Extracting Action Hierarchies from Action Labels and their Use in Deep Action Recognition

Konstantinos Bacharidis, Antonis A. Argyros

Institute of Computer Science,
Foundation for Research and Technology – Hellas (FORTH)

AND

Department of Computer Science,
University of Crete – Hellas

OVERVIEW

We utilize linguistic semantic associations between action label sentences, to formulate shallow Action Trees, and explore ways to incorporate this coarse-to-fine hierarchy in existing deep neural network architectures, evaluating the impact on recognition accuracy and learning speed.

MOTIVATION

Label sentence linguistic correlations indicate potential similarities in appearance/motion representations of action video sequences. Action granularity level relates to label sentence size and complexity. Actions with similar coarse action motifs tend to share verbs with high semantic similarity, and differentiate based on finer object-related criteria.

MAIN IDEA

We combine off-the-self part of speech taggers with custom action-oriented syntax rules to detect motion verbs in label sentences. We define two-level Action Trees by clustering labels based on action verb semantic similarities relying on word-embedding associations. We reshape existing deep neural networks for this coarse-to-fine classification task, and utilize the representations learned for the coarser action set as complementary to the ones learned for the fine-grained action set, resulting in accuracy and learning speed improvements.

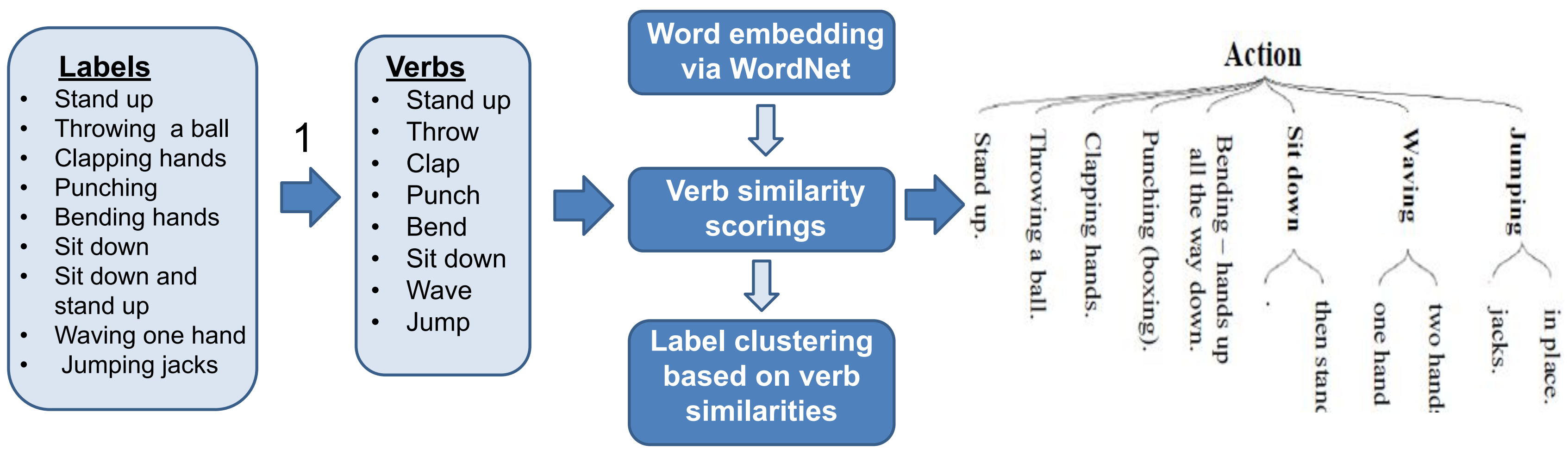
PROPOSED METHOD

1 Assign part-of-speech tags & refine tags with custom action-oriented syntax rules, account for words with multiple semantic interpretations (e.g. screw: noun or verb, take off & take out).

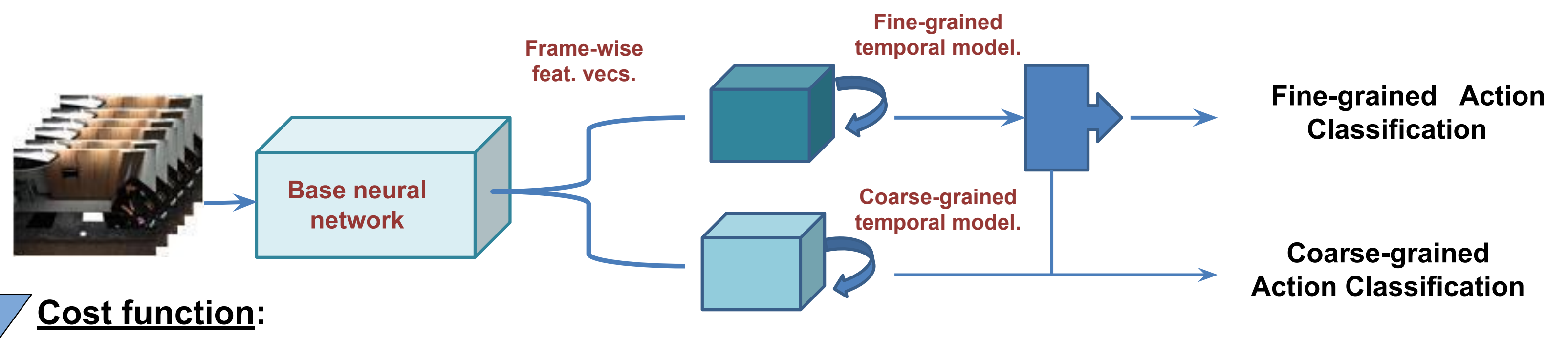
- **Verbs**: discriminate between cases of the same verb when followed by an ad-position or a particle (at, on, out, over per, that, up, with)
- **From noun to verb**: acceptable action description is formed as

verb + adposition/ particle + noun

2 Cluster classes based on verb commonalities or having verbs with high semantic content similarity (based on Word Net-defined embed-dings).



3 Reshape architecture to comply with the dual learning task (a) coarse-, (b) fine-grained class sets. Split into sub-networks for each task, add coarse representation to fine-grained.


Frame-wise feat. vecs.
Base neural network
Fine-grained temporal model.
Coarse-grained temporal model.
Fine-grained Action Classification
Coarse-grained Action Classification

Cost function:

$$C = -\frac{1}{N}\sum_{n=1}^{N}\left[\sum_{k=1}^{K} T^{gn}_{n,k} log\left(Y^{gn}_{n,k}\right) + \sum_{l=1}^{L} w_l T^{fn}_{n,l} log\left(Y^{fn}_{n,l}\right)\right]$$

EXPERIMENTAL RESULTS

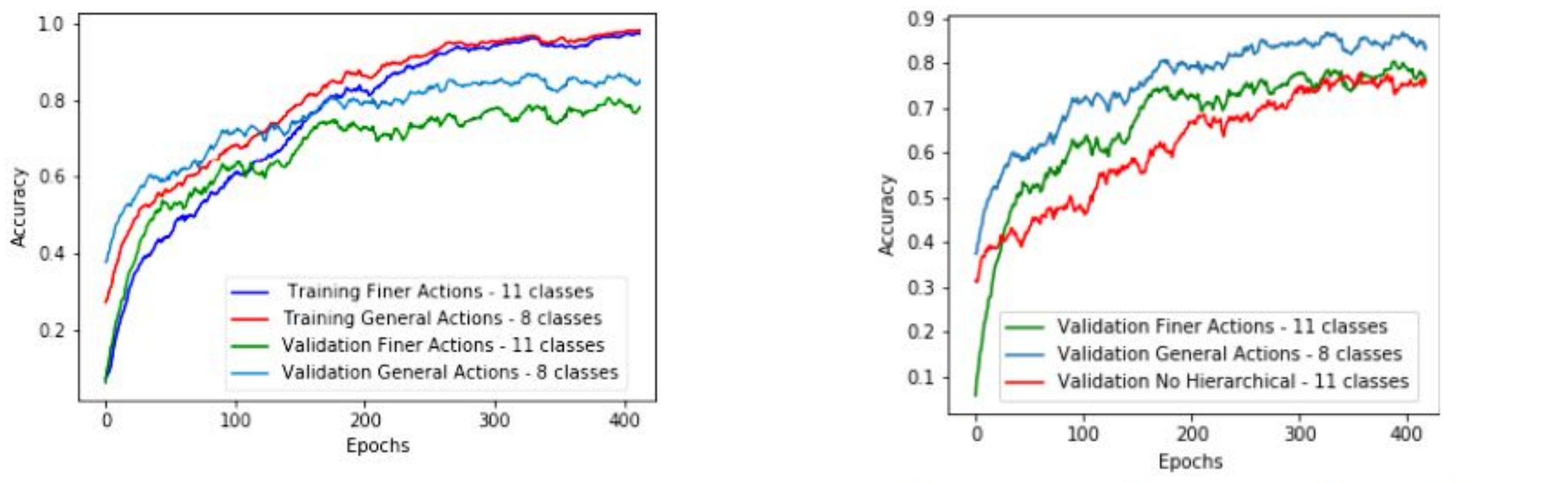
- **Datasets**: (a) *MHAD* (11-classes), coarse-grained (simple and small label sentences), (b) *J-HMDB* (21-classes), mid-range (medium sized and moderate label sentence complexity, (c) MPII Cooking Activities (64-classes), fine-grained (large and complex label sentences)

	Datasets		
	MHAD	*J-HMDB*	*MPII Cooking*
Num unique verbs	9 verbs	19 verbs	42 verbs
Avg num verbs/lbl	1.128 verb/lbl	1.0 verb/lbl	1.188 verbs/lbl
Avg lbl length	3.182 PoS/lbl	1.333 PoS/lbl	2.297 PoS/lbl
Avg asc via verb	0.545 asc/lbl	0.286 asc/lbl	1.656 asc/lbl
Max/min asc verb	1/0 asc	2/0 asc	5/0 asc
Num finer labels	11	21	64
Num Gen labels	8	18	36

- **Impact on accuracy**: hierarchical design improves accuracy up between 4 – 6% compared to a non-hierarchical one

Architecture	Datasets (mAcc. (Coarse, Fine)%)		
Design	*MHAD*	*J-HMDB*	*MPII Cook*
NH-BiLSTM	(-, 64.17)%	(-, 36.28)%	(-, 29.45)%
H-BiLSTM	(82.50, 70.25)%	(45.68, 42.61)%	(60.70, 35.40)%
NH-I3D	(-, 89.61)%	(-, 72.38)%	(-, 48.18)%
H-I3D	(98.75, 96.38)%	(78.47, 76.10)%	(70.47, 54.30)%

- **Impact on learning speed**: hierarchical design increases learning speed in earlier epochs.

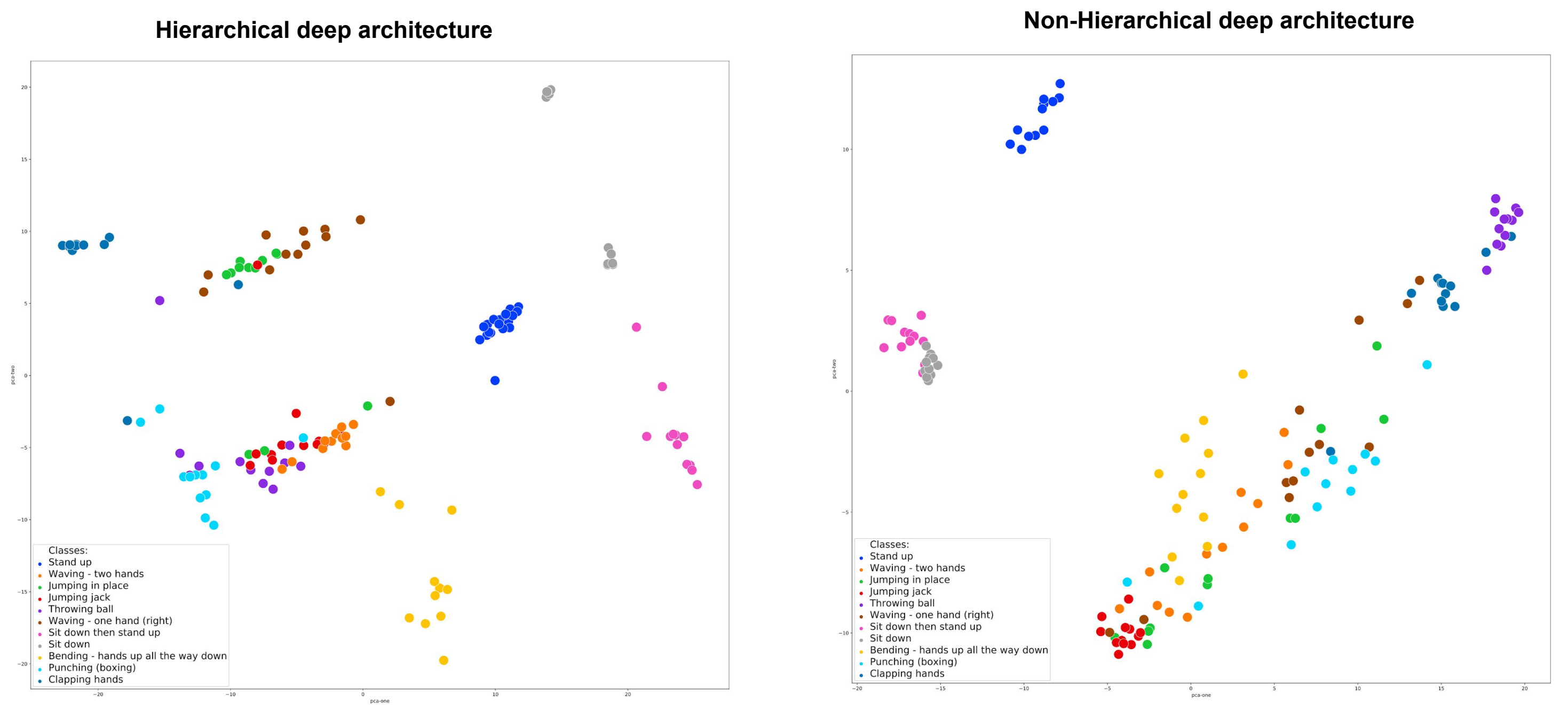


(a) Hierarchical DNN (b) Hierarchical and Non-Hierarchical DNN

- **Coarse fusion level selection & accuracy impact**: (a) late coarse (probabilistic) to mid level fine-grained, (b) mid level coarse to mid level fine-grained.

Architecture	Datasets (mAcc. (Coarse, Fine)%)		
Design	*MHAD*	*J-HMDB*	*MPII Cook*
H-BiLSTM	(82.50, 70.25)%	(45.68, 42.61)%	(60.70, 35.40)%
HFP-BiLSTM	(86.35, 65.46)%	(42.41, 39.55)%	(36.84, 28.19)%
H-I3D	(98.75, 96.38)%	(78.47, 76.10)%	(70.47, 54.30)%
HFP-I3D	(91.35, 82.89)%	(67.17, 60.46)%	(60.34, 37.55)%

- **Visualize learned representation:** using 1st and 2nd PCA components, MHAD dataset



CONCLUSIONS

- Linguistic analysis on action label sentences provides useful insights regarding potential correlations between labels.
- Exploiting linguistic similarities allows action decomposition in a coarse-to-fine scheme, reformulating the recognition to a multi-task learning problem.
- The hierarchical linguistic representation of the action, when mimicked in the deep NN design, leads to recognition accuracy increase.
- Exploiting the learned coarser representation in the finer learning process, allows for faster learning in the earlier learning epochs.
- Datasets with more complex and larger label sentence sets could potentially benefit more by a more elaborate linguistic analysis approach, that is able to express the underlying semantic content similarities.

FORTH
INSTITUTE OF COMPUTER SCIENCE
For more information
contact {kbach, argyros}@ics.forth.gr

We acknowledge the **NVIDIA** Corporation for the donation of a Quadro P6000 GPU.
This work was partially supported by **EU H2020** project **Co4Robots** (Grant No 731869).