THE ROMANIAN ACADEMY

School of Advanced Studies of the Romanian Academy

"Simion Stoilow" Institute of Mathematics

PH.D. THESIS SUMMARY

Human Sensing in Videos

Supervisor:
C.S. I Dr. Cristian Sminchișescu

Author:
Mihai Fieraru

Bucharest, 2024

Thesis Contents

Chapter 1

Introduction

Building intelligent systems that understand the world and are useful to humans requires developing algorithms that can perceive people from visual data. In this thesis, we address the three-dimensional reconstruction of humans from a monocular image or video. This is an important problem, with the potential of enabling numerous applications, such as novel human-computer interfaces, robot navigation and interactions, security and surveillance, healthcare or entertainment. Yet, sensing 3d humans solely from monocular cameras poses several challenges: depth ambiguity, partial views, occlusions with the scene or self-occlusions, large appearance variation due to diverse human bodies, articulations, garments and various types of physical contact.

Our first contributions focus on understanding 3d human interactions. We introduce models for 3d contact signature prediction and show how their use in an optimization setting improves reconstruction of people in close proximity. We also build two large datasets for training and evaluation purposes and propose methodology for recovering the ground-truth pose and shape of interacting people in a controlled setup.

Next, we study the reconstruction of human self-contact. We propose models for predicting self-contact signatures from monocular images and show how they can be leveraged to improve human reconstruction accuracy in an optimization framework. To support learning and evaluation, we collect two large datasets, one captured in the laboratory and one consisting of in-the-wild images with 3d self-contact annotations.

We then propose a method for end-to-end learning of 3d reconstruction of multiple interacting humans under weak supervision. We introduce a novel unified model for self-collision and interpenetration losses and use both self-contact and interaction contact losses directly into the learning process. Our model obtains state-of-the-art results even when no 3d supervision is used.

Lastly, we present the first automatic system performing 3d human sensing for fitness training, made possible by the collection of a motion capture dataset of more than 37 types of exercises. The system estimates 3d human motion, segments exercise

repetitions and identifies the deviations between standards learnt from trainers and the execution of a trainee, offering quantitative feedback in natural language.

To support research in this field, we release all datasets collected in this thesis, together with evaluation servers and public benchmarks.

Chapter 2

Reconstructing Three-Dimensional Models of Interacting Humans

This chapter is based on the paper [1] "Three-Dimensional Reconstruction of Human Interactions." by Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu, published in The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020. which was further augmented as the current version and is currently under submission as a journal (also available as arXiv preprint [2]).

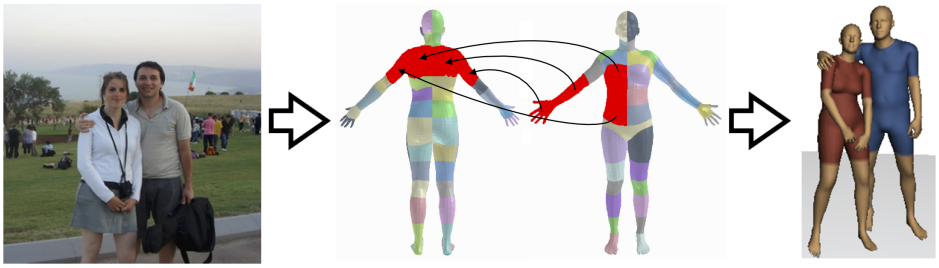

Figure 2.1: Monocular 3d reconstruction, constrained by contact signatures, preserves the essence of the physical interaction between people and supports behavioral reasoning.

In this paper, we propose a first set of methodological elements to address the reconstruction of interacting humans by relying on recognition, segmentation, mapping, and 3d reconstruction. More precisely, we break down the problem of producing veridical 3d reconstructions of interacting humans into (a) contact detection, (b) binary segmentation of contact regions on the corresponding surfaces associated to the interacting people; (c) contact signature prediction to produce estimates of the potential many-to-many correspondence map between regions in contact; and (d) 3d reconstruction under augmented losses built using additional surface contact constraints given a

contact signature.

Figure 2.2: Annotated text describing the motion and fitting of the GHUM model to two interacting subjects in the CHI3D dataset. All 4 stacked views are displayed for 3 frames (left to right in temporal order): a hugging sequence (top), a posing sequence (bottom). Subject in green is the subject wearing the markers. Subject in brick-red does not wear any markers on clothing.

To train models and evaluate the techniques, we introduce two large datasets. We capture CHI3D, a lab-based 3d motion capture repository containing 631 sequences containing $2,525$ contact events, $728,664$ ground truth skeletons, as well as annotated temporal extent of the physical contact and a textual description of the interaction motion in each video sequence. We also gather FlickrCI3D, a dataset of $11,216$ images, with $14,081$ processed pairs of people, and $81,233$ facet-level surface contact correspondences.

Besides, we propose methodology to obtain the ground truth pose and shape of the interacting people in the CHI3D dataset. By leveraging information from motion sensors, multi-view RGB cameras and a 3d scanner, but also from contact annotations,

pose priors and physical constraints, we achieve ground-truth level 3d reconstructions (see fig. 2.2).

We publicly release the ground-truth motion sequences in multiple formats (GHUM [3] and SMPLX [4] parameters, Human3.6m [5] 3d joints) at `ci3d.imar.ro` and implement an evaluation server on a hidden test set, together with a public benchmark, with the purpose of advancing the state of the art in 3D human reconstruction of close contact interactions.

We evaluate the performance of the contact detection task and obtain an average accuracy of 0.846, with 0.844 for the "contact" class and 0.848 for the "no contact" class. Table 2.1 shows the evaluation of our model *ISP* for contact segmentation and signature estimation using the intersection over union ($\text{IoU}_{N_{reg}}$) metric, computed for different region granularities. Table 2.2 shows that annotated contact information improves the accuracy of the reconstruction.

Method	$\text{IoU}_{75}\uparrow$		$\text{IoU}_{37}\uparrow$		$\text{IoU}_{17}\uparrow$		$\text{IoU}_{9}\uparrow$	
	Segm.	Sign.	Segm.	Sign.	Segm.	Sign.	Segm.	Sign.
ISP full	**0.318**	**0.082**	**0.365**	**0.129**	**0.475**	**0.248**	**0.618**	0.408
ISP w/o semantic 2d features as input	0.300	0.073	0.350	0.116	0.465	0.240	**0.618**	**0.410**
ISP w/o jointly learning contact segm.	-	0.072	-	0.124	-	0.218	-	0.383
ISP w/o masking out corresp. outside the estimated segm. mask	-	0.075	-	0.124	-	0.230	-	0.385
Human performance	0.456	0.226	0.542	0.370	0.638	0.499	0.745	0.635

Table 2.1: Results of our *ISP* model for contact segmentation and signature estimation, evaluated on FlickrCI3D for different region granularities on the human 3d surface (from 75, down to 9 regions). We ablate different components of our full method to illustrate their contribution. Human performance represents the consistency values between annotators.

	Grab	Hit	Handshake	Holding hands	Hug	Kick	Posing	Push	OVERALL
Optim.	P↓ T↓	P↓ T↓	P↓ T↓	P↓ T↓	P↓ T↓	P↓ T↓	P↓ T↓	P↓ T↓	P↓ T↓
Loss	C↓	C↓	C↓	C↓	C↓	C↓	C↓	C↓	C↓
$L^\star$	**117 390**	**119 367**	**97 388**	101 **380**	**174 400**	**140 419**	**139 364**	**117 381**	**125 368**
	19 (**3.5**)	**8** (**4**)	**12** (**3**)	**20** (**3**)	**62** (**45**)	**32** (**7**)	**41** (**11**)	**14** (**4**)	**26** (**10**)
$L^\star$ **w/o** L_G	121 416	128 396	99 406	**100** 389	180 424	155 460	140 377	124 399	131 408
[6]	459 (366)	426 (363)	377 (305)	373 (274)	368 (328)	550 (464)	388 (327)	425 (369)	421 (350)

Table 2.2: Comparison between using the contact consistency loss during optimization ($L^\star$) and not using it ($L^\star$ **w/o** L_G). 3d human **pose** (**P**) and **translation** (**T**) estimation errors, as well as mean (median) 3D **contact distance** (**C**), expressed in mm, for the CHI3D dataset. Our full optimization function, with the geometric alignment term on contact signatures, decreases the pose and translation estimation errors and the 3D distance between the surfaces annotated to be in contact.

Chapter 3

Learning Complex 3D Human Self-Contact

This chapter is based on the paper [7] "Learning Complex 3D Human Self-Contact." by Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu published in the Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, pages 1343–1351, 2021.

Most monocular 3d human reconstruction systems do not directly infer human self-contact, although its central role in correctly recognizing the subtleties of many iconic poses or gestures is widely acknowledged perceptually.

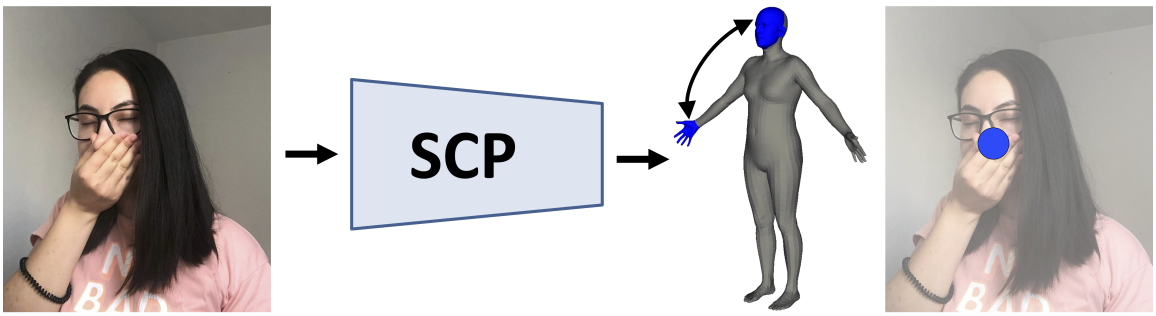

Figure 3.1: Our self-contact prediction network (*SCP*) estimates the body regions in contact, their correspondences and the self-contact positioning in image space.

To overcome some of the shortcomings of existing, self-contact agnostic, 3d reconstruction methods, we propose to represent self-contact explicitly and show how the resulting models can assist behavioural understanding in applications assessing face touching. Our models learn to predict the image location of contact in order to assist the detection of body regions in self-contact, as well as their signature, defined as the correspondences between regions on the surface of a human body model that touch. Conditioned on such detailed estimates, self-contact can be recovered correctly in the

3d reconstruction. To train models and for large-scale quantitative evaluation, we collect and annotate two large scale datasets containing images of people in self-contact. HumanSC3D is an accurate 3d motion capture dataset containing $1,032$ sequences with $5,058$ contact events and $1,246,487$ ground truth 3d poses synchronized with images captured from multiple views. We also collect FlickrSC3D, a dataset of $3,969$ images, containing $25,297$ annotations of body part region pairs in contact, defined on a 3d human surface model, together with their self-contact localisation in the image.

The main contributions of the paper are as follows:

- Introduce a first principled model to detect self-contact body regions and their signature. Our novel deep neural network *SCP* is assisted by an intermediate self-contact image localisation (branch) predictor, leveraged both in training, for local feature selection, and in testing, by enforcing consistency with the estimated 3d contact signature. Evaluation results are shown in Table 3.1.

- Novel, task-specific, large scale, valuable community datasets capturing people in self-contact, together with dense annotations of a 3d body model to capture the surface regions in contact, as well as image annotations associated to the observed points of contact. The data is made available for research purposes.

- Quantitative (see Table 3.2) and qualitative (see Fig. 3.2) demonstration of metrically more accurate and perceptually veridical 3d reconstructions based on self-contact signatures.

- A foundation for a large class of applications that would benefit from accurate 3d self-contact representations, such as, health monitoring of possible infections when hands touch parts of the face (mouth, nose, eyes) in hospitals or during a pandemic, or subtle behavioral understanding of gestures for robot-assisted therapy of children with autism, to name just a few.

	$\text{IoU}_{\mathbf{75}}\uparrow$		$\text{IoU}_{\mathbf{37}}\uparrow$		$\text{IoU}_{\mathbf{17}}\uparrow$		$\text{IoU}_{\mathbf{9}}\uparrow$	
Method	Segm.	Sign.	Segm.	Sign.	Segm.	Sign.	Segm.	Sign.
SCP	**0.469**	**0.301**	0.507	**0.339**	0.591	**0.442**	0.693	**0.550**
ISP [1] (adapted for self-contact)	0.462	0.133	0.503	0.186	0.583	0.305	0.688	0.460
Human performance	0.528	0.422	0.564	0.475	0.664	0.579	0.768	0.692

Table 3.1: Results of our self-contact segmentation and signature estimation *SCP* on FlickrSC3D, evaluated for different region granularities on the human 3d surface (from 75, down to 9 regions). Human performance represents the consistency values between annotators.

Optim. loss	W/o chair - standing				W/o chair - sitting				W/ chair				Overall			
	P↓	T↓	V↓	C↓	P↓	T↓	V↓	C↓	P↓	T↓	V↓	C↓	P↓	T↓	V↓	C↓
L	**94**	**408**	**77**	**13**	**116**	**424**	**93**	**27**	**107**	**426**	**85**	**24**	**98**	**414**	**80**	**16**
L w/o L_G	106	419	121	210	145	436	147	183	132	432	123	189	114	423	124	203

Table 3.2: 3D human pose (P), translation (T), vertex (V) estimation errors, as well as mean 3d contact distance (C), expressed in mm, for the HumanSC3D dataset. Using the full optimization function, with the geometric alignment term on annotated self-contact signatures, decreases the pose, translation and vertex estimation errors as well as the 3d distance between surfaces annotated as being in contact.

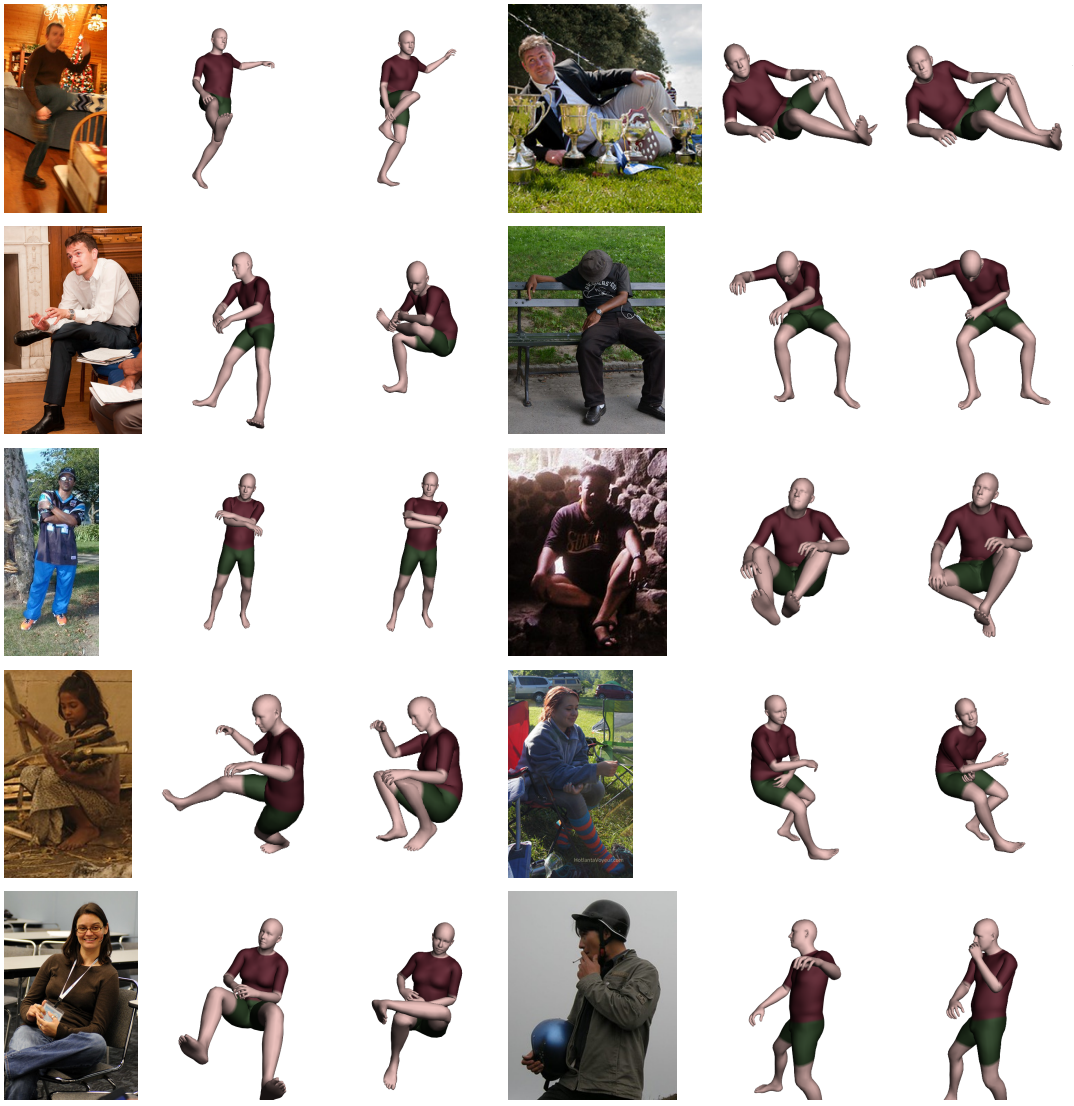

Figure 3.2: 3D pose and shape reconstructions using our annotated self-contact data. Original image (left). Reconstruction without considering the self-contact and the associated loss (center). Reconstruction that uses the self-contact annotations and the corresponding loss (right).

Chapter 4

REMIPS: Physically Consistent 3D Reconstruction of Multiple Interacting People under Weak Supervision

This chapter is based on the paper [8] "REMIPS: Physically Consistent 3D Reconstruction of Multiple Interacting People under Weak Supervision." by Mihai Fieraru, Mihai Zanfir, Teodor Szente, Eduard Bazavan, Vlad Olaru, and Cristian Sminchisescu, published in Advances in Neural Information Processing Systems, volume 34, pages 19385–19397. Curran Associates, Inc., 2021.

The three-dimensional reconstruction of multiple interacting humans given a monocular image is crucial for the general task of scene understanding, as capturing the subtleties of interaction is often the very reason for taking a picture. Current 3D human reconstruction methods either treat each person independently, ignoring most of the context, or reconstruct people jointly, but cannot recover interactions correctly when people are in close proximity. In this work, we introduce **REMIPS**, a model for 3D Reconstruction of Multiple Interacting People under Weak Supervision. **REMIPS** can reconstruct a variable number of people directly from monocular images. At the core of our methodology stands a novel transformer network that combines unordered person tokens (one for each detected human) with positional-encoded tokens from image features patches (see Fig. 4.1). We introduce a novel unified model for self- and interpenetration-collisions based on a mesh approximation computed by applying decimation operators. We rely on self-supervised losses for flexibility and generalisation in-the-wild and incorporate self-contact and interaction-contact losses directly into the learning process. With **REMIPS**, we report state-of-the-art quantitative results (see Table 4.1) on common benchmarks even in cases where no 3D supervision is used.

Additionally, qualitative visual results (see Fig. 4.2) show that our reconstructions are plausible in terms of pose and shape and coherent for challenging images, collected in-the-wild, where people are often interacting.

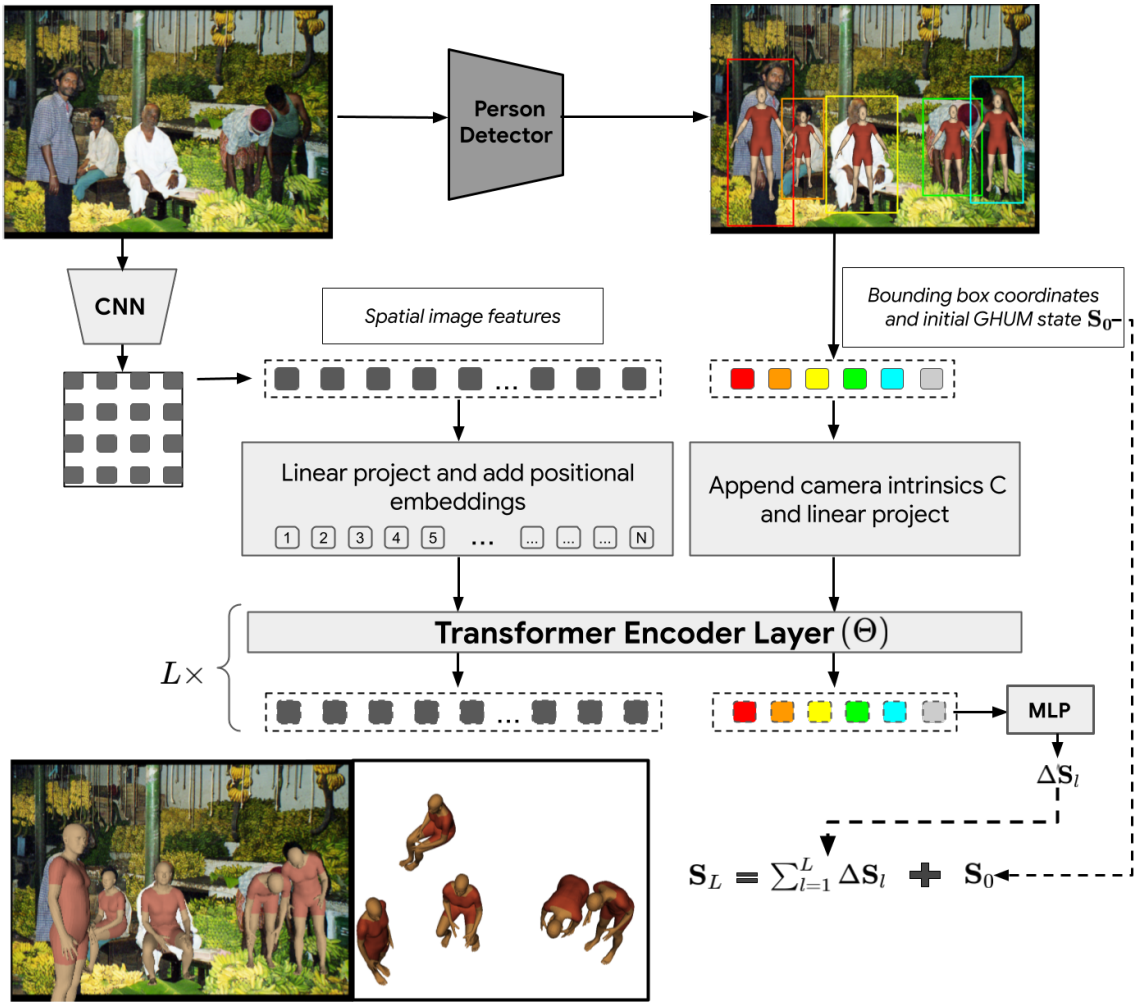

Figure 4.1: Overview of **REMIPS**, our proposed architecture to reconstruct the 3D pose and shape of multiple interacting people. Starting from a single input image, we use an off-the-shelf detector to extract the human bounding boxes. We create a sequence of person tokens from these detections to which we attach an initial GHUM state estimate $\mathbf{S}_0$. On a separate branch, starting from the image, we run a backbone convolutional neural network architecture and create an additional sequence of spatial image feature tokens, $\mathbf{F}_s$. We concatenate the two sequence representations and iteratively refine this joint representation through a single transformer encoder layer for a number of L stages. At the end of each stage l, we collect the transformed representation for the token sequence associated with the people and apply an MLP to regress the residual GHUM state estimates $\Delta\mathbf{S}_l$. Our final estimation is given by the weighted sum of all the residual state updates and the initial state. The network is trained weakly-supervised on various datasets with 2D annotations. We use contact and collision losses defined over the recovered geometries to ensure physical plausibility.

Method	MPJPE ↓	MPVPE ↓	Translation Error ↓	#2D	#3D
Fieraru *et al.* [1]	125.4	–	368.0	N/A	N/A
Jiang *et al.* [9]	136.0	N/A	N/A	100K	300K
REMIPS (ours)	**120.8**	**134.7**	**284.1**	115K	0

Table 4.1: Performance on the **CHI3D** [1] dataset for multiple person pose and shape reconstruction methods. In columns 2, 3, 4 we show the mean per joint position error (**MPJPE**), the mean per vertex position error (**MPVPE**) and the translation error. All errors are reported in mm and are relative to the root joint. Our method has lower errors compared to the other optimization and inference based methods. We also compare the number of **#2D** and **#3D** annotations used as supervision during the training. Our models use no **#3D** and achieve better performance on the challenging dataset **CHI3D** [1].

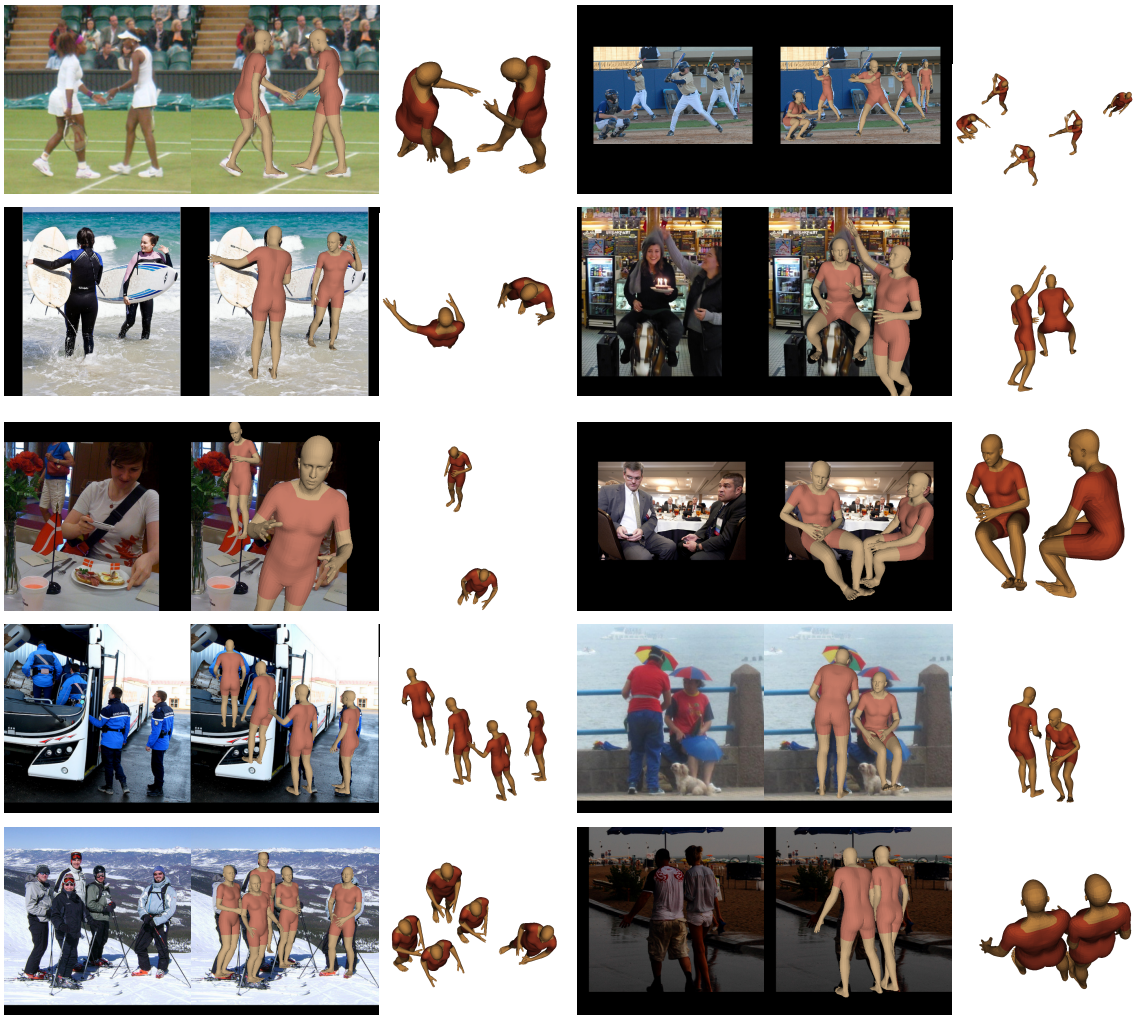

Figure 4.2: 3D human pose and shape predictions on the **COCO** validation set (rows 1-5) for in-the-wild images. We show the initial image together with an overlaid reconstruction of the meshes as well as a rendering from a different viewpoint which better illustrates the physical consistency of the **REMIPS** reconstructions.

Chapter 5

AIFit: Automatic 3D Human-Interpretable Feedback Models for Fitness Training

This chapter is based on the paper [10] "AIFit: Automatic 3D Human-Interpretable Feedback Models for Fitness Training." by Mihai Fieraru, Mihai Zanfir, Silviu-Cristian Pirlea, Vlad Olaru, and Cristian Sminchisescu, published in The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2021.

In this paper we propose *AIFit*, the first AI-enhanced training system for fitness. The system is able to reconstruct 3d human pose over time, count repetitions, and automatically provide localized feedback, visually grounded in images of the trainee, and phrased in natural language displayed on a screen (see Fig 5.1).

In order to support research and evaluation, we introduce *Fit3D*, a large-scale dataset of over 3 million images and ground truth 3d motion capture poses, collected from 13 subjects (including one licensed fitness instructor and one advanced fitness subject), observed by 4 different RGB cameras, together with 3d scans of each subject. The dataset features 37 exercises consisting of simple and compound motions, covering all major muscle groups and articulation types, including, among many others, warm-ups, barbells, dumbbells, push-ups, or yoga.

Our proposed methodology includes large-scale monocular and multi-view evaluation of 3d human pose reconstruction for fitness training using *Fit3D*, models for automatic identification of exercise repetitions, as well as methods to compare instructors' and trainees' performances according to statistical policies defined over mined features (passive and active) defining the exercise, and carrying most of its motion energy. Our statistical coach is governed by a global parameter that models how critical it is in regard to a student's performance. In practice, the parameter helps the coach adapt to a student's level of fitness (i.e. beginner vs advanced vs expert) or to the expected

accuracy of the underlying 3d pose reconstruction method. Finally and importantly, our statistical coach provides easy to understand, visually grounded spatio-temporal feedback, in natural language. A system overview is shown in fig. 5.2.

Figure 5.1: Textual and visual feedback produced by our *AIFit* on real world videos, captured with a regular smartphone camera. We use MubyNet-FT to estimate the 3d pose of the trainee. For each example, we show the following: an image with the identified error of the trainee (***top row***), the 3d reconstruction of the trainee (***second row***), the corresponding image with the correct execution of the instructor (***third row***) and the textual feedback (***bottom row***). The two examples on the (***left***) show active features feedback, while the two on the (***right***) show passive features feedback. Notice generalization to various humans in different environments and camera viewpoints.

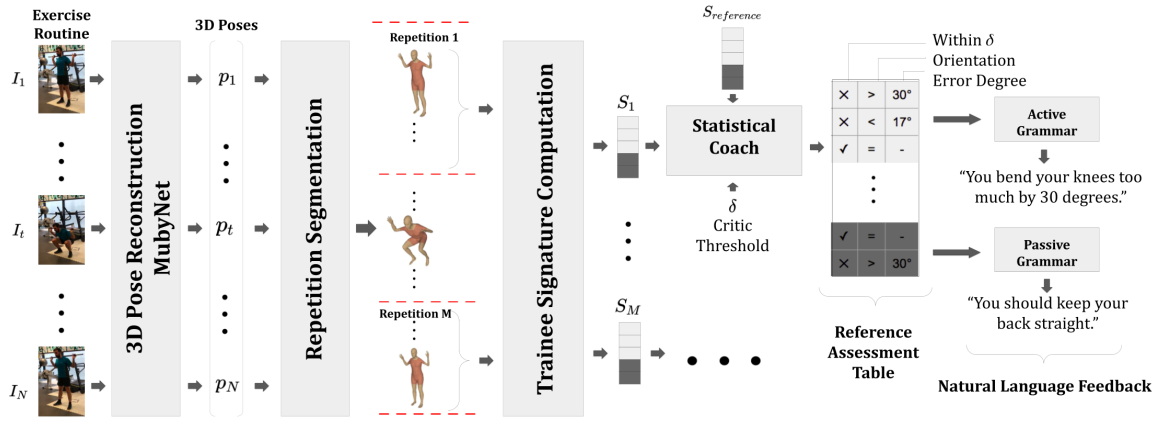

Figure 5.2: *AIFit* overview. Given a video of a trainee performing an exercise, (a) the system performs **3d pose reconstruction** in each frame and then (b) applies **repetition segmentation** to automatically count the number of 3d pose repetitions and determine each repetition interval. Next, **exercise modelling** (c) computes an *exercise signature* using the angular features of each repetition of the trainee (see fig. 5.3 for a detailed view). (d) The **statistical coach** compares each repetition signature against the instructor reference signature under a critic threshold that allows for different degree of error. The results of the comparison are populated into a **reference assessment table** specifying which deviations are greater than the critic threshold, the sign of the deviation and the degree of error. Finally, based on the table, e) *AIFit* produces **natural language feedback** for the trainee, using either an active or a passive grammar.

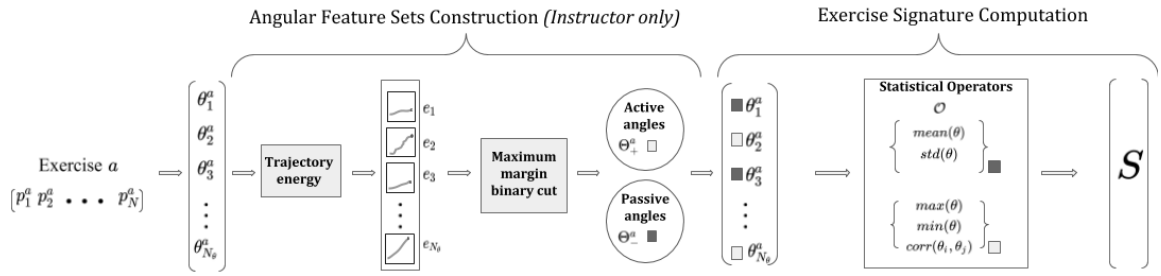

Figure 5.3: *Exercise Modelling:* **(Left) Active and passive angular feature sets construction** (instructor only). For an *exercise* a and for each angular feature function, we integrate its motion trajectory over the instructor's sequence of 3d poses, and get the motion energy of each feature function. We cluster the energies into two sets, active Θ^a_+ (associated with high energy) and passive Θ^a_- (associated with low energy) by using a maximum margin binary cut. **(Right) Exercise signature computation**. Both for trainees and instructor exercises, a signature is produced from the computed angular features, corresponding cluster assignments (derived from instructor exercises) and predefined statistical operators (applied to each of the two sets of angular features).

Bibliography

[1] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-dimensional reconstruction of human interactions. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[2] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Reconstructing three-dimensional models of interacting humans. *arXiv preprint arXiv:2308.01854*, 2023.

[3] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Ghum & ghuml: Generative 3d human shape and articulated pose models. In *CVPR*, 2020.

[4] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, 2019.

[5] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3d human sensing in natural environments. *PAMI*, 2014.

[6] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3d pose and shape estimation of multiple people in natural scenes - the importance of multiple scene constraints. In *CVPR*, 2018.

[7] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Learning complex 3d human self-contact. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1343–1351, 2021.

[8] Mihai Fieraru, Mihai Zanfir, Teodor Szente, Eduard Bazavan, Vlad Olaru, and Cristian Sminchisescu. Remips: Physically consistent 3d reconstruction of multiple interacting people under weak supervision. In M. Ranzato, A. Beygelzimer,

Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 19385–19397. Curran Associates, Inc., 2021.

[9] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5579–5588, 2020.

[10] Mihai Fieraru, Mihai Zanfir, Silviu-Cristian Pirlea, Vlad Olaru, and Cristian Sminchisescu. Aifit: Automatic 3d human-interpretable feedback models for fitness training. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021.