Master 2 Internship subject:

Monocular Reconstruction of 4D Human Models from Video

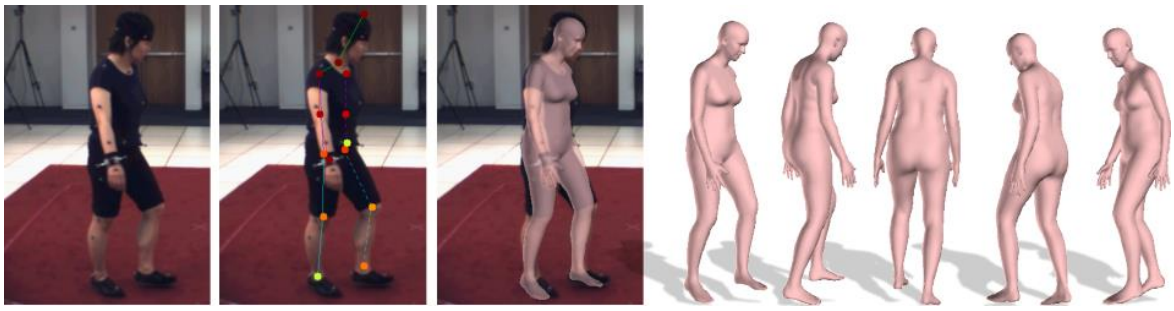

Figure 1: A model-based method can efficiently constrain the solution space for the monocular reconstruction of human body [9].

Hosting institute

ICube Laboratory (Le laboratoire des sciences de l'ingénieur, de l'informatique et de l'imagerie : The Engineering science, computer science and imaging laboratory) at the University of Strasbourg is a leading research center in Computer Science, with more than 300 permanent researchers, with the recently opened AI graduate school supported by the French government.

Work place and salary

The internship will take place in MLMS (Machine Learning, Modélisation & Simulation) research team located at the hospital site of the laboratory, 10 min walking distance to the city center of Strasbourg, which is a UNESCO world heritage site.

Salary: 500€/month approximately for a duration of 6 months.

Supervisors

— director: Hyewon Seo (ICube, Univ. Strasbourg)

— co-supervisor: Cédric Bobenrieth (ICube, ECAM)

Staring date

February – March 2021.

Context

The robust three-dimensional reconstruction of face and body from one or more images has been an open problem for decades, with many exciting application areas. Initially, efforts were focused on facial reconstruction [1][2][3], and later evolved into the reconstruction of body [4-11]. A common way to capture such models is to use calibrated multi-view passive cameras [4-8] to merge a sparse or dense set reconstructed depth images into a single mesh, but size and cost of such multi-view systems prevent their use in consumer applications.

In more unconstrained and ambiguous settings, such as in the monocular image or video, priors in the form of template model or parametric model are often used, which help to constrain the problem significantly. While *generative* methods reconstruct the moving geometry by optimizing the alignment between the projected model and the image data [8][9], *regressive* methods [3][11][12][16] train deep neural networks to infer shape parameters of a parametric body model from a single image. Despite remarkable progress, reconstruction of 4D humans, i.e. space-time coherent 3D models has not been fully addressed yet, with most existing algorithms operating in a frame-by-frame manner.

In this internship, we will focus on the reconstruction of space-time coherent deforming geometry of entire human body from video input. The problem is particularly challenging since such 4D data is typically of high dimension both spatially and temporally. We will approach the problem by combining a parametric model such as SMPL [13] with recent deep learning techniques that learn to predict both the shape and the motion of the human body in its parametric space.

Objectives

Our work will be inspired by recent progress on deep autoencoders that approximate an identity mapping by coupling an encoding stage with a decoding stage to learn a compact latent representation of reduced dimensionality. With its appealing characteristic that these are unsupervised, i.e. no labeled data is required, autoencoders have been used to tackle a wide range of tasks, including face recognition [14], real-time 2D-to-3D alignment [15], and face model reconstruction [16].

The main objective is to develop a novel, model-based autoencoder that will learn to jointly regress a set of model parameters (identity shape, pose-dependent shape) based on a skinned template, as well as camera parameters to the foreground segmented from the input video. Among others, SMPL [13] representation is considered as our model: the body model is parameterized by the pose vector $\boldsymbol{\theta}$ and shape vector $\boldsymbol{\beta}$, with a template mesh $\boldsymbol{M}$ whose pose-dependent deformation is computed using a linear blend skinning function. It will further include camera parameters, the orientation $\mathbf{T} \in SO(3)$ and the position $\mathbf{t} \in R^3$ of the camera. More specifically, the 3D body $M(\boldsymbol{\beta}, \boldsymbol{\theta})$ will be rendered using a full perspective projection $\prod: R^3 \rightarrow R^2$ that maps from camera space to screen space. To enable training, implementing a backward pass may be required, i.e. the computation of the gradients of the projection function with respect to the parameters. Developing a robust loss function that includes spatiotemporal regularization along with the data error will also be an important part of this work. Evaluation and comparison of the performance to the state-of-the-art methods is strongly recommended, whenever applicable.

Candidate profile

— Master student in Computer Science or in (Applied) Mathematics

— Solid programming skills in deep learning platforms: Tensorflow/Pytorch

— Background in geometric modeling and statistics

— Good communication skills

Application

Send your CV and your academic transcripts (Bachelor and Master courses) to seo@unistra.fr.

References

[1] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In Proc. SIGGRAPH, pages 187–194. ACM Press/Addison-Wesley Publishing Co., 1999.

[2] Tran, Luan & Liu, Xiaoming. (2018). Nonlinear 3D Face Morphable Model. 7346-7355. 10.1109/CVPR.2018.00767.

[3] Deng, Yu & Yang, Jiaolong & Xu, Sicheng & Chen, Dong & Jia, Yunde & Tong, Xin. (2019). Accurate 3D Face Reconstruction with Weakly-Supervised Learning: From Single Image to Image Set. WORKSHOPS (CVPRW) (2019): 285-295.

[4] J. Carranza, C. Theobalt, M. A. Magnor, and H.-P. Seidel. Free-viewpoint video of human actors. In ACM Transactions on Graphics, volume 22, pages 569–577. ACM, 2003.

[5] J. Starck and A. Hilton. Surface capture for performance-based animation. IEEE Computer Graphics and Applications, 27(3), 2007.

[6] E. De Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H.-P. Seidel, and S. Thrun. Performance capture from sparse multiview video. In ACM Transactions on Graphics, volume 27, page 98. ACM, 2008.

[7] N. Hasler, H. Ackermann, B. Rosenhahn, T. Thormahlen, and H.-P. Seidel. Multilinear pose and body shape estismation of dressed subjects from image sets. In IEEE Conf. on Computer Vision and Pattern Recognition, pages 1823–1830. IEEE, 2010.

[8] B. Allain, J.-S. Franco, and E. Boyer. An Efficient Volumetric Framework for Shape Tracking. In IEEE Conf. on Computer Vision and Pattern Recognition, pages 268–276, Boston, United States, 2015.

[9] 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. In European Conf. on Computer Vision. Springer International Publishing, 2016.

[10] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. (2018). Video Based Reconstruction of 3D People Models. CVPR 2018 Spotlight, IEEE Conference on Computer Vision and Pattern Recognition 2018 (CVPR).

[11] Dibra, Endri & Jain, Himanshu & Oztireli, Cengiz & Ziegler, Remo & Gross, Markus. (2017). Human Shape from Silhouettes Using Generative HKS Descriptors and Cross-Modal Neural Networks. 5504-5514. 10.1109/CVPR.2017.584.

[12] Richardson, Elad & Sela, Matan & Or-El, Roy & Kimmel, Ron. (2017). Learning Detailed Face Reconstruction from a Single Image. 5553-5562. 10.1109/CVPR.2017.589.

[13] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: a skinned multi-person linear model. ACM Trans. Graph. 34, 6, Article 248 (November 2015), 16 pages.

[14] M. Kan, S. Shan, H. Chang, and X. Chen. Stacked Progressive Auto-Encoders (SPAE) for Face Recognition Across Poses. 2014.

[15] J. Zhang, S. Shan, M. Kan, and X. Chen. Coarse-to-Fine Auto-Encoder Networks (CFAN) for Real-Time Face Alignment. 2014.

[16] Tewari, Ayush & Zollhöfer, Michael & Kim, Hyeongwoo & Garrido, Pablo & Bernard, Florian & Pérez, Patrick & Theobalt, Christian. (2017). MoFA: Model-based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction.