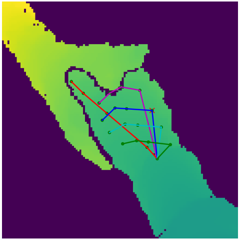
Imperial College London



Fusing Spatial and Temporal Models for Joint Hand Pose Estimation and Action Recognition

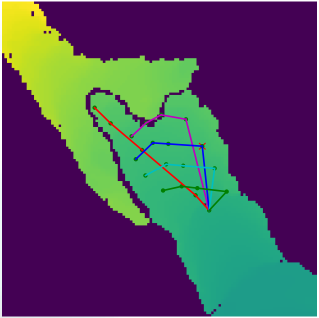
Haaris Mehmood

24 June 2019

Presentation outline

Imperial College London

Brief Introduction



Key Areas in HCI and CV Research

Hand pose (or joints) estimation (HPE): best performance using depth-maps [GYBK17]

Hand action recognition (HAR): best performance using pose [GYBK17]

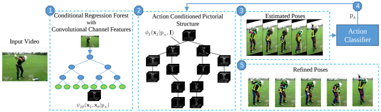
Objectives

Key Requirements

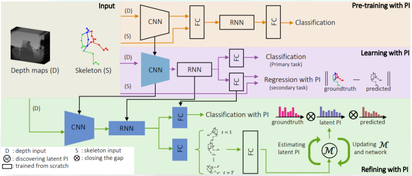
- Produce two outputs (pose and action) from a single input (depth)
- Improve predictions of one output using ‘cues’ from the other.
- Create a general framework that can make use of state-of-art methods from both HPE and HAR research.

Imperial College London

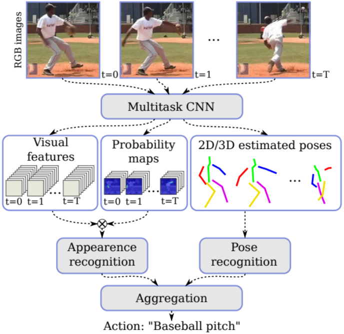
Prior Works



Pictorial Structure [IGG16]



Latent Space Refinement [SK17]



Multitask Learning [LPT18]

Imperial College London

Notation

Common Symbols

Depth: $\boldsymbol{U} \in \mathbb{R}^{W \times H}$, $\mathbf{U}_{seq} \in \mathbb{R}^{T \times W \times H}$

Hand Pose: $\boldsymbol{v} \in \mathbb{R}^{3J}$, $\boldsymbol{V}_{seq} \in \mathbb{R}^{T \times 3J}$

Action: $\boldsymbol{w} \in \mathbb{R}^{C}$

Hidden State: $\boldsymbol{h} \in \mathbb{R}^{L}$

Sequence Length: T, Width: W Height: H

Joints: J, Action Classes: C, Hidden Dimension L

Imperial College London

Choosing a Baseline

Solution

Concatenate HPE and HAR sequentially.

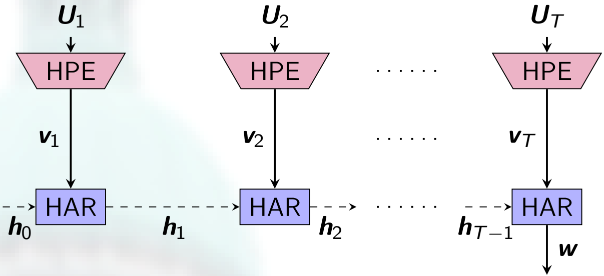
A 'well-known' general method to arbitrary fuse spatial and temporal models.

Suitable approach for many fields (action recognition, image captioning, video description) [DHG+15]

Pre-train strong HPE [OL17] and HAR baselines [ZLX+16] individually, then fine-tune as baseline model.

Imperial College London

Baseline Architecture



A time-unrolled view of the baseline architecture.

Imperial College London

Baseline Results

Baseline Variant	Pose Error	Action Accuracy
HPE	14.5mm	–
HAR (GT Pose)	–	72.3%
Untrained Baseline	14.5mm	59.0%
Trained Baseline	10.9mm	68.0%
Error Gap	–	4.3%

Imperial College London

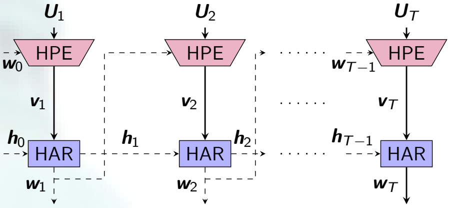
Extending the Baseline

Possible Extension

Can we feed the predicted action at $t = 1$ ($\boldsymbol{w}_1$), to improve pose predictions ($\boldsymbol{v}_2$) and action predictions ($\boldsymbol{w}_2$) for the next time step?

Imperial College London

Architecture



The architecture for the sequential action feedback method.

Key Concerns

How to extend HPE to accept conditioning?

How to pre-train such an HPE?

What proportion of (noisy) action to supply as feedback?

Imperial College London

Pre-training HPE

What strategy to use for such an HPE?

Ground truth action gives poor results with noisy action at fine-tuning

Our solution: supply redundant information as a 'place-holder' during pre-training

A hybrid mix of both leads to divergence

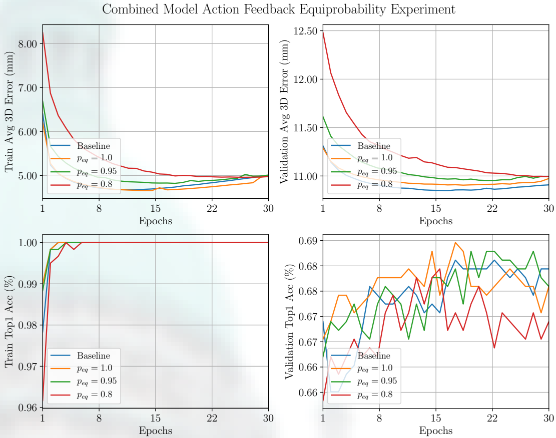
Solution

Only pre-train HPE using $\boldsymbol{w}_{eq} = [\frac{1}{45}, \ldots, \frac{1}{45}]$.

Use $\boldsymbol{w}_{eq}$ or $\boldsymbol{w}_{pred}$ randomly based on probability p_{eq} during fine-tuning.

Imperial College London

Results



Varying p_{eq} during fine-tuning, results as worse for lower p_{eq}

Imperial College London

Results

Why is low p_{eq} bad?

Error gets propagated through time, thus positive feedback is likely.

Too low p_{eq} cannot be used as pre-training done using $p_{eq} = 1.0$.

Next steps

What if action is supplied only at the end of the sequence?

How can we better utilise the two HPEs?

Imperial College London

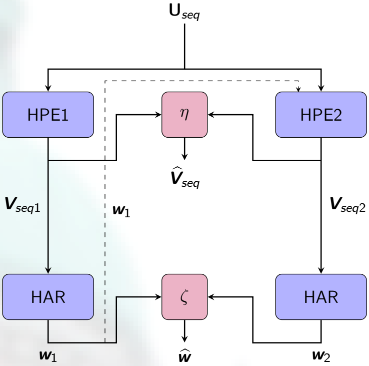
A New Approach

Solution

- Make use of two HPEs in an ensemble setting.
- Improve robustness of HAR by showing it variations of input for the same output
- Improve feedback method by supplying at the end of sequence and also to only one HPE.

Imperial College London

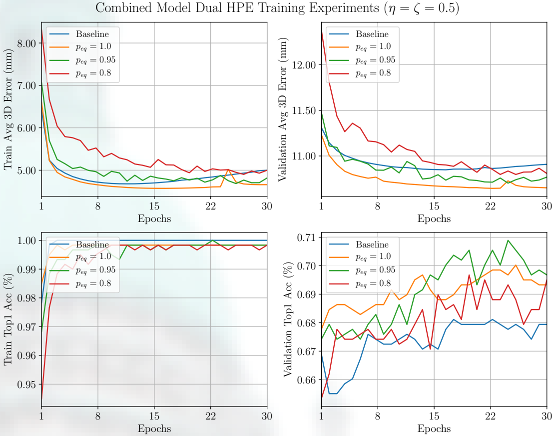
Architecture



The proposed improved architecture

Imperial College London

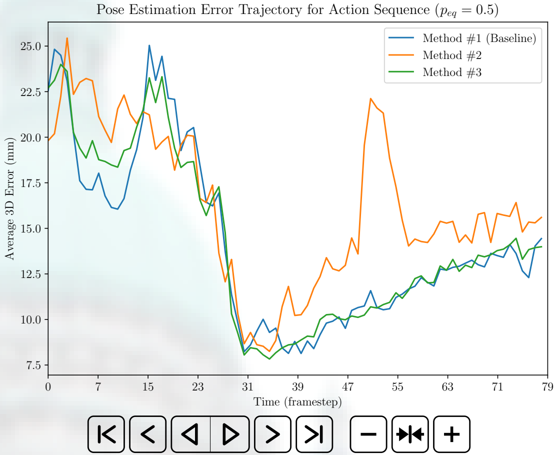
Results



Training and validation curves for different p_{eq} values

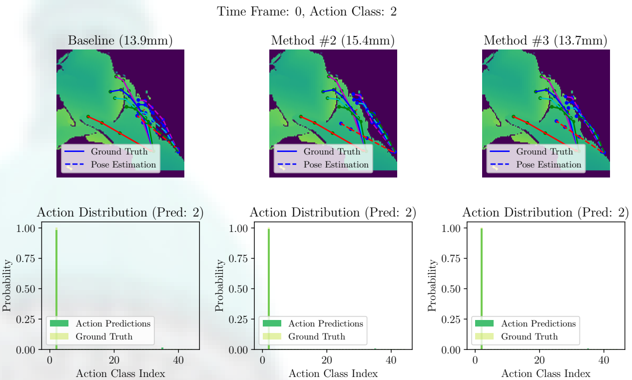
Imperial College London

Effect of action feedback on 3D error of a sequence



Animation best viewed in Adobe Reader

Visual Results



Changes in action sequence predictions and pose estimation

Animation best viewed in Adobe Reader

Conclusion

Summary of Results

Model	Error (mm)	Accuracy (%)
HAR Standalon (GT Pose)	–	72.3
HPE Standalone	14.49	–
Baseline (No Train)	14.49	59.0
Baseline (Train)	10.87	68.0
Our Method #1	10.95	68.2
Our Method #2	**10.69**	**71.3**

Imperial College London

Conclusion

What's Next?

Action feedback is a non-trivial problem, main issue is how to train the model to accept noisy action in a stable way.

Future work can include injecting noise into the action vectors.

Feedback is a worthwhile idea to pursue further (improvements shown).

The two-part model (HPE+HAR) can naturally lead to implicit data augmentation for downstream tasks (HAR).

Improvements are shown in a general way without specifying what HPE or HAR to use.

Imperial College London

Thank you

Imperial College London

Accepting Action Feedback

Requirements

Need to modify HPE in the least intrusive and most general way possible

Solution

Use ‘feature-wise linear modulation’ [PSdV+17] as a general method.

Given input $\boldsymbol{x}$ and conditional vectors γ & $\boldsymbol{\beta}$, $\boldsymbol{y} = \gamma \odot \boldsymbol{x} + \boldsymbol{\beta}$

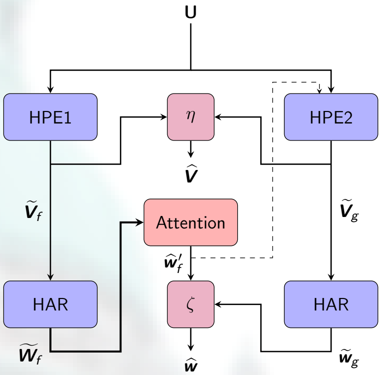
Temporal Attention

Refining Action

Can we further refine action information to only focus on the most discriminative action vectors?

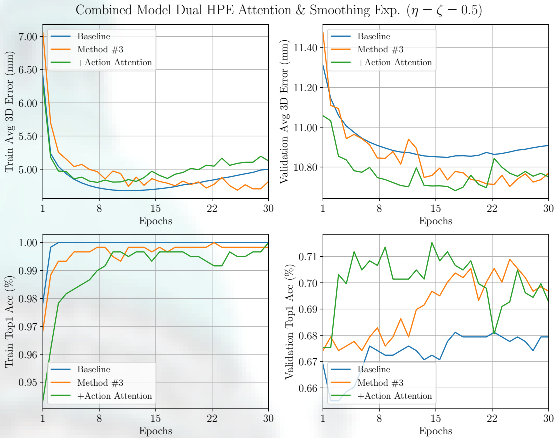
Imperial College London

Adding Attention...



Imperial College London

Adding Attention...



Effect of using temporal attention. Final reported scores are with attention.

Imperial College London

References I

J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, T. Darrell, and K. Saenko, "Long-term recurrent convolutional networks for visual recognition and description", in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 07-12-June, 2015, pp. 2625–2634, ISBN: 9781467369640. DOI: 10.1109/CVPR.2015.7298878. arXiv: 1411.4389.

Imperial College London

References II

G. Garcia-Hernando, S. Yuan, S. Baek, and T.-K. Kim, "First-Person Hand Action Benchmark with RGB-D Videos and 3D Hand Pose Annotations", , Apr. 2017. arXiv: 1704.02463.

U. Iqbal, M. Garbade, and J. Gall, "Pose for Action - Action for Pose", , Mar. 2016. arXiv: 1603.04037.

D. C. Luvizon, D. Picard, and H. Tabia, "2d/3d pose estimation and action recognition using multitask deep learning", in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, 2018. DOI: 10.1109/CVPR.2018.00539. arXiv: 1802.09232.

Imperial College London

References III

M. Oberweger and V. Lepetit, "DeepPrior++: Improving Fast and Accurate 3D Hand Pose Estimation", in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, vol. 2018-Janua, IEEE, Oct. 2017, pp. 585–594, ISBN: 978-1-5386-1034-3. DOI: 10.1109/ICCVW.2017.75. arXiv: 1708.08325.

E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. Courville, "FiLM: Visual Reasoning with a General Conditioning Layer", , 2017. arXiv: 1709.07871.

Imperial College London

References IV

Z. Shi and T. K. Kim, "Learning and refining of privileged information-based RNNs for action recognition from depth sequences", in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-Janua, 2017, pp. 4684–4693, ISBN: 9781538604571. DOI: 10.1109/CVPR.2017.498. arXiv: 1703.09625.

Imperial College London

References V

W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, and X. Xie, “Co-occurrence Feature Learning for Skeleton based Action Recognition using Regularized Deep LSTM Networks”, , Mar. 2016, ISSN: 1938-2367. DOI: 10.1007/SpringerReference_61203. arXiv: 1603.07772.