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

Konstantinos Bacharidis, Antonis Argyros

Computer Science Department, University of Crete, Greece

and

Institute of Computer Science, FORTH

25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
Milan, Italy 10 | 15 January 2021

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ
UNIVERSITY OF CRETE

INSTITUTE OF COMPUTER SCIENCE

Problem Statement

ICPR 2020
25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
Milan, Italy 10 | 15 January 2021

- **HAR datasets** smaller and less diverse compared to image recognition datasets
  -> *training data bottleneck in deep neural network learning*
- **Data limitation**: Multi-modal model designs utilizing language, audio or other sensory information
- **Language- infused designs**:
  - Script data introduced to speech recognition models
  - Constrained to a limited action set ->Action datasets with script data are scarce
- **Action/ activity labels**: source of linguistic information *present in every dataset*
  - Contain motion motif(s), object presence, visual relationships
  - Motion motif commonalities -> common verbs or verbs with high semantic similarity
  - Object commonalities -> common nouns or nouns with high semantic similarity

Proposed method:
Define Action Granularity Tree from Action Labels

- **Part-of-speech detection**: use a part-of-speech (POS) tagger from the Natural Language ToolKit (NLTK), to classify words into lexical categories

- **Tag refinement**: refine with *syntax rules* to account for words with multiple semantic interpretations (e.g. screw : *noun/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, etc.)

  - ***From noun to verb***: acceptable action description format

    verb + adposition/particle + noun

- **Cluster** label sentences based on POS commonalities or high semantic content similarity.

  - Similarity : defined with a form of distance between the verb word embeddings in WordNet.

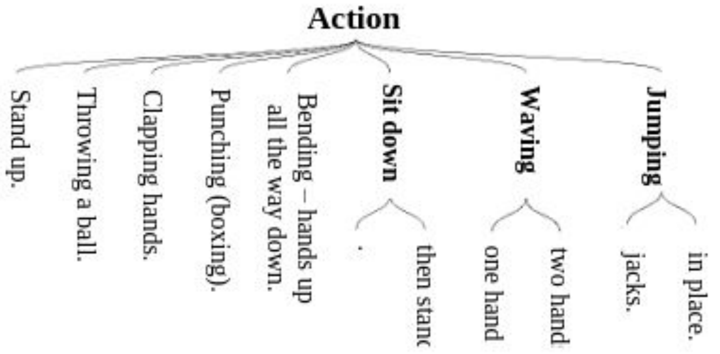


Fig. : Action hierarchy generated with the application of the proposed verb-centered lexical analysis on the class labels of the MHAD dataset, [*Jumping in place, Jumping jacks, Bending - hands up all the way down, Punching, Waving - two hands, Waving - one hand, Clapping hands, Throwing a ball, Sit down then stand up, Sit down, Stand up.*]

Incorporate Action Hierarchy in DNN designs

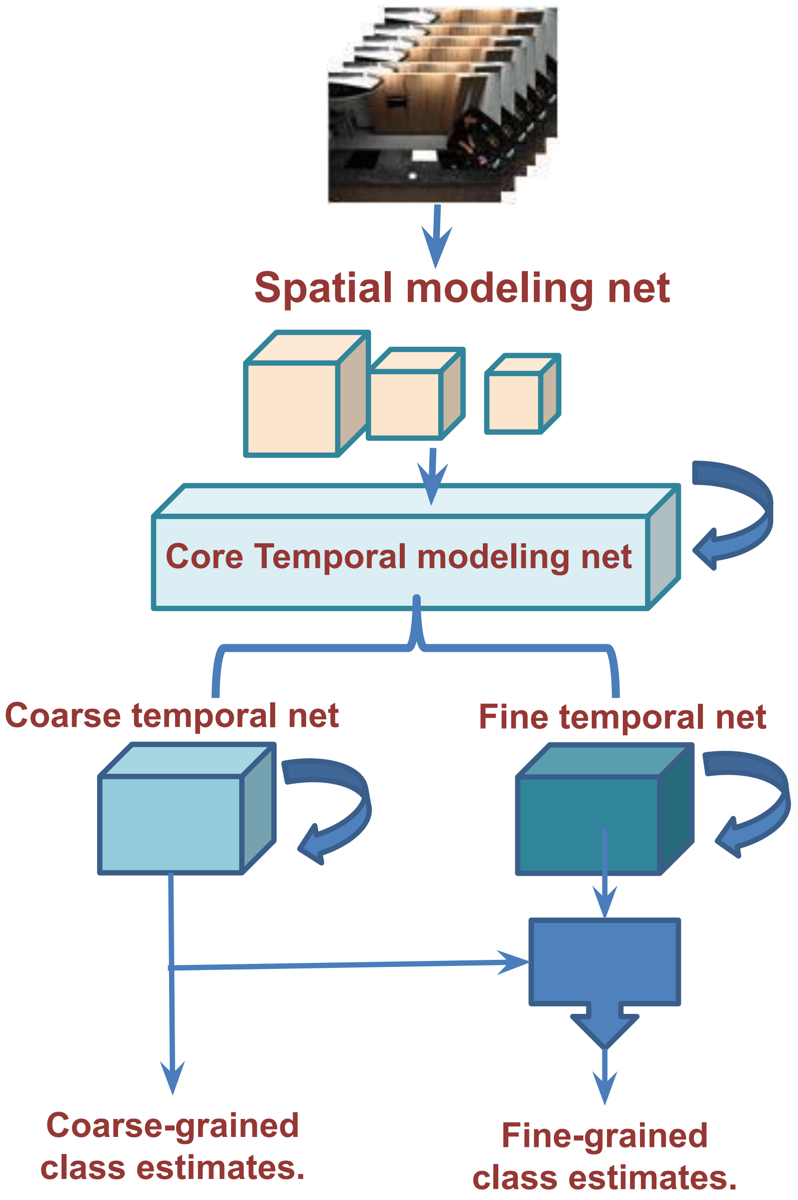
- **DNN design directions:**
  - <u>Modify</u> the temporal modelling sub-network.
  - <u>Mimic</u> the *N*-level action granularity with a set of subnets, one for each level (coarse-to-fine).
  - <u>Introduce</u> the learned representations of the coarser ones to finer sub-nets, using
    - *Skip-connections & feature vector concatenations*
- **Shallow action hierarchy** cost function:

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

with w_l : vector of label associations of the fine-grained action classes,

(T_n^{gn}, T_n^{fn}) : ground-truth labels for coarse- and fine-grained actions sets,

(Y_n^{gn}, Y_n^{fn}) : the estimated action classes



Experimental Results

- **Datasets:** MHAD (11-actions), J-HMDB (21-actions), MPII Cooking Activities (64-actions)

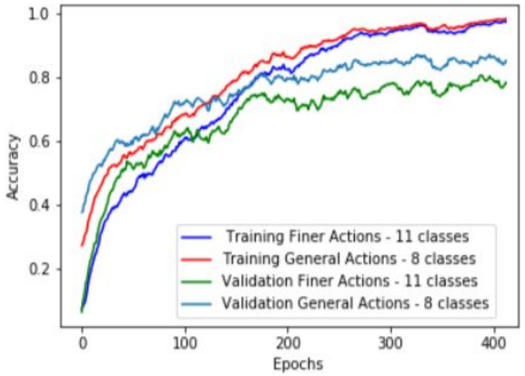
	Datasets		
	MHAD	*J-HMDB*	*MPII Cooking*
Num unique verbs	9 verbs	19 verbs	42 verbs
Avg num verbs/lbl	1.128 verb/lbl	1.0 verb/lbl	1.188 verbs/lbl
Avg lbl length	3.182 PoS/lbl	1.333 PoS/lbl	2.297 PoS/lbl
Avg asc via verb	0.545 asc/lbl	0.286 asc/lbl	1.656 asc/lbl
Max/min asc verb	1/0 asc	2/0 asc	5/0 asc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

- **Accuracy:**

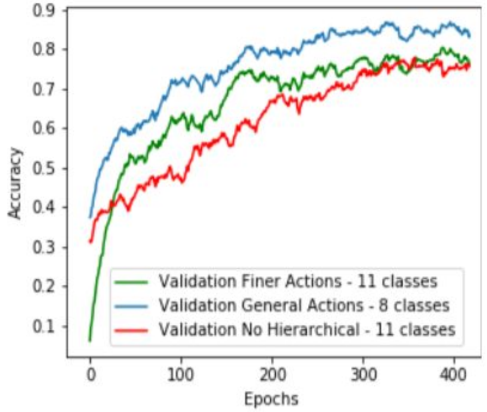
Architecture Design	**Datasets** (mAcc. (Coarse, Fine)%)		
	MHAD	*J-HMDB*	*MPII Cook*
NH-BiLSTM	(-, 64.17)%	(-, 36.28)%	(-, 29.45)%
H-BiLSTM	(82.50, 70.25)%	(45.68, 42.61)%	(60.70, 35.40)%
NH-I3D	(-, 89.61)%	(-, 72.38)%	(-, 48.18)%
H-I3D	(98.75, 96.38)%	(78.47, 76.10)%	(70.47, 54.30)%

TABLE: Action recognition performance for MHAD, JHMDB and MPII datasets between hierarchical (H) and non-hierarchical (NH) deep architecture designs.

- **Learning speed difference:**



(a) Hierarchical DNN

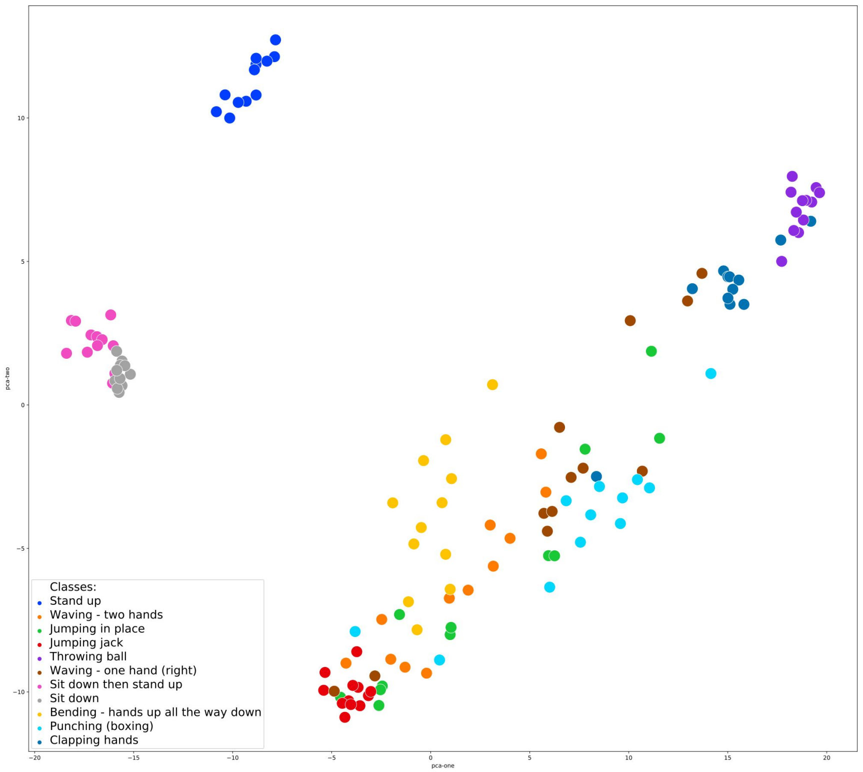


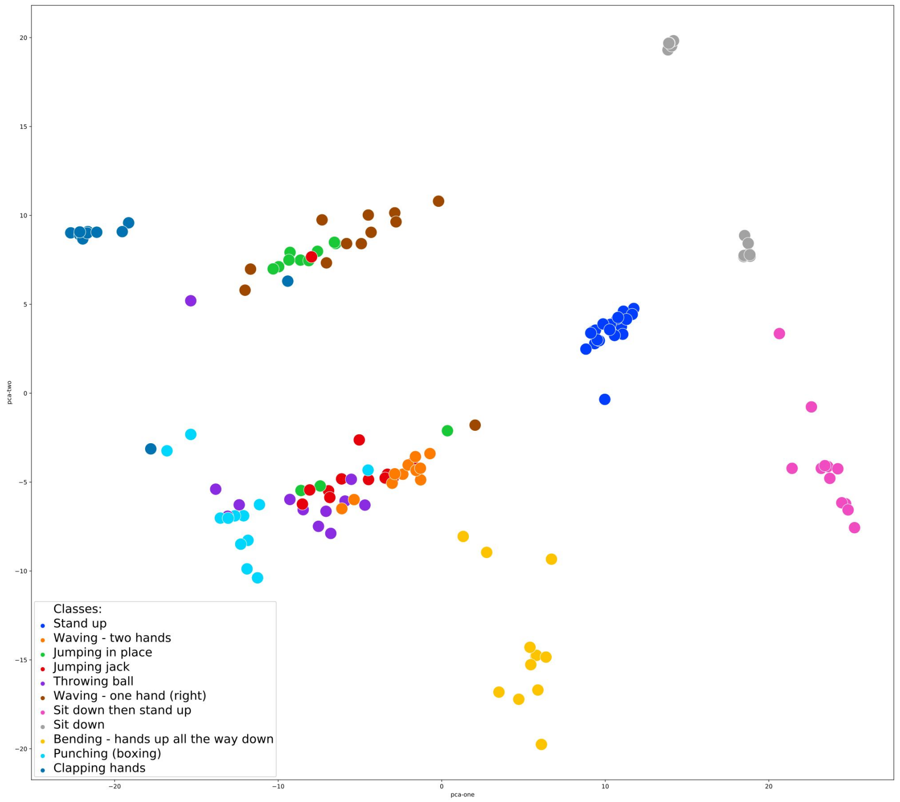
(b) Hierarchical and Non-Hierarchical DNN

- **Observations on the impact of hierarchical design:**
  - 4-6% score improvement in every deep model & dataset case
  - Increases learning speed in the earlier epochs of learning

Experimental Results

- **Visualization of learned representations with PCA:**





Classification layer, MHAD dataset (a) Non-Hierarchical , (b) Proposed Hierarchical model design

Conclusions & Future Work

Conclusions:

- **Action labels** contain a considerable amount of **action-related information**.
- Identifying action class similarities a priori using the linguistic description provides useful insights regarding action complexity and hierarchy.
- Mimicking this hierarchical structure in a deep model design, **leads to learning speed and accuracy improvement (up to 6%)**, compared to a non-hierarchical design.
- Despite the increase in the number of hyperparameters (+20-24%), learning at early stages is faster compared to a non-hierarchical design.

Future Work:

- Datasets with **complex actions require elaborate linguistic analysis** to capture the semantics contained in the action labels.
- Investigate the effect of increasing the levels of action granularity in
    - accuracy and learning speed
    - layer level fusion selection

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ
UNIVERSITY OF CRETE

FORTH
INSTITUTE OF COMPUTER SCIENCE

ICPR 2020
25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
Milan, Italy 10 | 15 January 2021

Thank you!

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ
UNIVERSITY OF CRETE

FORTH
INSTITUTE OF COMPUTER SCIENCE