3 차원 인체 모델을 사용한 휴먼 모션 전이 기법

Herman, 박인규
인하대학교 정보통신공학과
herman.wsyi@gmail.com, pik@inha.ac.kr

Motion Transfer Using 3D Human Parameters

Herman, In Kyu Park
Department of Information and Communication Engineering, Inha University

요 약

모션 전이 기법은 주어진 모션 시퀀스를 타겟 대상의 움직임에 적용하는 기법이다. 사실적인 모션 전이를 위해서는 소스와 타겟 휴먼의 포즈, 형태 및 카메라 정보를 기반으로 한 모션 정보가 필요하다. 본 논문은 최근 3 차원 인체 모델링에서 우수한 성능을 보인 SMPL 을 이용하여 정교한 모션 정보를 추출하고 이를 통한 모션 전이를 수행 할 수 있는 기법을 보인다. 소스와 타겟의 SMPL 매개 변수를 사용하여 모션 정보를 나타내고 이를 통해 encoder 로부터 추출된 특징 맵을 변형하여 모션 전이를 수행한다. 제안하는 기법의 정성, 정량적 분석을 보이고 휴먼 모션 전이 기법에 대한 향후 연구 방향을 제시한다.

1. Introduction

The task of human motion transfer is to impose a source motion sequence to the target appearance. The motive is creatively driven, where the objective is to generate a dancing video of a target person corresponding to the movement of the source video, which allows any person to dance as long as we have the motion information, such as a dance video tutorial. That is, given single or multiple target images, an output sequence is produced where the target appearance imitates the motion of the source sequence.

To do this, we need a human body representation that models a clear relationship between joints, where shape and pose information should be disentangled. This brings us to SMPL, first introduced by Bogo et. al. [1], a parametric model commonly used in 3D human reconstruction tasks. It consists of three groups of parameters; joint rotation, shape information, and camera parameters. These parameters are used by the SMPL layer, where it will output a 3D body mesh with 6890 vertices that correspond to the parameters. The parametric approach allows for a low-dimensional representation of a 3D human body.

In this work, we want to explore the feasibility of using SMPL parameters for motion transfer. The motivation being the excellent availability, adoption, and research progress of SMPL-based methods.

The main challenge of this approach is to consolidate motion information within the network. Simply creating a correspondence between the parameters and the target appearance would require re-training every time the target appearance changes. The motivation is to infer motion information from SMPL parameters that modulate the features, disentangling the neural network's task of motion transfer from appearance reconstruction.

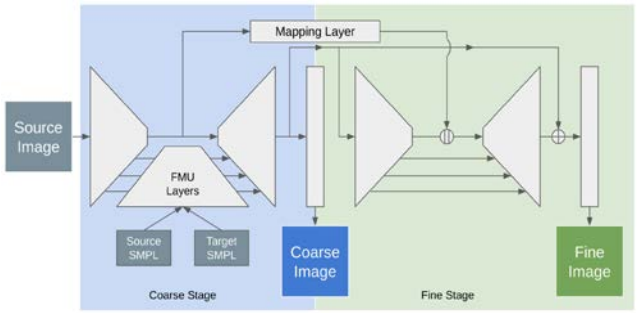

Figure 1: Diagram of the Network. The source image is fed into the coarse network to provide appearance information, while the Feature Modulator Unit layers (FMU) take the source and target SMPL parameter to compute motion information.

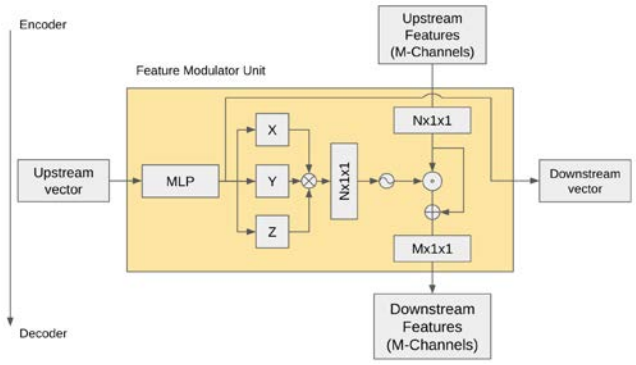

Figure 2: Diagram of Feature Modulator Unit (FMU). The FMU computes the motion information tensor M using the upstream vector V through MLPs, which is then also passed to the deeper layers of FMU. We then apply the Hadamard operation between M and the upstream features F_u to produce downstream features F_d.

2. Proposed Method

We employ a U-Net-based architecture [2] as our backbone due to its skip connections carrying both low and high-level information from the encoder to the decoder. Instead of using the skip connections directly, we want to incorporate the SMPL parameters as motion information. The consolidated motion information is applied to the skip connection features. Therefore, the objective of the decoder is to construct an image based on the input image and the transformed features.

We split the task into two distinct parts as illustrated in Figure (1). The first part is the coarse stage, which deals with feature transformation and creating the overall human body shape. The second part is the fine stage, which improves upon the coarse results to refine the shape and add more detail to the image.

The Feature Modulator Unit (FMU) is the heart of this experiment, as its primary purpose is to consolidate motion information from two sets of SMPL parameters that represent the source and the target parameters. The process consists of five elements. The first two elements are the upstream and downstream motion vectors to allow gradients to flow.

The third element is the process of representing motion information M within the unit. As seen in Figure (2), we compute three vectors to create the motion tensor M with the same size as the feature tensor.

We use three different layers to produce a linearly separable ZYX tensor. The tensor is then passed into a convolutional layer before applying the Tanh activation function to produce M.

Finally, the final two elements are the upstream and downstream image features. For each FMU in layer i, M will perturb the upstream features F_u to create the downstream features F_d using the following formula:

$$F_d^i = \left(M^i \odot F_u^i\right) + F_u^i \quad (1)$$

where $\odot$ is the Hadamard product between two matrices. Since we deploy the FMU at each level, both global and local features are capable of being perturbed by the network.

We employ several standard image reconstruction losses for our training objective, which are L_1, L_{DSSIM}, and L_{LPIPS}. We also define an additional term L_{LOC} as a local image loss based on location (x, y) and parameters (κ_x, κ_y) that controls the cropping size. We compute the local losses on the face, body, and hand regions. The final objective is a weighted sum of all the loss terms:

$$L = \lambda_1 L_1 + \lambda_2 L_{LPIPS} + \lambda_3 L_{DSSIM} + \lambda_4 L_{LOC} \quad (2)$$

3. Implementation

We use a subset of the iPER dataset by Liu et. al. [3] for training and evaluation. We use subject number 9 for training and doing the self-reconstruction evaluation, while we use subject number 8 for doing cross-appearance reconstruction evaluation. We have two training scenarios for the model, one conditioned on one appearance and one conditioned on multiple appearances. We use only appearance number 8 for the former scenario, while we use all the available appearances for the latter scenario.

4. Experimental Results

The self-reconstruction result works well for simple poses, especially with the model conditioned on the single appearance. When we train the model on a multi-appearance dataset, the performance suffers, particularly due to the severe loss of detail and texture. However, the shape is still well preserved, as reflected in the SSIM score.

Table 1: Quantitative results. 1 is the T-pose motion while 2 is random motion. Both are from Subject 9 with appearance 8. 3 is random motion from Subject 8.

Training Method	SSIM	NRMSE	PSNR
Single Appearance[1]	29.49	0.046	0.97
Single Appearance[2]	22.16	0.105	0.91
Single Appearance[3]	17.15	0.184	0.86
Multiple Appearance[1]	24.30	0.080	0.93
Multiple Appearance[2]	21.68	0.108	0.91
Multiple Appearance[3]	21.25	0.114	0.91

When we use a different subject for motion transfer, the single-appearance performance suffers greatly as it is simply memorizing the appearance of the training subject, while the multiple-appearance model performed more consistently. Based on qualitative results, it is apparent that the model is still conditioned on appearance when trained only on a single appearance. The multiple-appearance one performs much better in terms of generalization. However, it suffers from a lack of detail.

From the results, it is feasible to do subject-agnostic motion transfer. However, this approach most likely failed due to the inability to preserve detailed features after perturbation, thus causing the loss of detail. For future works, we can adopt a sequence-based model to allow for the hidden states to act as memory, thus allowing the network to adapt better to unseen data.

5. Conclusion

We propose to use SMPL parameters to modulate image features to achieve subject-agnostic motion transfer. While our method manages to do subject-agnostic shape reconstruction when trained on multiple appearances, it fails to capture fine detail. This is because of the inability to preserve features after perturbation. The encoder-decoder structure is also prone to overfitting if trained for too long. Future works include adding a memory mechanism to the network using a sequence-based architecture to alleviate this problem.

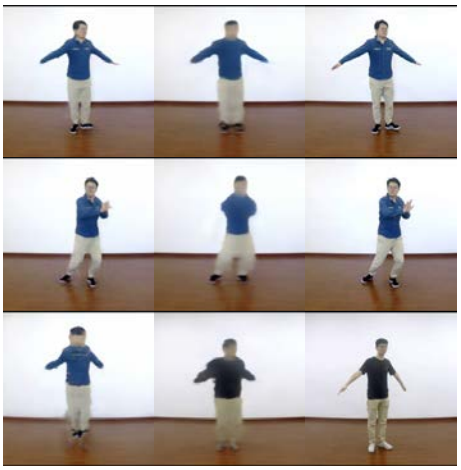

Figure 3: Qualitative Results. Left to right: single appearance, multiple appearance, ground truth. Top to bottom: Self T-pose, self random motion, cross random motion. Note how the network trained on single appearance memorizes Subject 9's appearance

Acknowledgement

This work was supported by Samsung Research Funding Center of Samsung Electronics under Project Number SRFCIT1901-06. This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (2020-0-01389, Artificial Intelligence Convergence Research Center (Inha University)).

References

[1] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. Proc. European Conference on Computer Vision, 2016.

[2] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. Proc. MICCAI, 2015

[3] W. Liu, Z. Piao, J. Min, W. Luo, L. Ma, and S. Gao, "Liquid warping GAN: A unified framework for human motion imitation, appearance transfer and novel view synthesis," Proc. IEEE International Conference on Computer Vision, 2019.