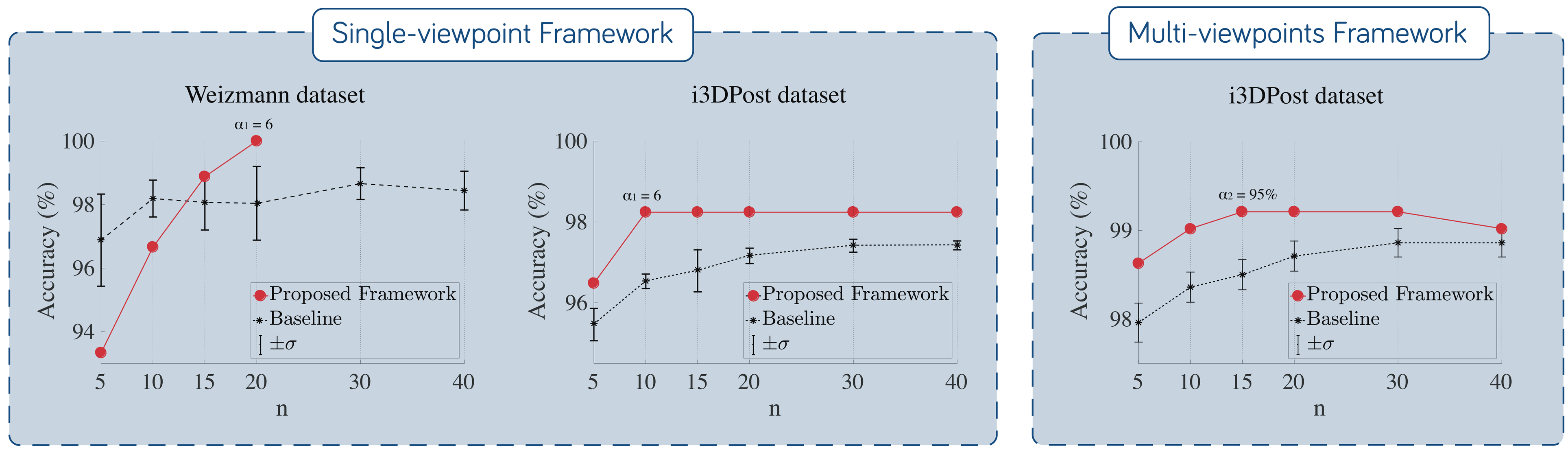
Key facts

- **Gesture analysis**: to leverage local information for global action information retrieval
- **Searching local gestures**: 3D-HOG features clustered with *Hierarchical Clustering*
- No longer randomly chosen **Local (Gestures) Library**
- **Cross-location Gesture Composition Learning**: with *L2 Regularized Logistic Regression*

4 Results

Comparison between baseline and proposed frameworks performance

Experimental Setting: *Leave-one-actor-out* (robustness to appearance changes)



Legend

- n: clustering parameter ∝ library size
- σ: standard deviation
- α_1: number of principal components (PCA)
- α_2: explained variance percentage (PCA)

Conclusions

- **Outperforming results** in all studied cases
- **Stable performance** over different trainings
- Higher accuracy for smaller n (best values)

Comparison between dataset state-of-arts and proposed frameworks performance

Experimental Setting: *Leave-one-actor-out* (robustness to appearance changes)

Weizmann Dataset

Method	Actions	Accuracy	n	α_1
Proposed Framework	10	**100%**	20	6
Gorelick et al.	10	**100%**	-	-
Jiang et al.	10	**100%**	-	-
C. Li et al.	9	97.53%	-	-
Ahsan et al.	9	97.5%	-	-
Ahsan et al.	10	94.26%	-	-

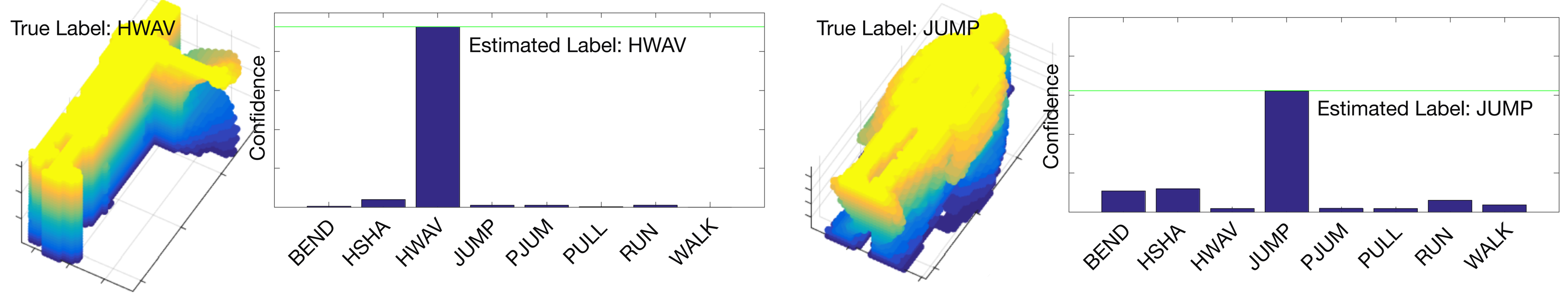
i3DPost Dataset

Method	Actions	POV	Accuracy	n	α_2
Proposed Framework	8	8	**99.60%**	30	95%
Proposed Framework	6	8	**99.73%**	30	99%
Castro et al.	6	2	99.00%	-	-
Iosifidis et al.	6	8	98.16%	-	-
Iosifidis et al.	8	8	96.34%	-	-
Hilsenbeck et al.	6	8	92.42%	-	-

Conclusions

State-of-art results (Weizmann) and **outperforming** results (i3DPost)

Examples



5 References

[1] S. Herath, M. Harandi, F. Porikli, "Going deeper into action recognition: A survey", Image and Vision Computing, vol 60, pp 4–21, 2017.

[2] D. Weinland, M. Özuysal, P. Fua, "Making action recognition robust to occlusions and viewpoint changes", LNCS - Lecture Notes in Computer Science, vol 6313, no. part 3, pp. 635-648, 2010.

VIDEO Human Behaviour Analysis Demo

VIDEO Multimodal Signal and Information Processing Team