

A MULTI-TASK LEARNING APPROACH ENHANCED BY ACTION PROMPTS AND PRIOR KNOWLEDGE FOR 3D HUMAN POSE ESTIMATION

Hong-Wen Chen, Chow-Sing Lin*

Dept. of Computer Science and Information Engineering, National Univ. of Tainan, Tainan, Taiwan(R.O.C)

*mikelin@mail.nutn.edu.tw

Abstract

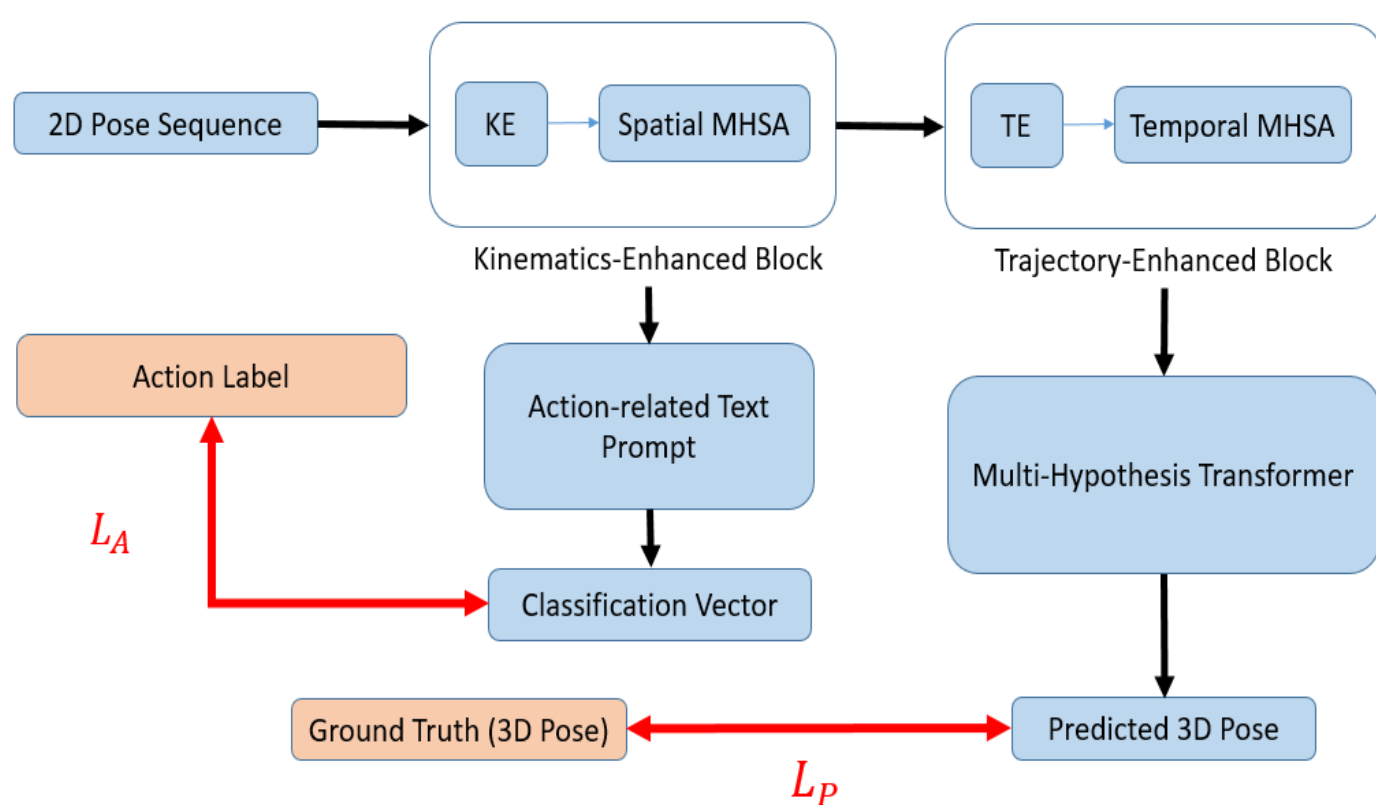
Estimating 3D human pose (3D HPE) from monocular video is challenging due to inherent ambiguity. We introduce the Action-Prompt Enhanced Multi-hypothesis Transformer (APEMT), a novel multi-task framework designed to boost 3D HPE accuracy by integrating rich semantic priors. APEMT employs a Transformer-based architecture with two core enhancements: 1) Kinematic and Trajectory Prior Modules ensure generated pose hypotheses are physically plausible. 2) An auxiliary task uses CLIP semantic prompting to fuse action features with external knowledge from textual prompts, providing robust contextual guidance. Preliminary results on the Human3.6M dataset, using CPN-detected 2D inputs, demonstrate the efficacy of APEMT and superior structural accuracy, achieving a P-MPJPE of 48.0mm.

Results

Table 1 Preliminary Performance Evaluation of the APEMT Framework on the Human3.6M Dataset

Model	MPJPE(P1)	P-MPJPE(P2)	N-MPJPE(P3)
APEMT	62.4mm	48.0mm	60.4mm

We present the preliminary result of our APEMT model on the Human3.6M dataset, using CPN-detected 2D inputs. As detailed in Table 1, the model achieves an MPJPE (P1) of 62.4mm. Crucially, it demonstrates superior structural accuracy with a significantly lower P-MPJPE (P2) of 48.0mm. This substantial 14mm reduction between P1 and P2 strongly validates the effectiveness of our integrated kinematic motion priors and CLIP semantic guidance in capturing correct human posture and exhibiting robustness to input noise. Furthermore, the N-MPJPE of 60.4 mm confirms the model's stable shape reconstruction ability regardless of body scale. Collectively, these results demonstrate APEMT's efficacy in generating structurally plausible 3D pose estimates.



Overall architecture of APEMT

Loss Function

$$\mathcal{L}_{total} = \alpha \mathcal{L}_P + \beta \mathcal{L}_A$$

$$\mathcal{L}_P = \| P_{final} - P_{GT} \|_2$$

$$\mathcal{L}_A = \text{CrossEntropy}(C_{pred} - C_{GT})$$

Conclusion

In this paper, we introduce the APEMT, a novel dual-task framework for 3D Human Pose Estimation (HPE). Our model enhances accuracy by integrating kinematic and trajectory priors to enforce physical plausibility, and uses CLIP semantic prompting via an auxiliary task to provide high-level contextual constraints. Preliminary result on Human3.6M confirms the efficacy of APEMT. Specifically, the low P-MPJPE validates that the combined guidance successfully regularizes the ambiguous 3D space, yielding structurally accurate and robust pose predictions.

Acknowledgements

This research work was partially supported by National Science And Technology Council under grand no. NSTC 113-2221-E-024 -006.