Spatio-temporal tendency reasoning for human body pose and shape estimation from videos

Reviewer #1

Weaknesses 1: The authors failed to cite as well as compare with key recent and highly relevant papers (Wei et. al. and Lee et. al.) solving same task and claim to perform superior to the proposed method.

A1: Thanks for your review. We note that the method of Lee and Wei et al. outperforms our method in reconstruction accuracy, but are poorer in smoothness. Although our tasks is the same, our method is designed to be practical for situations such as extreme lighting. Meanwhile, we were unable to obtain qualitative results for Lee and Wei et al.'s method because their code are not publicly available and incomplete. Second, the training methods are different. Wei et. al. and Lee et. al. additionally used the AMASS dataset annotated with SMPL for training. But in our method, no AMASS dataset is involved in training.

Weaknesses 2: The proposed modules are not well motivated/justified in terms of architectural choices and seems quiet bulky while no discussion was given on computation time needed to train.

A2: Thanks for your review. Sorry, we will add the intuition of modular design to the paper. TTR increases the receptive field through the residual connection of different fragments. For example, for F_4, the equivalent receptive field of the final output F_4^o is 4 times larger than that of F_1. This enables the network to reason details and global tendency in finer-grained fragments. STE simulates human motion by constructing a sequence of adjacent frame differences to enhance spatial tendency. We compared the parameter size and training time with TCMR. Parameters: 108.89M(TCMR) vs. 127.19M(Ours); training time: about 1.5h (TCMR)vs. 1.8h(Ours).

Weaknesses 3: Detailed ablation is not given for the effect of the different integration strategies.

A3: Thanks for your review. In the integration strategy, we also did experiments using only the self-integration phase or only the cross-integration with a PA-MPJPE of 62.7(self) vs. 62.3(cross) vs. 61.6(ours). The results show that self-integration and cross-integration phases work best when combined.

Weaknesses 4: Qualitative video results were not provided in the supplementary to evaluate effectiveness of temporal coherence of fitted SMPL models.

A4: Thanks for your review. We have provided qualitative video results of our method on Github as a supplement. At the same time, we will also present the experimental results and details mentioned above in the form of additional materials.

Reviewer #3

Weaknesses 1: The paper proposes a complicated framework for the temporal-spatial tendency reasoning. However, the intuition behind the design is not very clear. For example, the intuition of the spatial representation offset M1, similarly for M2, the motivation of transforming the feature representation to the Fourier domain and then transform back, and also the detailed description of the 'Integration' operation.

A1: Thanks for your review. Motion information is an important clue for understanding human behavior in videos. Our designed STE module exploits the temporal differences of adjacent frame-level features to focus on motion features while suppressing irrelevant information in the background. We observe that the overall appearance information of the image changes slowly over time. The pixel value of the human motion area changes more than the static area block. So we use M_1, M_2 to approximate human motion to enhance the spatial tendency. For M_2, since the Fourier transform has integrity, we use the overall difference method. When designing the module, we also explored operations such as Gaussian filtering in the Fourier domain, and we found that the effect was not ideal. Human motion can be considered as high-frequency information in the frequency domain. We consider that the FFT-IFFT operation can capture the high-frequency motion information of the human body to compensate for the lack of motion features in the time domain. Secondly, this operation can restore the spatial features as much as possible while enhancing the high-frequency motion features. Integrated detailed operations: We pass each feature to the ReLU activation function and a fully connected layer to change the size of the channel dimension to 2048. The output features are then adjusted to 256 and concatenated by a shared fully connected layer. The concatenated features are passed to several fully connected layers, followed by a softmax activation function, which produces the corresponding attention values a_1, a_2, a_3. The final feature $F_{STR} = a_1 F_{ttr} + a_2 STEF_1^o + a_3 STEF_2^o$.

Weaknesses 2: The design of the STE shares some spirit with the SENet, while the SENet is much simpler. I wonder how much the performance will drop if the STE is replaced with a simpler version like SENet. Similarly, how much the performance will drop if we replace the integration strategies with a simpler version of feature aggregation.

A2: Thanks for your review. SENet is designed for image classification tasks. The differences between STE and SENet: (1) When SENet is applied to spatiotemporal features, each frame of the video is processed independently without considering temporal information. (2) SENet uses its own global features to calibrate different channels, while our STE is a spatial tendency enhancement module that captures motion-related features. We investigate experiments where SENet replaces STE. PA-MPJPE: 62.5 (SENet) vs.

61.6 (STE). Experiments show that compared with SENet, STE can better capture human motion features to enhance spatial tendency. We also investigate experiments on alternative ensemble strategies for common feature aggregation. PA-MPJPE: 62.6 (plain aggregate) vs. 61.6 (ours). The integration strategy considers self-fusion on the basis of ordinary aggregation. The aim is to introduce an integration of the temporal dimension, thereby enhancing human features that weaken over time.

Reviewer #4

Weaknesses 1: Comparison against [A] needs to be added, which achieves better results than this submission in some of the metrics. Also, although the authors claim L304-306 that ”the performance on the challenging dataset being particularly impressive”, the improvement on that dataset does not seem too large compared to TCMR.

A1: Thanks for your review. We will add a quantitative comparison and analysis with the method of Lee et al. in the paper. However, the method of Lee et al. does not disclose the code, so we cannot obtain qualitative results of the method. Secondly, our method targets scenes such as extreme lighting, which is different from the method of Lee et al., so we cannot evaluate qualitative comparison results in extreme lighting scenes. It is worth mentioning that our method achieves good results in the crowded dataset 3DPW-Crowd. This also illustrates the robustness of our method, MPJPE:89.9(ours)vs.97.3(tcmr) MPVPE: 113.5(ours)vs.125.1(tcmr).

Weaknesses 2: The method seems to be quite complex, so it is unclear if it adds much computational cost compared to e.g., TCMR. Furthermore, according to the ablation study, the main gain comes from the integration strategies, whilst some of the modules add a marginal gain in exchange for increased complexity. Some runtime analysis would have been helpful too.

A2: Thanks for your review. We compared the parameter size and training time with TCMR. Parameters: 108.89M(TCMR) vs. 127.19M(Ours); training time: about 1.5h(TCMR) vs. 1.8h(Ours). The parameter increase of our model relative to TCMR and VIBE is reasonable.

Weaknesses 3: It is hard to understand the design choices behind the Spatial Tendency Enhancing, as the authors mostly describe the architecture instead of explaining the motivation for those steps. L111-114 in Section 1 tries to explain the intuition behind those elements, but an expanded explanation for the motivation of the design choices when explaining the approach (i.e., Section 2.3) would make it quite easier to understand.

A3: Thanks for your review. Sorry, we reduced the motivation to describe the architecture due to paper space limitations. For motivation, see Reviewer 3: Q1. We will add motivation to subsequent versions of the paper.

Weaknesses 4: I would have liked to see some limitations or failure cases shown. Also, I think a separate related work from the introduction would have given a better overview of past methods.

A4: Thanks for your review. Thanks for your suggestion, we’ll add limitations or failure cases, and separate descriptions of related work.