Qeios

Peer Review

Review of: "Optimizing Human Pose Estimation Through Focused Human and Joint Regions"

Qilin Li[1]

1. Institute of Electrical and Electronics Engineers (IEEE), Piscataway, United States

The manuscript presents a novel dual-stream framework, VREMD, for video-based human pose estimation, focusing on enhancing visual representations and disentangling motion dynamics. The proposed methods show significant improvements over existing state-of-the-art techniques on multiple benchmark datasets. However, there are areas that could be further clarified or improved to enhance the manuscript's quality and impact.

1. The description of the Human-Keypoint Mask Enhanced (HKME) module and the Bidirectional Motion Disentanglement (BMD) module is quite detailed. However, the explanation of the deformable cross attention mechanism could be further clarified. Specifically, the mathematical formulation in Equations (3) and (4) might benefit from a more intuitive explanation or a diagram illustrating the process. This would aid readers in understanding how the mechanism selectively focuses on regions centered at the target person's body.
2. The manuscript provides a comprehensive comparison with state-of-the-art methods in Tables 1, 2, and 3. However, the discussion of these results is somewhat brief. The authors should elaborate more on why their method outperforms others, highlighting specific contributions (e.g., the effectiveness of the dual-mask mechanism or the bidirectional separation strategy) that lead to these improvements.
3. The ablation studies in Tables 4, 5, and 6 are valuable in demonstrating the impact of individual components. However, the presentation could be enhanced by discussing the implications of these results in more depth. For example, in Table 5, the authors could explain why the combination of the human mask and the keypoint mask leads to the best performance, providing insights into the synergistic effects of these components.

 doi.org/10.32388/OGMH9J

4. The implementation details section mentions the use of a Vision Transformer (ViT-L) backbone pretrained on the COCO dataset. It would be beneficial to specify the exact configuration of the ViT-L model used (e.g., number of layers, attention heads) to ensure reproducibility. Additionally, details about the training schedule (e.g., batch size, learning rate schedule) could be provided to give a more complete picture of the experimental setup.
5. The loss function described in Equation (5) is standard for pose estimation tasks. However, the authors could explore the impact of using more advanced loss functions, such as focal loss or wing loss, which have been shown to improve keypoint localization accuracy in previous works. A comparison of the performance with different loss functions could provide additional insights into the robustness of the proposed method.

Declarations

Potential competing interests: No potential competing interests to declare.