SCAT: Stride Consistency with Auto-regressive regressor and Transformer for hand pose estimation

Daiheng Gao[1], Bang Zhang[1], Qi Wang[1], Xindi Zhang[2], Pan Pan[1], Yinghui Xu[1]
[1]DAMO Academy, Alibaba Group [2]Queen Mary University of London

2021 ICCV OCTOBER 11-17 VIRTUAL

Abstract

We propose a 3D hand pose estimation approach: SCAT, which does not depends on any parametric hand models. SCAT offers a new representation for measuring hand poses meanwhile ensures inter-frame smoothness and removing motion jittering. Extensive experiments show that our SCAT can generate more accurate and smoother 3D hand pose estimation results compared with the SOTA methods.

Method

Although the dominating trend of using 3D parametric model (MANO) is unstoppable, this paper aims at one point: does model like MANO indeed helps a lot to hand pose estimation result? Since MANO is *"learned from around 1000 high-resolution 3D scans of hands of 31 subjects in a wide variety of hand poses"* Through experiments on the MANO-based model [37, 50] in complex gestures: hand heart, metal horns and etc, we found it is difficult for the MANO-based model to reconstruct a satisfactory gesture. The reason in obtaining an inferior estimation comes down to the limited and monotonous exemplars that are used to construct the parametric model.

(1) Inspired by METRO [27], we use transformer encoder in Constructing meaningful and mighty interrelationship between hand joints. we obtain coarse predictions C_{coarse} through adding the transformer encoder's output (offset O) on pre-defined 21 basis coordinates (mean M) extracted from a standard hand template mesh. Through this mean-plus-offset strategy, we obtain a biomechanical-plausible prediction that is empowered by M.

(2) Moreover, we present a novel pose length regularization loss that encourages good conditioning in the mapping from feature maps F_i to the offsets of 3D hand pose O_i, which is the stride consistency as we proposed in title.

(3) Following the successful practice of HMR [21], we develop a coarse-to-fine strategy to obtain the fine-grained 3D predictions.

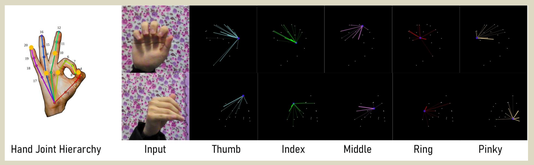

Fig 1 Visualization of relationship between joints through choosing representative joint of each finger: 4 (for thumb), 5 (for index), 10 (for middle), 13 (for ring), 20 (for pinky), where brighter color indicates stronger correlations.

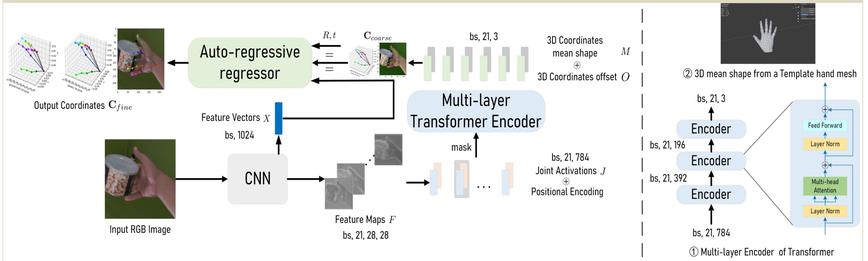

Fig 2 Overview of the proposed framework. SCAT receives a single RGB image and yields the feature maps F and feature vector X through a trainable CNN. Then we flatten the 2D feature maps to 1D feature vector J of fixed 21 joints, after performing positional encoding to J, a multi-layer transformer encoder is used to regress the 3D coordinates offset O, the structure of which is depicted in the right. Next, we extract the mean shape M from a pre-defined hand mesh template, after adding the offset O to the mean shape M , the satisfied 3D coordinates of 21 key joints are obtained by an auto-regressive manner.

Experiments

I. Compare with current SOTA

Method	AUC of PCK ↑		
	DO [39]	STB [49]	RHD [51]
Ge *et al.* [11]	-	**0.998***	0.920
Yang *et al.* [47]	-	0.996	0.943
Baek *et al.* [2]	0.650	0.995	0.926
Z & B [51]	-	0.948	0.675
Xiang *et al.* [46]	0.912	0.994	-
Zhou *et al.* [50]	0.948	0.898	0.856
Spurr *et al.* [38]	0.820	-	0.920
Rong *et al.* [37]		0.992	0.934
Li *et al.* [26]	0.860	0.996	**0.960***
Ours$_{coarse}$	0.892	0.977	0.915
Ours$_{fine}$	**0.951**$^{\Delta,*}$	**0.994**$^{\Delta}$	**0.954**$^{\Delta}$

Tab 1 Comparison with state-of-the-art methods on three public datasets (DO, STB, RHD).

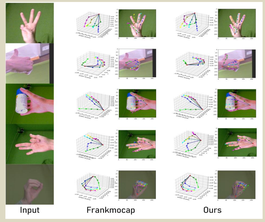

Fig 3 Comparison to the current method Frankmocap.

II. Ablation Study

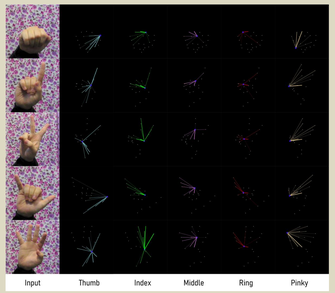

Fig 4 Visualization of inter-joint relationship.

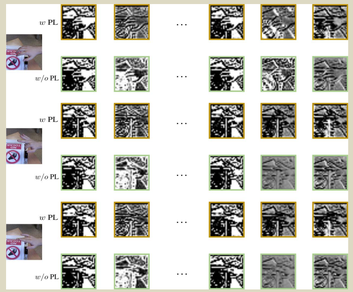

Fig 5 Ablation study of Pose Length (PL) Regularization. PL Regularization is serve to remove motion jittering, which is in accord with stride consistency.

Conclusion

We propose a simple yet effective method for 3D hand pose estimation from a single RGB image, SCAT. By utilizing a simple mean shape of a template hand mesh and the strong correlation modeling capacity bring from the transformer encoder, a reasonable and reliable 3D hand pose is predicted. Futhermore, we come up with novel pose length regularization in the pose estimation field to ensures a smoother prediction through time went on, without any temporal priors needed, which greatly enhanced our frame-based pose estimation method. We do lot of ablation study and readers can refer to the paper for more detail.