International Journal of Computer Vision (2022) 130:199–200
https://doi.org/10.1007/s11263-021-01548-7

EDITORIAL

Preface to the Special Issue on Human Pose, Motion, Activities and Shape in 3D

Manuel J. Marín-Jiménez[1] · Javier Romero[2] · Hao Li[3] · Grégory Rogez[4]

Accepted: 6 November 2021 / Published online: 4 January 2022

Current computer vision algorithms and deep learning-based methods can detect people in images and estimate their 2D pose with remarkable accuracy. However, understanding humans and estimating their pose and shape in 3D is still an open problem. The ambiguities in lifting 2D pose to 3D, the lack of annotated data to train 3D pose regressors in the wild and the absence of a reliable evaluation dataset in real world situations make the problem very challenging.

Since very recent papers achieved impressive results in tasks like body reposing using purely 2D based techniques, we also wanted to challenge the need of explicit 3D techniques and data in such computer vision problems related to humans.

Therefore, in this special issue we aimed at gathering high quality publications on 3D human pose estimation from RGB images and videos, and related topics such as 3D human shape estimation from images or activity recognition from 3D skeletal data.

This Special Issue received a total of 19 submissions, where 10 have been accepted for publication after a rigorous review process. All selected papers underwent at least one round of revisions. More than 60 reviewers have participated in the selection of the papers—we thank all of them for their great contribution.

The selected papers of this Special issue cover a variety of topics including 3D motion capture, with (Chatzitofis et al.) and without markers (Chen et al.), and its temporal prediction (Mao et al.); human action recognition from 2D images by using synthetic 3D humans (Varol et al.) or from skeleton data (Gupta et al.); monocular 3D human pose estimation (Wang et al.; Liu et al.) and body reconstruction (Madadi et al.); 3D sign language production (Saunders et al.); and 3D face registration (Bahri et al.).

We briefly summarize the contribution of each accepted paper below.

In the paper "DeMoCap: Low-Cost Marker-Based Motion Capture," by Chatzitofis et al., a new human motion capture system is proposed that combines traditional optical marker-based techniques with deep learning approaches, based on low-cost depth-sensors. To resolve the limitations of previous approaches, such as erroneous capture due to the marker occlusions or mislabeling from marker swapping, the proposed approach directly regresses the 3D pose from multi-view depth maps in a differentiable way.

In "SportsCap: Monocular 3D Human Motion Capture and Fine-grained Understanding in Challenging Sports Videos" by Chen et al., the authors propose a motion capture system from regular sports video. Unlike previous methods, the authors estimate both 3D human pose and fine-grained action simultaneously. To achieve this, the authors propose a new dataset composed of 110,000 frames from 640 videos, manually annotated with skeleton joints, their visibility, semantic attributes and sub-motions. The system is evaluated in their proposed dataset as well as existing ones, comparing favorably to state-of-the-art systems.

The paper "Multi-level Motion Attention for Human Motion Prediction" by Mao et al. exploits human motion repeatability to improve motion prediction. More specifically, they introduce "motion attention" which compares motion sub-sequences instead of individual frames. To generalize better, this attention is implemented at different granularity levels (i.e. pose, part and joint motion attention).

✉ Manuel J. Marín-Jiménez
mjmarin@uco.es

Javier Romero
javier.romero@tuebingen.mpg.de

Hao Li
hao@hao-li.com

Grégory Rogez
gregory.rogez@naverlabs.com

[1] University of Cordoba, 14071 Córdoba, Spain

[2] Facebook Reality Labs, 08026 Barcelona, Spain

[3] Pinscreen Inc., UC Berkeley, Berkeley 90025, USA

[4] NAVER LABS Europe, 38240 Meylan, France

The presented method outperforms previous methods which restrict their attention models to individual frames.

Varol et al. proposes to generate synthetic data in order to learn action recognition models that better generalize to unseen viewpoints in "Synthetic Humans for Action Recognition from Unseen Viewpoints." Instead of rendering recorded MoCap sequences of the target actions, the authors leverage recent advances in 3D human pose estimation to obtain synthetic data by re-rendering SMPL sequences estimated from real videos with variations of viewpoints, cameras, background, human texture and shape.

In the paper "Quo Vadis, Skeleton Action Recognition?," by Gupta et al., the authors present a study on skeleton-based human action recognition in the wild. They introduce three new datasets: Skeletics-152, Skeleton-Mimetics and Metaphorics. State-of-the-art models are evaluated on the existing NTU-120 dataset and selected to evaluate their performance on non-controlled scenarios. The results on these new datasets indicate that further work is needed to solve the problem of skeleton-based human action recognition in the wild.

"Learning a Robust Part-aware Monocular 3D Human Pose Estimator via Neural Architecture Search," by Wang et al., presents a novel part specific model for monocular 3D pose estimation which is learned via neural architecture search. The searched model is efficient and compact and can automatically select a suitable decoder architecture to estimate each human body part.

In the paper "View-Invariant, Occlusion-Robust Probabilistic Embedding for Human Pose" by Liu et al., the authors propose to solve the fundamental ambiguity of lifting a 2D pose to 3D, by mapping a 2D pose not to a single latent point but to a whole distribution using a probabilistic representation in a compact view-invariant embedding space. The paper demonstrates the importance of the proposed approach in different applications, including 3D pose retrieval in multi-view settings, action recognition and video alignment.

Madadi et al. proposes a method for estimating the complete body surface from a sparse set of landmarks in "Deep unsupervised 3D human body reconstruction from a sparse set of landmarks." The system achieves this by estimating the parameters of a generative model (SMPL) from a complete skeleton and a set of landmarks, inferred from an incomplete and noisy input. The method works in an unsupervised manner, meaning that it does not require pairs of landmarks and the corresponding body model parameters. It achieves comparable accuracy to previous optimization-based approaches (i.e. Mosh++) with a substantially smaller runtime.

The paper "Continuous 3D Multi-Channel Sign Language Production via Progressive Transformers and Mixture Density Networks," by Saunders et al., addresses the problem of sign language production (SLP). Given spoken language sentences, SLP tries to generate a valid sequence of signs (face, arm and hand posture combinations), in the target sign language, that looks natural and that can be understood by deaf people. In this work, an end-to-end model for SLP, coined Progressive Transformer, is able to handle variable length continuous sequences. It is evaluated on the challenging PHOENIX14T dataset. Although the model improved existing gloss-based models, the authors conclude that current sign productions still need improvement to be fully understandable by the deaf community.

Finally, in the paper "Shape My Face: Registering 3D Face Scans by Surface-to-Surface Translation," by Bahri et al., a new deep learning approach for 3D surface registration is proposed for human faces. The authors devise an auto-encoder model based on 3D point cloud convolutions and equipped with an attention module. This network solves the registration problem by learning a face representation at the same time.

Collectively, these ten papers illustrate the diverse range of issues associated with the field of automatic analysis of 3D humans in images and video, including 3D human pose estimation, motion capture, action recognition, sign language production and face registration.

We thank all the authors for participating in this special issue, and all the reviewers for their constructive suggestions. We specially thank the IJCV editorial team, and in particular Katherine Moretti, for their excellent support in producing this special issue.

Manuel, Javier, Hao and Gregory—October, 2021.