Multi-task human analysis in still images: 2D-3Dpose, depth map, and multi-part segmentation

Daniel Sánchez [1], Marc Oliu [1], Meysam Madadi [2], Xavier Baró [1,2], Sergio Escalera [2,3]

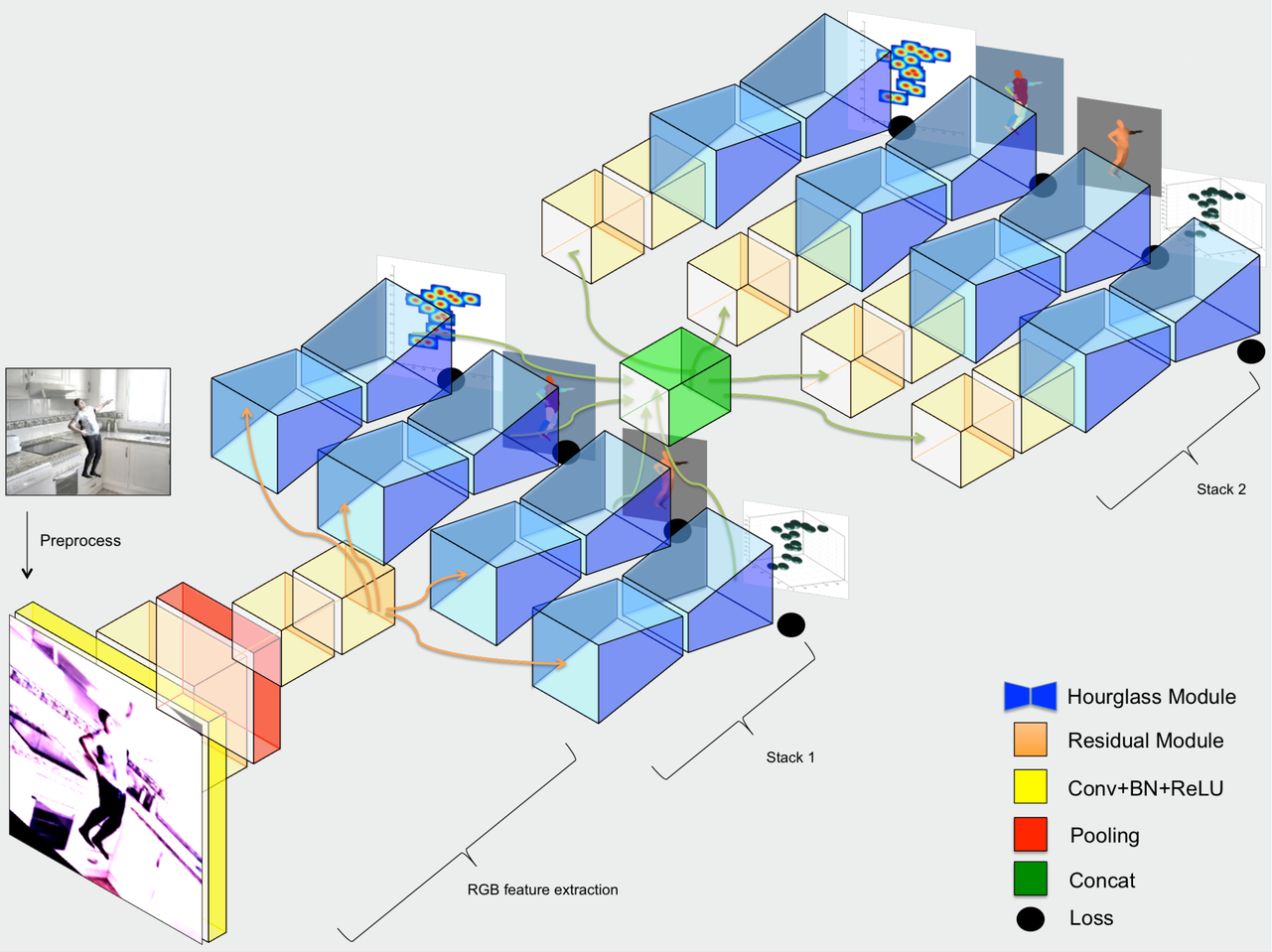
[1] Faculty of Computer Science, Multimedia and Telecommunication - Universitat Oberta de Catalunya, Spain,
[2] Computer Vision Center - Universitat Autonoma de Barcelona, Spain,
[3] Department of Mathematics and Informatics - Universitat de Barcelona, Spain

Abstract and Motivation

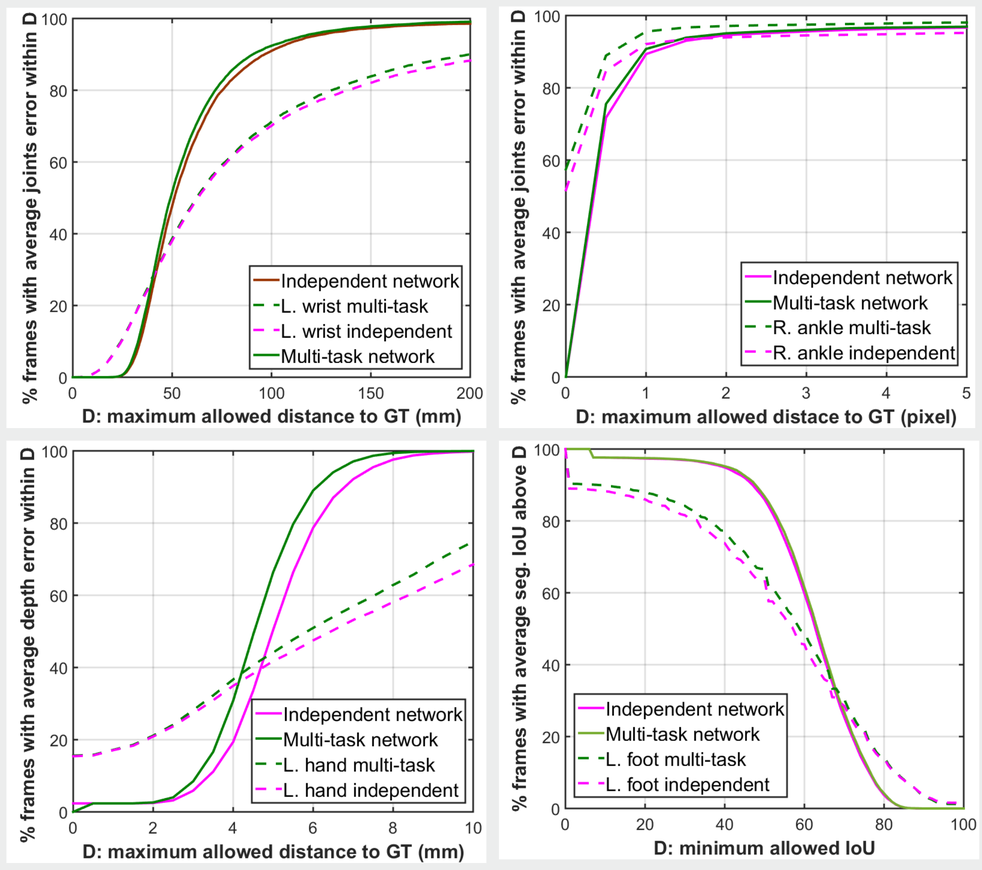
- In this work, we analyze **4** related human analysis tasks in still images in a **multi-task** scenario by leveraging such datasets. Specifically, we study the correlation of **2D/3D pose estimation**, **body part segmentation** and **full-body depth estimation**.
- While many individual tasks in the domain of human analysis have recently received an accuracy boost from deep learning approaches, multi-task learning has mostly been ignored due to a lack of data. New **synthetic datasets** are being released, **filling this gap** with synthetic generated data. Following recent work of **SURREAL dataset** [1], a new large-scale dataset consisting of realistic synthetic data, we evaluate all multi-task combinations in an end-to-end architecture.
- Finally, we present an analysis how training together these 4 related tasks **can benefit each individual task** for a better generalization.

Proposed Model

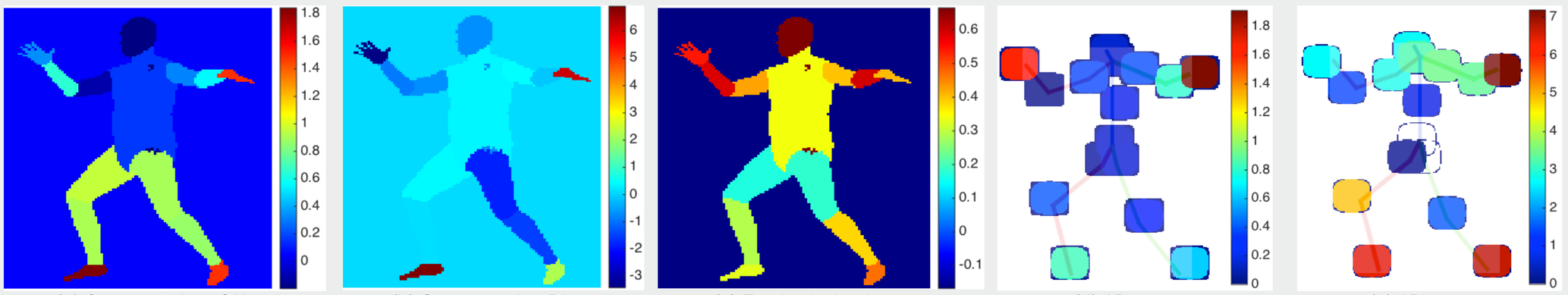
The proposed model combines 2D/3D pose, body part segmentation and full-body depth. These tasks are learned via the well-known Stacked Hourglass [2] module such that each of the task-specific streams shares information with the others. In this way, the outputs can be refined sequentially.



Experimental Results



Success rate error for the different tasks. For each task: isolated task vs best multi-task approach; and for joint/part with highest multi-task improvement, its isolated task vs multi-task score. From left to right: 3D pose, 2D pose, Full-body Depth map, Body Parts segmentation.



(a) Segmentation IOU: seg. vs 2D+seg.+depth

(b) Segmentation Pixel Accuracy: seg. vs 2D+seg +depth

(c) Full-body depth: depth vs 3D+seg.+depth

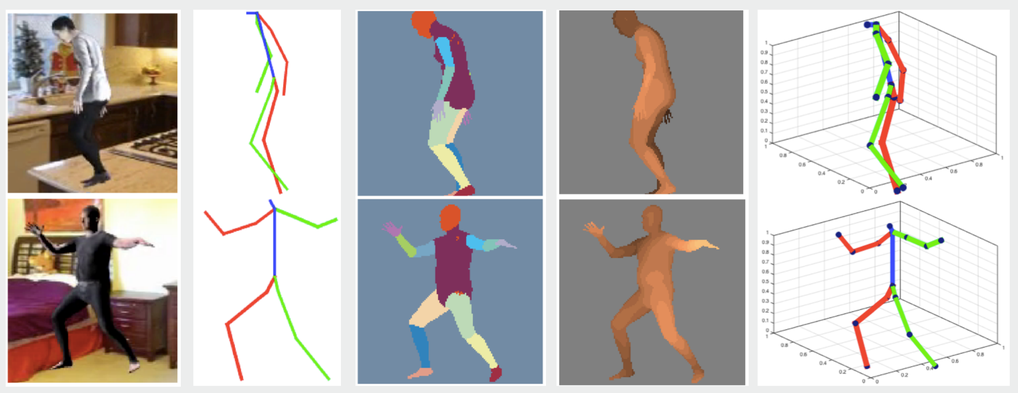
(d) 2D pose: 2D vs 2D/3D+seg.+depth

(e) 3D pose: 3D vs 3D+seg.+depth

Error visualization per each body part and task. The higher the value the higher the performance improvement for a particular metric of the best multi-task model compared to the baseline isolated task.

	Seg. (IoU)	2D pose (PCKh)	3D pose (MJD mm)
Varol et al [3] independent tasks	59.2	82.7	46.1
Varol et al. [3] multi-tasks	**69.2**	90.8	**40.8**
Ours - independent tasks	65.3	96.5	60.1
Ours - multi-tasks	66.1	**97.0**	57.0

STATE-OF-THE-ART COMPARISON ON SURREAL



(a) RGB (b) 2D pose (c) Body Parts (d) Depth (e) 3D pose

Samples from SURREAL dataset with the chosen modalities.

Conclusions

Results show 4 tasks benefit from the multi-task approach, but with different combinations of tasks: while combining all four tasks improves 2D pose estimation the most, 2D pose improves neither 3D pose nor full-body depth estimation. On the other hand, body parts segmentation can benefit from 2D pose but not from 3D pose. In all cases, as expected, the maximum improvement is achieved on those human body parts that show more variability in terms of spatial distribution, appearance and shape, e.g. wrists and ankles.

[1] G. Varol, J. Romero, X. Martin, N. Mahmood, M. Black, I. Laptev,C. Schmid. Learning from synthetic humans. CVPR, 2017.
[2] Newell, K. Yang, J. Deng. Stacked hourglass networks for humanpose estimation. ECCV, 2016.
[3] G. Varol, D. Ceylan, B. Russell, J. Yang, E. Yumer, I. Laptev, C.Schmid. Bodynet: Volumetric inference of 3d human body shapes.arXiv:1804.04875, 2018.