UNIVERSITAT POLITÈCNICA DE CATALUNYA

Doctoral Programme:

AUTOMÀTICA, ROBÒTICA I VISIÓ

Research Plan:

Context-aware human modeling

Enric Corona Puyané

Advisors:

Francesc Moreno Noguer, Dr.

Guillem Alenyà Ribas, Dr.

June 2020

Contents

1 Introduction and Motivation

Being able to understand the environment and inferring a 3D arrangement of the world is an essential ingredient for developing autonomous robots. In particular, if we want to design robots that can interact with humans, or move safely in human environments, we need to develop algorithms that can reconstruct the 3D position and shape of the visible people and are able to predict their motion in the near future. These systems need to be robust to occlusions, to the number of visible people and to the variability of the human pose. 3D pose/shape estimation and motion prediction algorithms also have a huge potential in other fields, with possible applications in virtual and augmented reality, cloth virtual try-on, games or entertainment. Moreover, if 3D understanding improves enough, robots can learn to perform new actions without explicit orders or commands, by imitating humans.

In this thesis, we want to investigate methods for human perception while taking context into account. The environment conditions human actions and motion and our actions are also designed to have an effect on the scene. In particular, we will use deep neural networks to build new algorithms that exploit geometric and semantic priors, currently not integrated within most state-of-the-art models. Semantic information can be explicitly given to the model during training, but ground truth human-object interaction data is scarce and, ideally, it should be learnt without any supervision. Moreover, this will allow to understand what contextual cues are learnt by the model to achieve a given task. For this, we will study the use of unsupervised training methods such as Variational Autoencoders or Generative Adversarial Networks. Geometric priors, in contrast, can be integrated via loss functions or other architectural designs that exploit how the 3D world is structured.

This work is developed under the National project "HuMoUR: Markerless 3D human motion understanding for adaptive robot behavior" TIN2017-90086-R. The thesis is aligned with the goal of the project, which first aims to develop computer vision algorithms for pose estimation and motion prediction. These algorithms will be used to implement new service robots that can perform complex manipulation for assisting tasks. In particular, the project plans to demonstrate our developments on the three scenarios of (a) Feeding a person, (b) brushing a persons hair and (c) help dressing a person. In all cases, the perception algorithms need to tackle strong body occlusions and be robust to pose/cloth diversity.

2 Objectives

The main goal of this doctoral thesis is to investigate contextual human perception in a variety of tasks, such as human pose/shape estimation and human motion prediction. We believe that current state-of-art methods for human perception tasks, such as pose and shape reconstruction or motion prediction, do not take contextual information into account. However, human actions are significantly conditioned by our context, mainly from relations with other people or objects we want to interact with. For this reason, we aim to incorporate geometric or semantic priors to improve on state-of-the-art methods. The specific research objectives of this thesis are:

- Designing algorithms that model human pose/shape, motion from a single input image.
- Developing context-aware architectures that can reason about human-object interactions. In this work, we plan to combine this line of work with unsupervised or semi-supervised learning methods, which few papers have proposed so far.
- Building solutions for human perception that work on in-the-wild input images, without any requirement of calibration step.

This topic presents a great research opportunity for the impact that 3D reconstruction and motion prediction can have in fields such as robotics, AR/VR, games or entertainment. We believe that we are in good position to pursue these challenges, given recent work done by the research group along this topic, and plan to perform some particular research collaborations with other experts on the field.

3 State of the Art

This section reviews the most relevant literature for the tasks we want to tackle in this thesis. The first subsection will review general deep learning research, then we will summarise the related work on human pose estimation and shape reconstruction, and finally review the most recent works on human motion estimation.

3.1 General Deep Learning Research

Over the past decade, deep learning has grown in popularity thanks to its breakthroughs in performance in a wide range of applications, specially in the fields of computer vision, natural language processing and robotics, amongst other fields where it is applied.

Although the theoretical concepts behind deep neural networks date back to the previous century [1, 2, 3], it was not until 2012 where they got popularised, when Krizhevsky *et al.* [4] won the ImageNet challenge [5] using a convolutional network architecture with critical recent improvements [6, 7]. Since then, the performance of deep learning algorithms in computer vision has continuously improved in a range of tasks.

For the current project we will also take advantage of general deep learning tools such as recurrent neural networks [8], transformers [9] and generative models, specially VAEs [10] and GANs [11]:

- Recurrent neural networks (RNNs) are a general architecture for neural networks that create a recursion between consecutive observations, using an internal representation state that contains the information about the sequence. These are commonly used in the task of processing data for multiple steps, as in the task of motion prediction.
- Recently, RNNs have been superseded in many tasks by transformers, which takes a set of pieces of information, such as words, and first encodes them to a more useful representation. Then, each representation weights the other ones in a so-called attention mechanism, that is much more interpretable and is being used in more and more deep learning areas.
- Generative models have recently allowed the generation of very realistic distribution of data, such as faces or bodies. A particular key architecture are VAEs, which are formed by an encoder and a decoder. The encoder maps the target data distribution to an encoding distribution z and the decoder does the opposite. By ensuring that z follows a Gaussian distribution during of both encoder and decoder, we can then sample from z to obtain a realistic target data distribution. These models are also used to make latent space more robust and regularise more complex processes.
- Generative Adversarial Networks (GANs) are a very successful kind of generative models, based on two networks where, a generator maps a distribution z to a

target distribution x, while a discriminator aims to identify whether x is real or was created by the generator. Both networks are jointly trained using a minimax game that allows the generator to learn a very good mapping from z to x.

3.2 Human Pose Estimation

Since the release of large-scale MoCap datasets [12, 13, 14], there has been a growing interest in the problem of estimating 3D human pose from single images [15, 12, 16, 17, 18, 19, 20, 21].

Single-person pose estimation methods follow two different lines of work. In the first one, a single-stage algorithm predicts 3D body joints position directly [22]. One of the first deep-learning-based works [23] proposed a joint model for body part detectors and pose regression. Pavlakos *et al.* [16] propose a U-Net architecture to recover joint-wise 3D heatmaps. Sun *et al.* [24] propose a regression approach using a bone-based representation that exploits human pose structure. In [25], they propose a differentiable soft-argmax operation that allows to train an hourglass network more efficiently. In the second line of work, algorithms learn a mapping from 2D estimated joint detections to 3D [26] Moreno-Noguer [26] propose to infer 3D pose via distance matrix regression. Yang *et al.* [27] propose to use an adversarial approach that ensures that estimated poses are antropomorphic.

When considering the task of 2D or 3D multi-person pose estimation, there are two opposite main approaches to structure model architectures, top-down [28, 29, 30, 31] or bottom-up models [32] . On the former, a human detector first estimates the bounding boxes of humans. Each detected human area is cropped and fed into the pose estimation network. The latest also follows a two-stage pipeline, where a model first localizes all human body keypoints in an input image first, and then groups them into each person using particular clustering techniques.

Mehta *et al.* [32] follow a bottom-up approach where they estimate three occlusion-robust location-maps [33]. They model the association between body keypoints using Part Affinity Fields [34], to allow predictions on multiple people.

However, most recent methods use a top-down architecture. For instance, Rogez *et al.* [35, 29] proposed an approach for 2D/3D Multi-Person pose estimation, where for each detected person, they classify the pose to one of the anchor clustered poses. They follow a coarse-to-fine approach by further regressing each anchor pose. More recently, Moon *et al.* [30] proposes an architecture where, on one side predicts the 3D absolute position of the root joint, and in a second branch reconstructs the relative pose.

3.3 Human 3D Reconstruction

After the tremendous improvement in 2D and 3D pose estimation from single images, many works have focused on reconstructing the 3D shape of humans. One of the major

works on this topic is the one from Loper *et al.* [36], where they introduce the SMPL model. This is a parametric human body representation that can be used to recover the full body shape and joints using a set of 10 body shape parameters and 72 pose parameters. The shape parameters were found by using PCA on a number of T-posed human 3D scans, and encode different shape aspects of the human body. The pose parameters are used to find the position of each joint, and they use a learned skinning process to change the body shape from T-pose to the final body pose. Due to simplicity and robustness of this model, several works [37, 38, 39] have used it for reconstruction of human bodies from single images.

At the end of the day, however, the SMPL model can only reconstruct unclothed people, but modelling clothing [40, 41] still represents a very challenging task for human reconstruction. Even though SMPL estimations can provide rough human shape reconstruction, several methods have tried to complement the parametric model via voxel-based [42] or implicit function [43] reconstructions. Deephuman [42] uses the SMPL template as the basis on which they train a 3D Auto-encoder that fills clothing details. To make predictions look realistic, they add wrinkles by modifying the vertices location via predicted normal maps. Implicit functions models [44, 45] were initially proposed for object reconstruction [46, 47, 45], and have already been used to complement with SMPL [43]. These methods can represent a much higher resolution density than voxel-based models, by classifying each 3D point as being in or outside the human mesh.

However, some template-free reconstruction methods have also been recently proposed. For instance, PIFu [48, 49] propose a pixel-based implicit function architecture that obtains very accurate reconstructions from single views, in experiments with simple body poses and front-parallel camera views. By using geometric cues, other works have tried to reconstruct 3D body shape in more in-the-wild setups. Pumarola *et al.* [12] propose a mesh-to-image mapping and back, to train the model via 2D convolutions, and Gabeour *et al.* [50] propose to predict a double depth map of a person. These contain the visible depth map and the depth map obtained when tracing each pixel to the furthest point of the person, and then combine them to create a 3D mesh.

3.4 Hand 3D Reconstruction

Most literature on 3D hand analysis is focused on estimating hand pose, represented by a skeleton with up to 21 joints. This problem has been studied for years, either taking as input RGB-D [51, 52, 53, 54, 55, 56] or RGB images [57, 58, 59, 60, 61, 62, 63, 64, 65]. As in the full-body pose estimation field, the community is also recently shifting to estimating hand 3D shape from RGB inputs [64, 62, 65, 66].

The SMPL parametric model has been extended for hand modelling, with the MANO layer [67], to be able to accurately represent hand pose and shape diversity. As shown in recent works *et al.* [62, 65, 66], the parametric model can also be used to estimate hand pose and shape, even in the case of hand-object interactions [65, 66].

3.5 Human Motion Prediction

Since the release of large-scale MoCap datasets [13, 14] and improvements on human pose representations [25, 36], there has been a growing interest in the task of human motion prediction problem.

Most approaches build upon RNNs [68, 69, 70, 71, 72, 73] that encode historical motion of the human and predict the future configuration that minimizes different sort of losses. One of the main works in this topic is the one from Martinez *et al.* [69], which model the velocity of the human body. They also introduce a model-free baseline, which yields very reasonable results under the L2 metric, and proves the difficulty of predicting realistic human motion. This phenomenon has been recently discussed by Ruiz *et al.* [74], that argue that L2 distance is not an appropriate metric to capture the actual distribution of human motion, and that a network trained using only this metric is prone to converge to a mean body pose. To better capture real distributions of human movement, some recent approaches use adversarial networks [11, 75] in combination with geometric losses [71, 70, 74, 76]. Apart from RNN-based models, Jain *et al.* [77] consider a hand-crafted spatial-temporal graph adapted to the skeleton shape, and Li *et al.* [78] use Convolutional Neural Networks to encode and decode skeleton sequences instead of RNNs.

This field closely follows the improvements on the field of human pose and shape estimation. Since recent works on human 3D reconstruction have improved significantly, motion prediction works also start considering full body shapes instead of skeleton representations. For instance, the work from Zhang *et al.* [79] uses the SMPL representation to both reconstruct shape from video and extend motion into the future.

4 Expected contributions beyond state of the art

In this chapter, we will discuss in more detail the topics that we want to tackle in this thesis, and describe the goal contributions.

4.1 Methodology

This subsection will review the methodology that we will follow in the thesis, for the target topics we want to tackle.

Human Pose and Shape Estimation There has been a huge amount of work in 2D and 3D human pose estimation in the past recent years, with state-of-the-art algorithms achieving good performances in on-the-wild datasets. For this reason, in this thesis we will mainly focus on 3D shape reconstruction from single images. The community has already been working on this field for a few years and some works obtained impressive results in particular instances. However, we find that there are a few challenges ahead that we think are key to enhance current methods:

- There are no in-the-wild datasets with clothed human shape annotations. Due to the difficulty of getting such dataset with accurate labelling, some works have proposed the use of physics engines for generating large-scale synthetic datasets. The most recent ones that contain clothed people are the 3DPeople dataset [12], released by this research group, and the CLOTH3D [80]. Even though both works try to solve this challenge, we believe there is a big margin of improvement. Both contain just a single piece of clothing which does not deform according to the motion of the human and, in the latter work, people do not contain hair or shoes either.
- Even though the SMPL model allows to predict a rough estimate of the human shape, it can only represent naked people. Many works have used this parametric model as a first estimation, due to its robustness, and have later exploited voxel-based [42] or implicit function [48, 43] representations to fill the missing cloth sections. But these approaches also have drawbacks, both require high computational resources and are far from working in real time. Also, they arguably lack the robustness of parametric models to *ie* predict realistic shape structures in not visible parts, for which some works have also used normal maps.

We have already worked on the topic for the initial case of hand pose and shape estimation from a single image [65]. In particular, we developed a method that predicts how the hand mesh should be placed to grasp a set of particular objects. The algorithm also looks at the context scenario to avoid interpenetration with other objects nor with

the table. We tackle the mentioned lack of realistic datasets by annotating our own set of realistic grasps for everyday objects, following a distribution of human grasp types [81].

To overcome the mentioned challenges in the case of full body reconstruction, we first want to design and extend a recent work on cloth parametric models [82]. Again, this assumes only one piece of clothing, although it provides a useful case example on how we can work towards cloth parametric models. Our vision is that a human can be represented first by the pose and shape SMPL parameters, and by each of the clothing parameters. In practice, we would have a few parameters for each garment type, and a binary encoding to represent what garment is the human actually wearing. This is fast, does not require expensive computation, and the forward process from representation to full body and garment meshes is fully differentiable.

For this, we need a significant amount of 3D human people wearing diverse clothing in a diverse range of poses. So we are planning to create and release a dataset of realistic clothed humans, dressed with multiple clothing that moves realistically according to the motion of the human. The dataset would be oriented to have very diverse types of clothing, realistic dressed people in single or in multiple-person scenarios.

Human Motion Prediction The ability to predict motion is directly linked to the capacity of understanding present human pose and shape. Typically, some works have used similar representation as in skeleton pose estimation, to process a group of observations that allow to project predictions into the future.

This topic generally draws much less attention to other topics related to human pose understanding. Therefore, in the first stages of the thesis, we have worked on this field to also gain insights of the current state-of-the-art in human 2D/3D pose and shape estimation. The work developed during this period [83] improves human motion prediction by considering other objects and people that the target person might interact with.

We plan to revisit this topic in the later stages of the doctoral thesis, after working on more pure human and cloth representations for 3D shape estimation. We hope we can apply the knowledge to represent clothed people and, similarly to recent work that only uses the SMPL model [79], process and project them into future predictions.

Human-Object Interaction We believe that Human-Object Interactions can provide a significant amount of information for improving human pose and shape estimation state-of-the-art. We plan to design models that exploit semantic cues to better understand the context around people motion and actions, along with their environment.

This field contains few datasets with manually annotated Human-Object interactions. Although the majority of the progress in the field has arrived via supervised learning, we believe we can provide very useful insights for Human-Object interactions learned via semi-supervised or unsupervised learning. In the two works developed so far [83, 65], we

have successfully exploited contextual information without any supervision. For instance in GanHand [65], where we study how humans should grasp objects depending on the setup of the scene, aiming for a reasonable variety of grasps types [81].

4.2 Expected Contributions

The long term goal of this work is to research on algorithms for human pose and shape estimation, motion prediction and scene understanding that would allow robots to be more autonomous and safe in human environments.

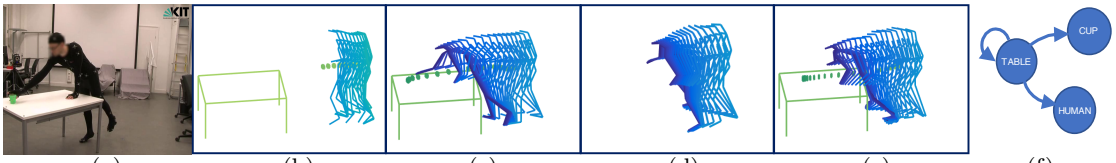
The short term goal is to investigate contextual human perception in a variety of tasks, such as human pose/shape estimation and human motion prediction. We next summarize the specific goals of our proposed research:

- Explore novel geometric and semantic priors that can integrate with deep learning models, and improve their performance
- Develop algorithms to push state-of-the-art in 3D reconstruction and human motion prediction
- Propose a model for garment representation that can help to represent clothed human meshes
- Study semi-supervised and unsupervised methods that allow to reason about Human-Object interactions
- Build solutions applicable to images "on the wild" which do not require any kind of calibration step.

5 Preliminary Results

The work performed in the first year of this PhD goes towards setting the base to build upon following research. Apart from reviewing literature, the first work studied the topic of human motion prediction, which allowed the student to review the literature and understand the state of the art representations of the human body. The second work was on hand reconstruction, and enabled the study of more complex 3D structures and body parameterisations.

5.1 Context-Aware Human Motion Prediction



(a) (b) (c) (d) (e) (f)

Figure 1: **Context-aware human motion prediction.** We propose a new work for motion prediction. By seeing a group of observations of the recent past (b), we aim to predict how the person will move in the following two seconds (c). In comparison with previous state-of-the-art works that do not take advantage of contextual information (d), we devise a context-aware approach that considers all objects and people in the scene to provide a more accurate prediction (e). The model is trained with no supervision on the interactions, which are autonomously predicted by the network (f).

The ability to predict and anticipate future human motion based on past observations is essential for interacting with other people and the world around us. While this seems a trivial task for a person, it is based on complex semantic understanding of the environment and the relations between all objects in it. Modeling and transferring this kind of knowledge to autonomous agents would have a major impact in many different fields, mainly in human-robot interaction [84] and autonomous driving [85], but also in motion generation for computer graphics animation [86] or image understanding [87].

In this work, we argue that current state-of-art algorithms [69, 72] lack to exploit the influence of the rest of the environment on the movement of the person. For instance, if a person is carrying a box, the configuration of the body arms and legs will be highly constrained by the 3D position of that box. Discovering the interrelations between the person and the object/s of the context, and how these interrelations constrain the body motion, is the principal motivation of this paper.

The context-aware motion prediction architecture models the interactions between all objects of the scene and the human using a directed semantic graph. The nodes of this graph represent the state of the person and objects, and the edges their mutual interactions. These interactions are iteratively learned with no ground truth supervision, via an attention model.

We evaluated our approach in the "Whole-Body Human Motion Database" [14] and in the CMU MoCap database [88]. We propose variations of the context-aware models, with different ways to process the context of the person, and with the additional capacity to estimate the motion of rigid bodies. We demonstrate that all context-aware models outperform previous baselines, both quantitatively and qualitatively. Also, we perform a qualitative study and show that the interactions predicted by the model between objects and people are coherent with the actions performed. We finally discuss the applicability of state-of-art motion prediction methods, with an ablation study of our models and baselines when considering noisy observations.

5.2 GanHand: Predicting Human Grasp Affordances in Multi-Object Scenes

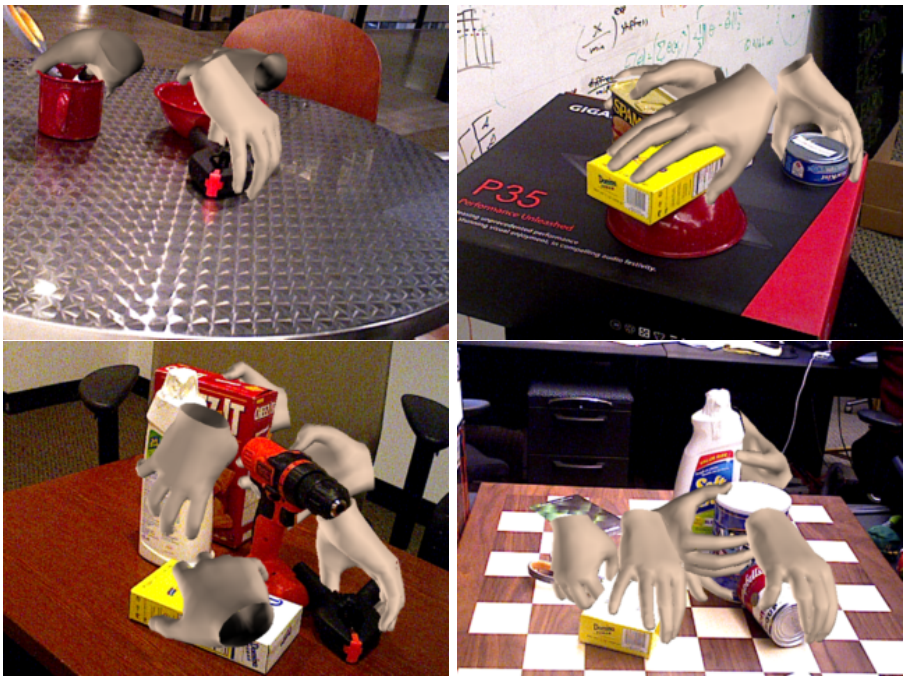


Figure 2: **GanHand: Context-aware human hand reconstruction.** In this work, we propose a method that estimates the hand pose and shape that could grasp the objects in the scene, given a single RGB image. The model reasons about the context, avoiding intersections with the table or other objects, and tries to imitate human grasp types [81]. This figure shows predictions by our model, where we show one grasp prediction for every object.

The problem of estimating 3D hand pose from monocular images has made major advances over the past few years [57, 60, 59, 89, 61, 90, 91]. Current approaches can estimate not only the 3D pose of the hand, but also its shape [64], even when manipulating an object [66].

In this work, we propose a new problem of estimating where the hand might be placed given the contextual information. In particular, we propose a new problem for the community: *given a single RGB image of a scene with an arbitrary number of objects, we aim to predict human grasp affordances*, such as predicting multiple plausible solutions of how a human would grasp each one of the observed objects. This knowledge can have an impact in several fields, such as human-robot interaction, virtual and augmented reality, and robot imitation learning.

In order to predict feasible human grasps, we introduce GanHand, a GAN architecture that takes one RGB image as input and estimates a distribution of grasps for each object on the scene. It first estimates the 3D shape/pose of the objects and predicts the best grasp type according to a taxonomy with 33 classes [81], for each object. It then refines the hand configuration given by the grasping class, through an optimization of the 51 parameters of the MANO model [67]. The model is trained to maximize the number of contact points between the object and the hand shape model while minimizing the interpenetration.

As discussed in Section 4, one of the major challenges in the task of human and hand shape estimation, is obtaining realistic datasets with accurate ground truth annotations. So we also propose and release the YCB-Affordance dataset that we created to train our network. This dataset is based on the 58 household objects of the YCB dataset [92], with *manually* annotated plausible human grasps according again to the taxonomy in [81]. The grasps of 21 objects are then transferred to 92 video sequences, depicting scenes with one or several still objects captured by a moving camera. The total number of annotated frames is 133,936, with more than 28M of realistic grasps, being the *largest dataset of human grasp affordances in real scenes* built so far.

An extensive evaluation on synthetic and real data demonstrates the robustness of GanHand [65] to predict realistic human grasps.

5.3 Cloth parametric model

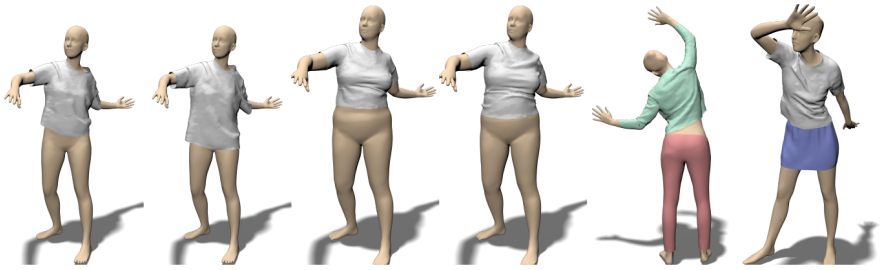


Figure 3: TailorNet [82] proposes the first cloth parametric model, extending SMPL. It can be used to represent a variety of T-shirt topologies, in any human pose or shape. However, scaling this model to new garments requires a significant amount of cloth samples, with enough variety on shape and pose. We plan to generate a large-scale dataset of humans wearing diverse clothes, that allows getting a cloth representation. The same dataset should allow training a model for clothed human shape reconstruction.

As discussed in Sec. 3, parametric models provide a significant amount of robustness and simplicity for human and hand shape representation that is not existing in clothing. For this reason, we are planning to create and release a dataset of realistic clothed humans, dressed with multiple clothing (eg. Coat on top of jumper, both on top of T-shirt), that can be used to find the parametric model of several clothing parts.

We plan to follow the contributions of [82], which are summarised in Fig. 3, to create the parametric model using few parameters per cloth on the unposed human. They can use the same SMPL skinning weights for moving the cloth into the desired person, and model wrinkles and other high-frequency details.. This work is in the very early stages and we have recently started to work on the dataset.

6 Work Plan

In this chapter we will present the different tasks to be carried out for the accomplishment of the objectives (see section 2) of this research plan, as well as an intended Gantt chart for their execution, figure 4. All tasks related to methods include development, testing on synthetic or real data, and, if possible, application on a real problem.

- **Task 1 - State of the Art Review**
  This task aims to meticulously study the current state of the art regarding human pose and shape estimation and motion prediction. Both literature and open source code will be continuously reviewed across the entire PhD to keep updated with new methods and identify future research lines.

- **Task 2 - Motion Prediction**
  In this task, we aim to understand human motion from a series of past observations, and be able to predict how the person is going to move in the following short-time period.

  Sub-task 2.1 - *Context-aware Human Motion Prediction.*

  Within this subtask, we explore the skeleton representation of human bodies on the task of human motion prediction. In particular, we design a model that can take advantage of the context of the person to better understand future motion. The context might include other people to whom the person is interacting with, or other objects, susceptible from manipulation.

  Sub-task 2.2 - *Fully parametric body motion prediction.*

  We aim to predict the motion of the full clothed body shape. Drawing inspiration from [79], we might represent the clothed human body as a group of body and cloth parameters, that we can extend into future observations.

- **Task 3 - Shape reconstruction**

  Our goal in this task is to be able to reconstruct human shape from a single image. This is subdivided into hand and full-body shape estimation.

  Sub-task 3.1 - *Pose and Shape estimation of human hand*

  Within this task, we focus on estimating the 3D pose and shape from the human hand for the task of grasping an object in a given context. The first part of this task was completed while interning in Naver Labs Europe during Summer of 2019, and we look forward to extend this work in a further work.

  Sub-task 3.2 - *Cloth parametric model*

  We aim to extend the standard SMPL representation to also consider different garments. We plan to use the clothing parametric model for full human reconstruction from a single image.

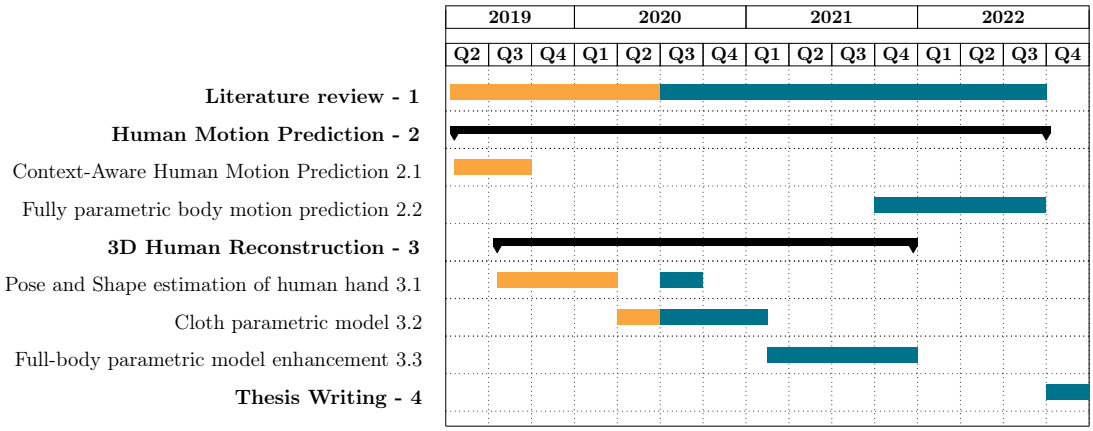


Figure 4: Intended work plan of the proposed thesis

Sub-task 3.3 - *Full-body parametric model enhancement*

After the previous work, the clothed SMPL parameterisation will still miss details that are critical for human characterisation, such as facial elements or hair. This subtask will focus in improving this representation.

- **Task 4 - Thesis Writing**
  The last task of this research is dedicated to the elaboration of the dissertation and the preparation of the public defense.

7 Resources

This Phd thesis will be developed in the “Perception and Manipulation" group at the *Institut de Robòtica i Informàtica Industrial (IRI CSIC-UPC)*[1]. This thesis is financed by the HuMoUR project, which aims to develop computer vision tools to estimate and understand human motion and pose. These algorithms will be used to implement new service robots that can perform complex manipulation for assisting tasks.

To train deep learning models, the student will have access to the “Visen” and “Tro” GPU servers at IRI. The GPU Servers operate with GPUs donated by NVIDIA™.

A research stay in an international center is also planned.

7.1 Publications

During the first year of the PhD, the two works described in 5.1 and 5.2 have been accepted at the *International Conference in Computer Vision and Pattern Recognition* (CVPR) 2020 [83, 65]. The latter [65] has been accepted as an Oral Paper and for a patent application [93]. We also plan to extend the work in two directions. First, a follow up journal version is in preparation and, second, we started a collaboration with Aalto University to implement the proposed ideas into robotic grasping.

The very preliminary work described in 5.3 is in preparation to be submitted at *Computer Vision and Pattern Recognition* (CVPR) 2021.

[1] `www.iri.upc.edu`

References

[1] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics*, vol. 5, no. 4, pp. 115–133, 1943.

[2] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[3] Y. LeCun, P. Haffner, L. Bottou, and Y. Bengio, "Object recognition with gradient-based learning," in *Shape, contour and grouping in computer vision*. Springer, 1999, pp. 319–345.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[6] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[7] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[8] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[10] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014.

[12] A. Pumarola, J. Sanchez-Riera, G. Choi, A. Sanfeliu, and F. Moreno-Noguer, "3dpeople: Modeling the geometry of dressed humans," in *ICCV*, 2019.

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

[14] H. S. Koppula, R. Gupta, and A. Saxena, "Learning human activities and object affordances from rgb-d videos," *IJRR*, 2013.

[15] 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 *ECCV*, 2016.

[16] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis, "Coarse-to-fine volumetric prediction for single-image 3d human pose," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7025–7034.

[17] C. Wang, Y. Wang, Z. Lin, A. L. Yuille, and W. Gao, "Robust estimation of 3d human poses from a single image," in *CVPR*, 2014.

[18] E. Simo-Serra, A. Ramisa, G. Alenyà, C. Torras, and F. Moreno-Noguer, "Single image 3d human pose estimation from noisy observations," in *CVPR*, 2012.

[19] E. Simo-Serra, A. Quattoni, C. Torras, and F. Moreno-Noguer, "A joint model for 2d and 3d pose estimation from a single image," in *CVPR*, 2013.

[20] F. Moreno-Noguer, "3d human pose estimation from a single image via distance matrix regression," in *CVPR*, 2017.

[21] E. Simo-Serra, C. Torras, and F. Moreno-Noguer, "3d human pose tracking priors using geodesic mixture models," *IJCV*, 2017.

[22] B. Tekin, I. Katircioglu, M. Salzmann, V. Lepetit, and P. Fua, "Structured prediction of 3d human pose with deep neural networks," *arXiv preprint arXiv:1605.05180*, 2016.

[23] S. Li and A. B. Chan, "3d human pose estimation from monocular images with deep convolutional neural network," in *Asian Conference on Computer Vision*. Springer, 2014, pp. 332–347.

[24] X. Sun, J. Shang, S. Liang, and Y. Wei, "Compositional human pose regression," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2602–2611.

[25] X. Sun, B. Xiao, F. Wei, S. Liang, and Y. Wei, "Integral human pose regression," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 529–545.

[26] F. Moreno-Noguer, "3d human pose estimation from a single image via distance matrix regression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2823–2832.

[27] W. Yang, W. Ouyang, X. Wang, J. Ren, H. Li, and X. Wang, "3d human pose estimation in the wild by adversarial learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5255–5264.

[28] G. Rogez, P. Weinzaepfel, and C. Schmid, "Lcr-net: Localization-classification-regression for human pose," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3433–3441.

[29] ——, "Lcr-net++: Multi-person 2d and 3d pose detection in natural images," *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[30] G. Moon, J. Y. Chang, and K. M. Lee, "Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 10 133–10 142.

[31] L. Kumarapu and P. Mukherjee, "Animepose: Multi-person 3d pose estimation and animation," *arXiv preprint arXiv:2002.02792*, 2020.

[32] D. Mehta, O. Sotnychenko, F. Mueller, W. Xu, S. Sridhar, G. Pons-Moll, and C. Theobalt, "Single-shot multi-person 3d pose estimation from monocular rgb," in *2018 International Conference on 3D Vision (3DV)*. IEEE, 2018, pp. 120–130.

[33] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt, "Vnect: Real-time 3d human pose estimation with a single rgb camera," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 1–14, 2017.

[34] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7291–7299.

[35] G. Rogez, P. Weinzaepfel, and C. Schmid, "Lcr-net: Localization-classification-regression for human pose," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3433–3441.

[36] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "Smpl: A skinned multi-person linear model," *ACM transactions on graphics (TOG)*, vol. 34, no. 6, pp. 1–16, 2015.

[37] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7122–7131.

[38] Y. Xu, S.-C. Zhu, and T. Tung, "Denserac: Joint 3d pose and shape estimation by dense render-and-compare," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 7760–7770.

[39] M. Kocabas, N. Athanasiou, and M. J. Black, "Vibe: Video inference for human body pose and shape estimation," *arXiv preprint arXiv:1912.05656*, 2019.

[40] E. Corona, G. Alenyà, A. Gabas, and C. Torras, "Active garment recognition and target grasping point detection using deep learning," *Pattern Recognition*, vol. 74, pp. 629–641, 2018.

[41] A. Gabas, E. Corona, G. Alenyà, and C. Torras, "Robot-aided cloth classification using depth information and cnns," in *International Conference on Articulated Motion and Deformable Objects*. Springer, 2016, pp. 16–23.

[42] Z. Zheng, T. Yu, Y. Wei, Q. Dai, and Y. Liu, "Deephuman: 3d human reconstruction from a single image," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 7739–7749.

[43] Z. Huang, Y. Xu, C. Lassner, H. Li, and T. Tung, "Arch: Animatable reconstruction of clothed humans," *arXiv preprint arXiv:2004.04572*, 2020.

[44] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "Deepsdf: Learning continuous signed distance functions for shape representation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 165–174.

[45] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, "Occupancy networks: Learning 3d reconstruction in function space," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4460–4470.

[46] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su *et al.*, "Shapenet: An information-rich 3d model repository," *arXiv preprint arXiv:1512.03012*, 2015.

[47] E. Corona, K. Kundu, and S. Fidler, "Pose estimation for objects with rotational symmetry," in *IROS*, 2018.

[48] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li, "Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 2304–2314.

[49] S. Saito, T. Simon, J. Saragih, and H. Joo, "Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization," *arXiv preprint arXiv:2004.00452*, 2020.

[50] V. Gabeur, J.-S. Franco, X. Martin, C. Schmid, and G. Rogez, "Moulding humans: Non-parametric 3d human shape estimation from single images," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 2232–2241.

[51] J. S. Supancic, G. Rogez, Y. Yang, J. Shotton, and D. Ramanan, "Depth-based hand pose estimation: Methods, data, and challenges," *Int. J. Comput. Vis.*, vol. 126, no. 11, pp. 1180–1198, 2018.

[52] C. Choi, S. Ho Yoon, C.-N. Chen, and K. Ramani, "Robust hand pose estimation during the interaction with an unknown object," in *ICCV*, 2017.

[53] M. Oberweger and V. Lepetit, "Deepprior++: Improving fast and accurate 3d hand pose estimation," in *ICCV Workshops*, 2017.

[54] S. Yuan, G. Garcia-Hernando, B. Stenger, G. Moon, J. Yong Chang, K. Mu Lee, P. Molchanov, J. Kautz, S. Honari, L. Ge *et al.*, "Depth-based 3d hand pose estimation: From current achievements to future goals," in *CVPR*, 2018.

[55] H. Hamer, K. Schindler, E. Koller-Meier, and L. Van Gool, "Tracking a hand manipulating an object," in *ICCV*, 2009.

[56] C. Keskin, F. Kıraç, Y. E. Kara, and L. Akarun, "Hand pose estimation and hand shape classification using multi-layered randomized decision forests," in *ECCV*, 2012.

[57] C. Zimmermann and T. Brox, "Learning to estimate 3d hand pose from single rgb images," in *ICCV*, 2017.

[58] P. Panteleris and A. Argyros, "Back to rgb: 3d tracking of hands and hand-object interactions based on short-baseline stereo," in *ICCV Workshops*, 2017.

[59] F. Mueller, F. Bernard, O. Sotnychenko, D. Mehta, S. Sridhar, D. Casas, and C. Theobalt, "Ganerated hands for real-time 3d hand tracking from monocular rgb," in *CVPR*, 2018.

[60] A. Spurr, J. Song, S. Park, and O. Hilliges, "Cross-modal deep variational hand pose estimation," in *CVPR*, 2018.

[61] U. Iqbal, P. Molchanov, T. Breuel Juergen Gall, and J. Kautz, "Hand pose estimation via latent 2.5 d heatmap regression," in *ECCV*, 2018.

[62] S. Baek, K. I. Kim, and T.-K. Kim, "Pushing the envelope for rgb-based dense 3d hand pose estimation via neural rendering," in *CVPR*, 2019.

[63] A. Boukhayma, R. d. Bem, and P. H. Torr, "3d hand shape and pose from images in the wild," in *CVPR*, 2019.

[64] L. Ge, Z. Ren, Y. Li, Z. Xue, Y. Wang, J. Cai, and J. Yuan, "3d hand shape and pose estimation from a single rgb image," in *CVPR*, 2019.

[65] E. Corona, A. Pumarola, G. Alenyà, F. Moreno-Noguer, and G. Rogez, "Ganhand: Predicting human grasp affordances in multi-object scenes," in *CVPR*, 2020.

[66] Y. Hasson, G. Varol, D. Tzionas, I. Kalevatykh, M. J. Black, I. Laptev, and C. Schmid, "Learning joint reconstruction of hands and manipulated objects," in *CVPR*, 2019.

[67] J. Romero, D. Tzionas, and M. J. Black, "Embodied hands: Modeling and capturing hands and bodies together," *SIGGRAPH*, vol. 36, no. 6, Nov. 2017.

[68] K. Fragkiadaki, S. Levine, P. Felsen, and J. Malik, "Recurrent network models for human dynamics," in *ICCV*, 2015.

[69] J. Martinez, M. J. Black, and J. Romero, "On human motion prediction using recurrent neural networks," in *CVPR*, 2017.

[70] L.-Y. Gui, Y.-X. Wang, X. Liang, and J. M. Moura, "Adversarial geometry-aware human motion prediction," in *ECCV*, 2018.

[71] E. Barsoum, J. Kender, and Z. Liu, "Hp-gan: Probabilistic 3d human motion prediction via gan," in *CVPR-Workshop*, 2018.

[72] D. Pavllo, D. Grangier, and M. Auli, "Quaternet: A quaternion-based recurrent model for human motion," *arXiv preprint arXiv:1805.06485*, 2018.

[73] E. Aksan, M. Kaufmann, and O. Hilliges, "Structured prediction helps 3d human motion modelling," in *ICCV*, 2019.

[74] A. H. Ruiz, J. Gall, and F. Moreno-Noguer, "Human motion prediction via spatio-temporal inpainting," in *ICCV*, 2019.

[75] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *ICML*, 2017.

[76] J. N. Kundu, M. Gor, and R. V. Babu, "Bihmp-gan: Bidirectional 3d human motion prediction gan," *arXiv preprint arXiv:1812.02591*, 2018.

[77] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena, "Structural-rnn: Deep learning on spatio-temporal graphs," in *CVPR*, 2016.

[78] Y.-L. Li, S. Zhou, X. Huang, L. Xu, Z. Ma, H.-S. Fang, Y.-F. Wang, and C. Lu, "Transferable interactiveness prior for human-object interaction detection," *arXiv preprint arXiv:1811.08264*, 2018.

[79] J. Y. Zhang, P. Felsen, A. Kanazawa, and J. Malik, "Predicting 3d human dynamics from video," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 7114–7123.

[80] H. Bertiche, M. Madadi, and S. Escalera, "Cloth3d: Clothed 3d humans," *arXiv preprint arXiv:1912.02792*, 2019.

[81] T. Feix, J. Romero, H.-B. Schmiedmayer, A. M. Dollar, and D. Kragic, "The grasp taxonomy of human grasp types," *Trans. on Human-Machine Systems*, vol. 46, no. 1, pp. 66–77, 2015.

[82] C. Patel, Z. Liao, and G. Pons-Moll, "The virtual tailor: Predicting clothing in 3d as a function of human pose, shape and garment style," *arXiv preprint arXiv:2003.04583*, 2020.

[83] E. Corona, A. Pumarola, G. Alenyà, and F. Moreno-Noguer, "Context-aware human motion prediction," in *CVPR*, 2020.

[84] H. S. Koppula and A. Saxena, "Anticipating human activities using object affordances for reactive robotic response," *TPAMI*, 2016.

[85] B. Paden, M. Čáp, S. Z. Yong, D. Yershov, and E. Frazzoli, "A survey of motion planning and control techniques for self-driving urban vehicles," *IV*, 2016.

[86] L. Kovar, M. Gleicher, and F. Pighin, "Motion graphs," in *SIGGRAPH*, 2008.

[87] Y. Chen, M. Rohrbach, Z. Yan, S. Yan, J. Feng, and Y. Kalantidis, "Graph-based global reasoning networks," *arXiv preprint arXiv:1811.12814*, 2018.

[88] C. G. Lab, "Cmu motion capture database," http://mocap.cs.cmu.edu/.

[89] Y. Cai, L. Ge, J. Cai, and J. Yuan, "Weakly-supervised 3d hand pose estimation from monocular rgb images," in *ECCV*, 2018.

[90] P. Panteleris, I. Oikonomidis, and A. Argyros, "Using a single rgb frame for real time 3d hand pose estimation in the wild," in *WACV*, 2018.

[91] M. Rad, M. Oberweger, and V. Lepetit, "Domain transfer for 3d pose estimation from color images without manual annotations," in *ACCV*. Springer, 2018.

[92] B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel, and A. M. Dollar, "The ycb object and model set: Towards common benchmarks for manipulation research," in *ICAR*, 2015.

[93] E. Corona, A. Pumarola, G. Alenyà, F. Moreno-Noguer, and G. Rogez, "Method for determining a grasping hand model," Jun. 9 2020, ES Patent App. 202030553.