SCIENTIFIC REPORT 2016 - 2018

Max Planck Institute for Intelligent Systems

Perceiving Systems Department

Stuttgart & Tübingen

MAX-PLANCK-GESELLSCHAFT

Contents

1 PERCEIVING SYSTEMS

1.1 Research Overview

Computer vision is often treated as a problem of pattern recognition, 3D reconstruction, or image processing. While these all play supporting roles, our view is that the goal of computer vision is to *infer what is not in the picture* – to recognize the unseen. This is different from the Aristotelian view that the goal of vision is "to know what is where by looking." We see vision as the process of inferring the causes and motivations behind the images that we observe; that is, we want to infer the story behind the picture.

The most interesting stories involve people. Consequently, our research focuses on understanding humans and their actions in the world. We aim to recover human behavior in detail, including human-human interactions, and human interactions with the environment.

Humans interact with each other and manipulate the world through their bodies, faces, hands and speech. If computers are to understand humans and our behavior, then they are going to have to understand much more about us than they currently do. For example, they need to recognize when we are picking up something heavy and might need help. They need to understand when we are distracted. They need to understand that changes in our behavior may signal medical or psychological problems.

To address this, we are developing the datasets, tools, models, and algorithms to recover human movement in unconstrained scenes at a level not previously possible. From single images or videos, we estimate full 3D body pose, including the motion of the face and the pose of the hands. We also recover the 3D structure of the world, its motion, and the objects in it so that human movement can be placed *in context.* We are not just interested in pose but also what the person is in contact with, what they are holding, where they are looking, who they are interacting with,

and what they may do next.

This is quite different from previous work in which the human body is treated in isolation, removed from the world around it, and 3D scene analysis happens on static scenes without humans. We think the interesting research problems involve analyzing human behavior when people are present in, and interacting with, the 3D world. By building 3D models of people and how they move, we are able to place them in context and reason about the goals behind their behavior and the physical contraints on this behavior.

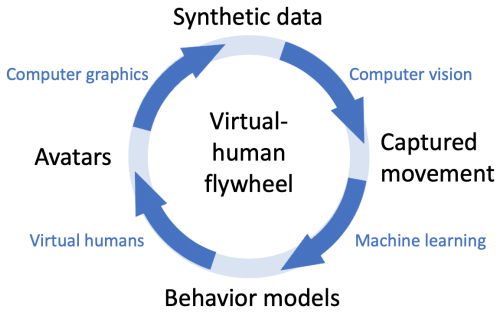


Figure 1.1: The virtual-human flywheel. By building better models of humans, we can simulate better training data, which leads to better vision algorithms, that help us gather more data about how humans behave, which helps us build better models of humans. This touches the core competencies of Perceiving Systems: Computer Vision, Machine Learning, Computer Graphics, and Virtual Humans.

To advance this agenda, Perceiving Systems combines computer vision with machine learning and computer graphics. For example, our 3D graphics models of the body enable us to generate training data for machine learning methods, which improve our computer vision algorithms [36, 52]. These improved algorithms give us better data from images and video with which to improve our graphics models, leading to a virtuous cycle as illustrated in Fig. 1.1.

This cycle is producing increasingly realistic virtual humans. We see the virtual human as more than a useful artifact. We see it as a testbed for evaluating our models of human behavior. If we can simulate a virtual human in a virtual world behaving in ways that are indistinguishable from a real human, then we assert that we have captured something essential about what it means to be human. This forces us to go beyond capturing human movement and to model the causes of that movement.

Over the history of Perceiving Systems, we have built the foundational technology for this effort. Specifically, we learn realistic 3D models of human body shape and pose deformation from thousands of detailed 3D scans. We have built many models, but SMPL[1] has become the de facto standard for research on human pose. To go beyond SMPL, we have learned a face model (FLAME) using a novel dataset of 4D facial sequences [13]. FLAME captures realistic 3D head shape, jaw articulation, eye movement, blinking, and facial expressions. Similarly, we developed MANO, a 3D hand model learned from around 2000 hand scans of different people in many poses [12].

In the last three years we have shown how to fit SMPL to image data and how to train deep networks end-to-end to extract full-body shape and pose from single images or video. This includes the following methods, which provide foundational tools for capturing and analyzing human motion in natural settings

- SMPLify [69],
- Unite the People [51],
- Human Mesh Recovery [40],
- Neural Body Fitting [30].

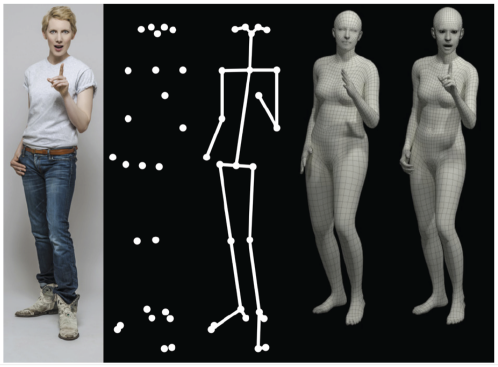


Figure 1.2: Bodies are not a collection of joints and bodies are not a skeleton. Bodies have shape, can move, can express emotion, and can interact with the world. Hence virtual bodies need faces and hands and the ability to move and use them. Our new SMPL-HF model (right) has the expressiveness needed to model human interactions with the world and between people. (Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A. A. A., Tzionas, D., Black, M. J., "Expressive Body Capture: 3D Hands, Face, and Body from a Single Image," CVPR 2019.)

[1] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, M. J. Black. SMPL: A Skinned Multi-Person Linear Model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)* **34** (6): 248:1–248:16, Oct. 2015

1.1.1 Expressive Bodies

SMPL, alone, however is not enough to understand human behavior. Thus, we have combined SMPL, FLAME and MANO into a single model, SMPL-HF, that goes beyond previous work to represent 3D body shape, pose deformations, facial expression, and hand pose in a unified model. Figure 1.2 illustrates how SMPL-HF is more expressive than previous representations of the body that are commonly used today. With SMPL-HF, for the first time, we can estimate information about the body together with hand-object interaction, gestures, and facial expression. We have developed algorithms to estimate the parameters of SMPL-HF from a single image and are working on extending these methods to video and RGB-D. This is the first step towards expressive motion capture in complex scenes that will underpin our future research on human behavior.

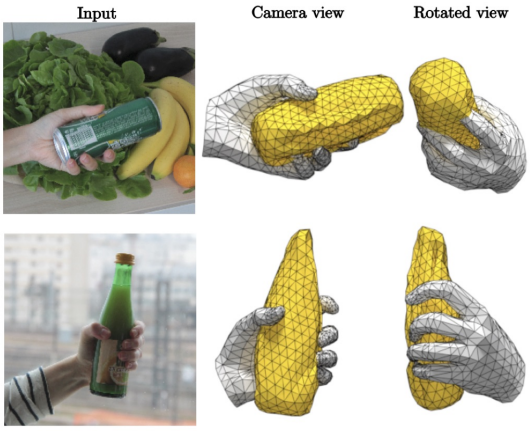


Figure 1.3: Hand pose and object shape are often estimated separately. These methods do not generalize to hand-object interaction. We jointly estimate both the 3D object and the 3D hand from a single image. This joint estimation enables us to penalize interpenetration and encourage contact, leading to more stable grasps. (Hasson, Y., Varol, G., Tzionas, D., Kalevatykh, I., Black, M. J., Laptev, I., Schmid, C., "Learning joint reconstruction of hands and manipulated objects," CVPR 2019.)

A key step in this direction is to analyze hand-object interaction in images and video. To that end, in collaboration with colleagues at INRIA, we have generated a novel synthetic training set of hands interacting with 3D objects; this is a good example of the virtual-human flywheel in action. Using this data, we train a novel neural network to estimate both 3D hand pose and 3D object shape. We observe that these two processes are synergistic and that estimating them together produces better results because occlusion and contact can be modeled. Specifically, we train the model in such a way that we penalize hand-object interpenetration and encourage contact when parts of the hand are close to an object surface. Despite training only on synthetic data, we obtain realistic 3D hand pose/shape and object shape as seen in Fig. 1.3. We show that, by incorporating constraints about hand-object interaction during training, we achieve more stable grasps than when training separate hand and object networks.

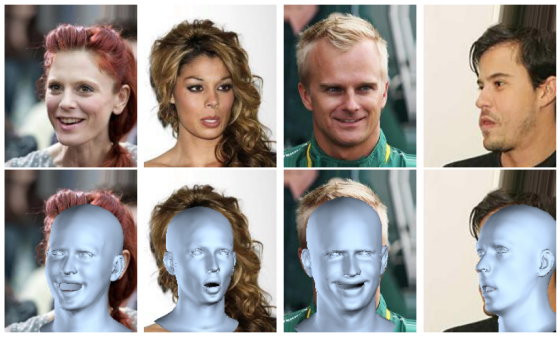


Figure 1.4: Without 3D supervision, RingNet learns a mapping from the pixels of a single image to the 3D facial parameters of the FLAME model. Top: Images are from the CelebA dataset. Bottom: estimated shape, pose and expression. (Sanyal, S., Bolkart, T., Feng, H., Black, M. J., "Learning to Regress 3D Face Shape and Expression from an Image without 3D Supervision," CVPR 2019.)

Like hands, faces and facial expressions are critical to understanding human behavior. The estimation of 3D face shape from a single image must be robust to variations in lighting, head pose, expression, facial hair, makeup, and occlusions. Robustness requires a large training set of in-the-wild images, which by construction, lack ground truth 3D shape. Consequently, to train a network without any 2D-to-3D supervision, we developed *RingNet*, which learns to compute 3D face shape from a single image (Fig. 1.4). Our key observation is that an individual's face shape is constant across images, regardless of expression, pose, lighting, etc. RingNet uses a novel loss that encourages the face shape to be similar when the identity is the same and different for different people. We achieve invariance to expression by representing the face using our FLAME face model. Once trained, our method takes a single image and outputs the parameters of FLAME, which can support the analysis of human behavior.

The above models capture the surface shape of the body and how it varies across people and with pose. To generalize to settings that we have never seen, we learn a physics-based model of

soft-tissue. We extend SMPL from a triangulated mesh model to a volumetric tetrahedral mesh. From 4D scans, we infer the material properties and thickness of the fat under the skin. We can then simulate soft-tissue dynamics and compression using finite-element methods [16].

Since humans wear a wide variety of clothing, we also develop models of clothing and how clothing drapes on the body. Specifically, ClothCap exploits 4D clothing scans to learn how clothing deforms with pose [15]. This enables us to retarget clothing from one person to people of other shapes.

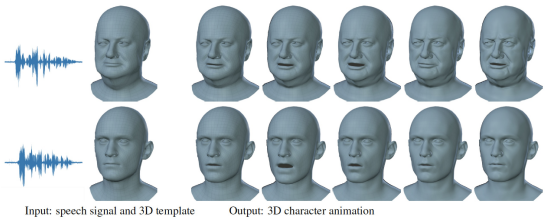


Figure 1.5: We learn a model (VOCA) that relates 3D facial deformations to speech signals. Given an arbitrary speech signal and a static 3D face mesh as input (left), VOCA outputs a realistic 3D character animation (right). (Cudeiro, D., Bolkart, T., Laidlaw, C., Ranjan, A., Black, M. J., "Capture, Learning, and Synthesis of 3D Speaking Styles," CVPR 2019.)

The field has focused on capturing the kinematics of the human body or the parameters of facial muscle activations. These have proven useful for many applications and can be used to animate models of the body. But, we are intersted in more. We want more semantic controls of human activity that relate goals and kinematics in context. As a first step in this direction we have captured an extensive database of 4D scans of people talking. We record both the facial shape and the speech and then train a neural network to relate the two. Using this method (VOCA), we can then animate any 3D face shape saying anything in any language and we can do so in a variety of speaking styles (Fig. 1.5). This is a step towards non-kinematic character control and opens up many avenues for the joint analysis of speech (and associated semantics, sentiment, emotion, etc.) and human behavior.

1.1.2 Bodies in Scenes

Humans and animals live in, and interact with, the 3D world around them. To understand humans then, we must understand the surfaces that support them and the objects with which they interact. To that end, we develop methods to estimate the structure and motion of the world from a single image, video, or multiple images. We approach this by combining unsupervised learning with physical knowlege of the world.

For moving scenes, we compute the optical flow representing the projection of the 3D motion field into the image. In doing so, we exploit the geometric structure of the problem to simplify it. If the scene is rigid and only the camera moves, then the optical flow is completely described by the depth of the scene and the camera motion. Real scenes, however, contain rigid structure and independently moving objects. To deal with this, we segment the scene into regions corresponding to the different types of motion [46]. To do so we exploit different constraints that are both geometric and semantic.

In the latter case, we know that certain objects like animals and cars can move independently while others, like buildings, cannot. Additionally objects like the road are typically planar and hence their motion is simply modeled. Thus, a semantic segmentation of the scene, provides information about what motions may be present where [76]. We argue that segmentation and motion estimation go hand in hand and we have explored methods to do both in a coupled way.

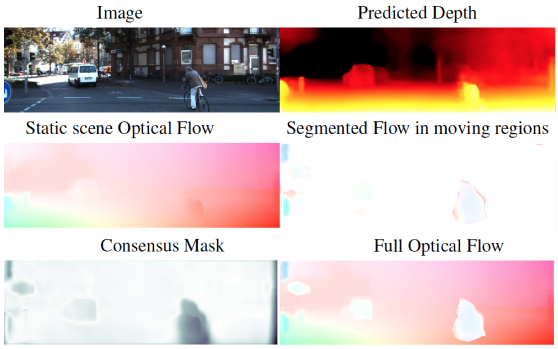


Figure 1.6: Unsupervised learning of depth, camera motion, motion segmentation (mask), and optical flow. Depth and camera motion gives rise to the flow in rigid regions. The networks learn to segment the non-rigid regions and compute flow in these. Everything is learned without supervision using a novel "collaborative competition" method. (Ranjan, A., Jampani, V., Balles, L., Kim, K., Sun, D., Wulff, J., Black, M. J., "Competitive Collaboration: Joint Unsupervised Learning of Depth, Camera Motion, Optical Flow and Motion Segmentation," CVPR 2019.)

Our most recent work focuses on the unsupervised learning of motion, scene depth, camera motion, and segmentation (Fig. 1.6). Training exploits *competitive collaboration* in which dif-

ferent neural networks vie to explain the motion in the scene [8]. To make this work, the physical constraints about motion in rigid scenes are critical. This geometric information allows us to learn depth from a single image without supervision. Motion makes this possible because scene structure is constant over time, allowing motion over time to inform the network about scene depth.

This work demonstrates that classical physical and geometric constraints that we already know about the world and its motion are not only compatible with deep learning but can play a critical role in enabling unsupervised learning. In a sense, the physics of the world provides a form of pre-existing supervision.

Our ongoing work combines people, scenes, and unsupervised learning to derive coherent explanations of the 3D world. We posit that the joint estimation of people and scenes will improve both.

1.1.3 Beyond Mocap

To understand human behavior, we must record that behavior across long periods of time, record interactions between people, and capture how people interact with their environment. Motion capture in lab environments is highly accurate but, necessarily, restricted in terms of realism and the length of sequences that can be captured. Vision-based methods for monocular mocap are designed to work in arbitrary scenes but still lack accuracy. No current method enables accurate capture of human motion in natural settings over long periods of time. Consequently, we are developing a range of motion capture technologies that can break free of the laboratory and capture more realistic behavior. As a first step, we are developing capture technologies that use IMUs, hand-held cameras, or swarms of drones to enable 3D human motion estimation in natural settings like outdoors, around town, or at home (Fig. 1.7).

The above methods, models, and datasets support our long-term goal of teaching computers to see us and, understand our behavior, and to mimic this behavior in virtual human avatars.

Figure 1.7: To take mocap out of the lab we use a small number of IMUs mounted on the body. Top: We train a neural network to regress full body pose from 6 IMUs. Bottom: We combine IMUs with a hand-held camera to obtain 3D poses in natural videos.

1.1.4 Impact and Outreach

While our focus is basic research on computer vision, graphics and machine learning, we want to have an impact beyond these academic disciplines. Consequently, we pursue collaborations that allow our work to have a broader impact.

Figure 1.8: Our body models are now used for problems in psychology and medicine. Our SMPL body model captures realistic body shape and can easily run in MR/VR applications. We have recently developed simple methods where clinicians can create avatars using only the controllers of an HTC Vive.

For example, we develop applications in medicine and psychology in collaboration with medical colleagues. We have collaborative efforts to relate the distribution of adipose tissue in the body to the risk of diabetes and cardiovascular disease [20]. We have developed a 3D model of infants and use it to track their movement to aid the early diagnosis of cerebral palsy [38]. Our 3D body model has also played an important role in understanding how women who suffer from anorexia nervosa see their body and the bodies of others [9–11]. We continue to collaborate with psychologists and doctors on a range of related topics and, through these collaborations, have developed tools like the Virtual Caliper (Fig. 1.8),

which makes it easy for practitioners to create realistic 3D body avatars and animate them in virtual reality (VR) or mixed reality (MR).
More information: https://ps.is.mpg.de/field/medicine-and-psychology

Figure 1.9: From a few snapshots of an animal, we reconstruct the detailed 3D shape. This can then be animated or 3D printed.

We are also collaborating with researchers on animal conservation. Specifically, we work with Wildbook on a project to protect the Grevy Zebras in Africa. We have developed technology that takes a few images of an animal and creates realistic 3D models (Fig. 1.9). With Wildbook, we are working on the first methods to analyze herd shape from photos and on developing drone-based surveying methods.
More information: https://ps.is.mpg.de/project/capturing-animal-shape

We are also active in patenting, technology licensing, and startups. We have spun off two companies that are using our 3D body model technology. One of these, **Body Labs Inc.**, was acquired by Amazon in 2017. The second, **Meshcapade GmbH**, started in 2018 and provides services and software for processing 3D scans and motion capture data.

We also make code and data available open source or for license. For example, over the last three years, our open source differentiable renderer (OpenDR) had a significant impact and has kick-started research on this topic. Additionally, our SMPL body model is now in wide use in both academia and industry for representing 3D body shape. Our code for estimating SMPL from images (SMPLify) has helped drive the field to solve this challenging problem.

Finally, we are responsible for, or contribute to, widely used datasets and evaluation benchmarks that help push the state of the art and provide a platform for industry to understand what works, how well, and why. We have played central roles in many influential datasets and evaluations in the field including Middlebury Flow, Sintel, KITTI, HumanEva, FAUST, JHMDB, and others. Over the last three years we have released several major datasets related to faces, hands, 3D bodies, clothing, animals, optical flow, and IMUs.

Datasets and code released in the last three years include: SMPL[ref:ps:smpl2015]; SMPLify [69]; FLAME [13] (model, fitting code, and registered meshes); CoMA [34] (code, model and data); MANO [12] (hand scans and model); Dynamic FAUST [55] (precise 4D scans in correspondence); 3D poses in the wild (3DPW) [32]; SURREAL [52] (synthetic humans for training deep networks); SlowFlow [54] (optical flow in real scenes); Unite the People [51] (training set for 3D human pose from images); BUFF [62] (body shape under clothing); SMAL [57] and SMALR [41] models of 3D animals; SMIL [38] infant body model; DIP [5] (code and data for 3D human pose from IMUs); AirCap [7, 14] (design and software for aerial motion capture).
More information: https://ps.is.mpg.de/code

1.1.5 About Us

The Perceiving Systems department was founded in January 2011 and today has about 50 members from all over the world. This includes support staff, technicians, students, guests, and scientists at various career stages. We have about 60 alumni including eight graduated Ph.D. students. Many of our alumni have gone on to academic positions, founded companies, or joined major research laboratories.

We take diversity seriously and, between 2016 and 2018, 30% of our papers had female first authors and 50% had at least one female author.

Figure 1.10: Perceiving Systems is highly international, diverse, and collaborative.

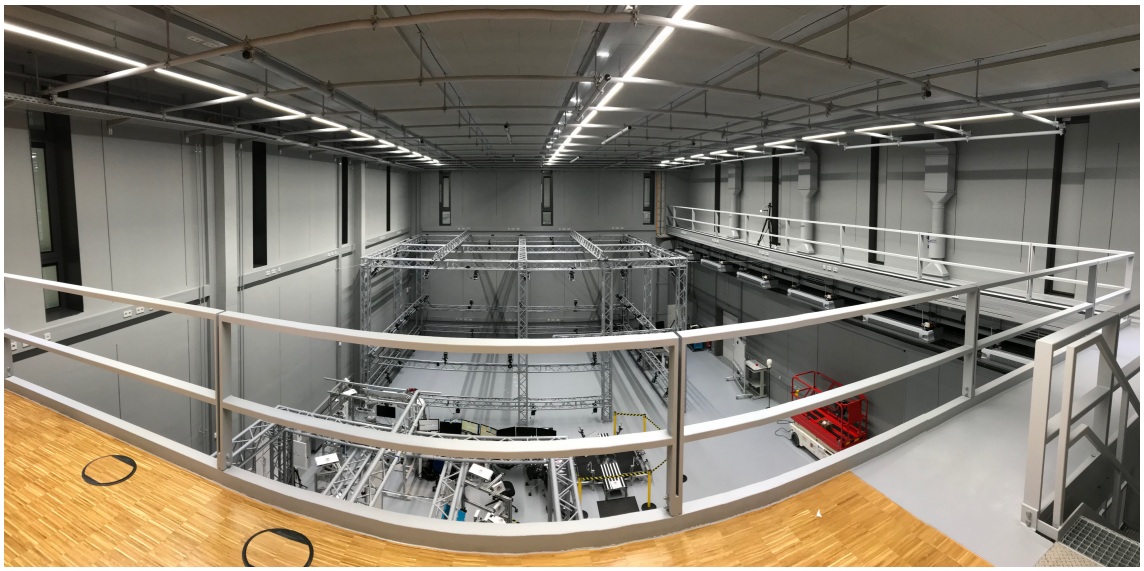
The department hosts two group leaders, Siyu Tang and Aamir Ahmad, who lead groups focused on "Holistic Vision" and "Robot Perception" respectively. Group leaders receive department funding and raise external funds to support their research. Details about these groups can be found on our website: https://ps.is.mpg.de/field/robot-perception-group and https://ps.is.mpg.de/field/holistic-vision-group

We have regular sabbatical and long-term visitors. Between 2016 and 2018 we hosted Cordelia Schmid (INRIA) as a Humboldt Professor, Andrew Blake (Samsung and FiveAI), and Hedvig Kjellström (KTH) as a sabbatical visitor. We also have a highly active visitor program and lecture series. We have had over 175 invited speakers, including many of the leaders in the field. A full list is here https://ps.is.tuebingen.mpg.de/talks

In what follows, we present a sampling of our research projects over the last three years. Our website provides information about many other projects as well as greater detail: https://ps.is.tue.mpg.de

A broader view of the department activities, including more of the social life, can be found on our Facebook page: https://www.facebook.com/PerceivingSystems/

1.1.6 Facilities



Two years ago, Perceiving Systems moved into the new building and we finally had laboratory space for our research on human shape and motion. In our Capture Hall we run a large and complex range of equipment and experiments in approximately 830 square meters of space. During 2016–2018 we captured 2,334,643 3D scans broken down into 938,017 full body scans, 1,082,752 face scans, 236,225 hand scans, and 77,649 foot scans. We know of no scanning effort of this size elsewhere in the world. Keeping this running is a professional staff of two human subjects coordinators, two scanner technicians, and three software engineers who support the custom software that processes all this data.

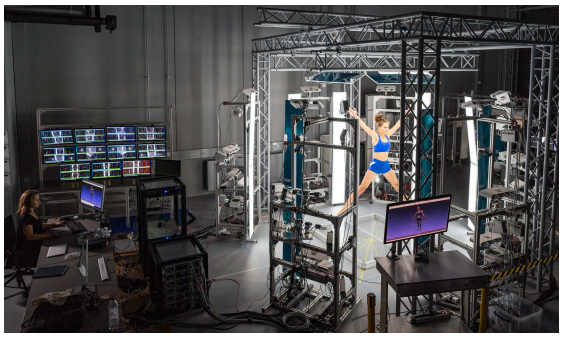


Figure 1.11: The 4D body scanner captures 3D meshes of the full body at 60 fps.

The centerpiece of this facility is our 4D body scanner (Fig. 1.11) made by 3dMD, which was the first of its kind that could capture the full range of human movement at 60 fps. At each time instant the system captures a 3D point cloud with about $150k$ points. We capture people both in minmal clothing and in normal street clothes. The scan data is then processed so that our 3D body model is aligned to the data using our own methods to produce detailed meshes that are in correspondence across people and poses.

Figure 1.12: The 4D face and hand scanner captures detailed facial or hand shape at 60 fps together with high-quality texture.

The full body scanner has limited resolution in the face region and, given the importance of the face for communication we purchased a dedicated 3dMD system that can capture the full 3D head in detail at 60 fps together with high-quality texture maps (Fig. 1.12). We also reconfigure this system to perform hand scanning but plan (in collaboration with the Haptic Intelligence department) to install a dedicated hand system as well to study hand-object interaction.

Figure 1.13: The 4D foot scanner captures a full 3D view of the foot during contact. The system images the bottom of the foot through a glass plate so that the deformation of the shape is captured. It is able to capture 4D sequences, revealing how feet interact with surfaces.

Feet receive much less attention than faces and hands but are literally the foundation of human movement. Many problems with the capture and animation of human avatars can be traced to the feet, which are typically approximated as rigid shapes. In fact, the feet are complex and highly deformable. If one wants to model the physical interactions between the body and the world, one needs a detailed foot model. Consequently, we have been capturing feet interacting with the ground, both in shoes and barefoot using a new 4D foot scanner (Fig. 1.13) created by 3dMD. We plan to add a detailed foot model to our 3D human model in the near future.

Figure 1.14: We recently added a 54-camera Vicon motion capture system to capture multiple people and complex motions.

While our 4D systems give unprecedented details of the human body in motion, the capture volume of these systems is very limited and this limits the kinds of motions we can study. In particular, we can only capture information about one person at a time. To study human interaction and more complex behaviors, we have installed a Vicon marker-based motion capture system with 54 high-resolution Vantage V16 cameras (Fig. 1.14). This system enables us to capture the motion of multiple people interacting, including the full body together with the face and hands. The high density of cameras minimizes problems with occlusion and the high resolution of the cameras means that we can resolve very small markers on the face and hands.

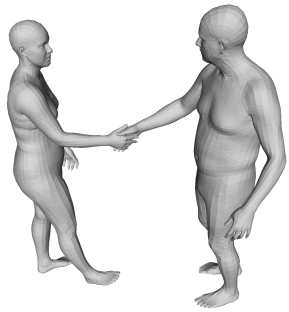


Figure 1.15: Capturing interaction. We are extending all of our capture and modeling methods to deal with human-human and human-object interaction. Here MoSh is extended to extract multiple people interacting and in contact. This is the beginning of a major effort to accurately capture and model contact and behavior.

While such a system gives sparse marker data, our MoSh technology[2] converts this into a full 3D representation of the actors and their movements. We are extending MoSh to capture hands and faces and to automatically fit multiple people (Fig. 1.15). At the same time we are designing new capture scenarios to study human communication, human-human contact, and human-object interaction.

Figure 1.16: As few as six inertial measurement units worn on the body provide enough information to infer human pose. This makes motion capture practical in many scenarios.

Going beyond the confines of the capture hall, we work extensively with IMU-based motion capture (Fig. 1.16). This includes full-body IMU

[2] M. M. Loper, N. Mahmood, M. J. Black. MoSh: Motion and Shape Capture from Sparse Markers. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)* **33** (6): 220:1–220:13, Nov. 2014

capture as well as hand pose capture. While IMUs enable us to capture people in natural settings, they have many drawbacks. We have developed several methods to make them more practical by training a method to estimate body pose from only 6 IMUs and combining IMUs with a hand-held video data to eliminate drift.

Figure 1.17: We have developed custom octo-copters with on board processing and cameras. These form the basis of our flying-motion-capture system, which aims to capture human and animal motion outdoors without any worn sensors.

IMUs also require the subject to cooperate and wear sensors that could affect their movement. This becomes impractical when we want to capture animal movement in the wild. To address this, we are developing a flying motion capture system based on micro-aerial vehicles that fly autonomously, track the subject in real-time, and then estimate the 3D pose of the body over time (Fig. 1.17). Our goal is to make this practical and accurate enough to be used for animal behavior analysis in the wild.

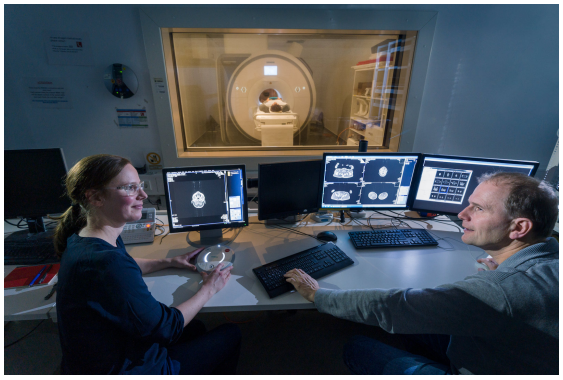


Figure 1.18: To look inside the body, we also scan people in an MRI scanner operated by the MPI for Biological Cybernetics.

Finally, the above work focuses on the outside of the body and its movement but we also capture the inside of the body using MRI scans in humans (Fig. 1.18) and CT scans in animals (rodents). We capture full-body MRI scans of human subjects lying down and 3D surface scans of them lying on a glass table. This enables us to relate the two datasets. Subjects are also scanned standing up so that we can learn how to relate such scans to the compressed shape of people lying down. Together with collaborators at the University Hospital Tübingen, we use this data to analyze the distribution of body fat and to relate this to body shape. In the case of rodents, we again collaborate with the University of Tübingen to model the relationship between the outer surface of the animal and the internal structure. The goal is to predict the motion of the bones and internal organs from only external observations.

1.2 Selected Research Projects

Expressive Body Models

Michael Black, Dimitris Tzionas, Timo Bolkart, Ahmed Osman, Vassilis Choutas, Georgios Pavlakos, Nima Ghorbani

Figure 1.19: We learn a new 3D model of the human body called SMPL+HF that jointly models the human body, face and hands. We fit the female SMPL+HF model to single RGB images and show that it captures a rich variety of natural and expressive 3D human poses, gestures and facial expressions.

Bodies in computer vision have often been an afterthought. Human pose is often represented by 10-12 body joints in 2D or 3D. This is inspired by Johannson's moving light displays, which showed that some human actions can be recognized from the motion of the major joints of the body. But the joints do not capture everything. The skeletal structure of the body is also a popular representation but is only approximate and is never actually observed in images.

In our work we have focused on 3D body shape, represented as a triangulated mesh. Shape gives us more information about a person related to their health, age, fitness, and clothing size. But shape is also useful because our body surface is critical to our physical interactions with the world. We cannot interpenetrate objects and they cannot interpenetrate us.

It has taken a few years for the field to catch on to this idea but now our SMPL[ref:ps:smpl2015] body model is widely used in research and industry. It is simple, efficient, posable, and compatible with most graphics packages. It is also differentiable and easy to integrate into optimization or deep learning methods.

While popular, SMPL has drawbacks. Pose deformations are non-local, the face does not move, the hands are rigid, there is no clothing and no hair. We are addressing these issues in ongoing work (see the theme on Clothing [15, 62] and projects on Faces [13] and Hands [12]). Our latest work is putting bodies, faces and hands together in a simple model that can be fit to data or animated. Like all our body models, we train this from scans of people to capture the realism and statistics of the population [55].

Such models provide the foundation for our analysis of human movement, emotion, and behavior.

More information: https://ps.is.mpg.de/project/expressive-body-models

Faces and Expressions

Michael Black, Timo Bolkart, Anurag Ranjan, Soubhik Sanyal, Tianye Li, Javier Romero, Cassidy Laidlaw

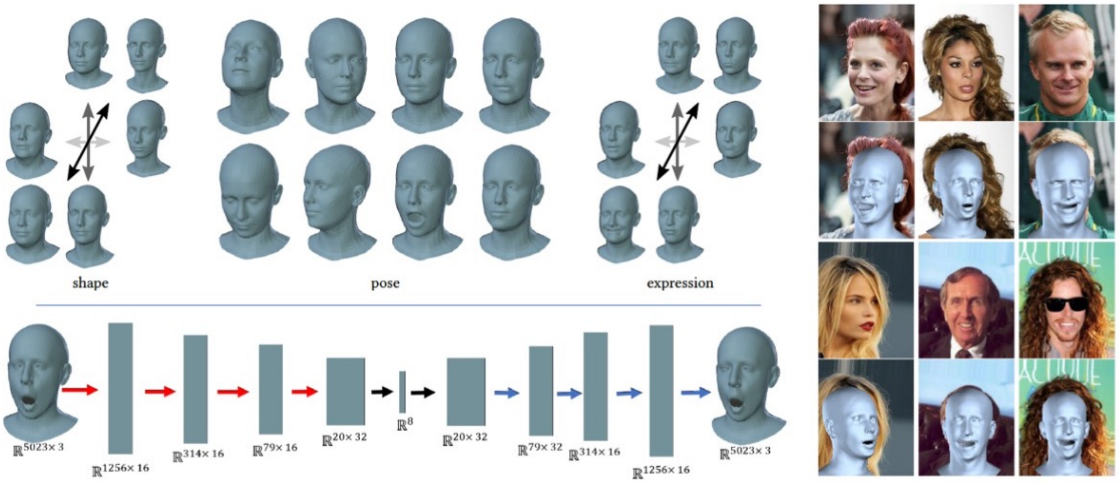


Figure 1.20: 3D Face Analysis: Top left: FLAME [13] captures face shape, pose, and expression with a linear model. Bottom left. CoMA models non-linear deformations using a novel mesh auto-encoder [34]. Right: Learning 3D shape, pose and expression from 2D images without 3D supervision.

Faces, their shape, and their motion are essential to communication. Consequently, we want a model of the face that can capture the full range of face shapes and expressions. Such a model should be realistic, easy to animate, easy to fit to data, and should support inference about human emotion and speech. Additionally we need the tools to estimate faces, their shape, pose, expression, gaze, and movement from images.

To that end, we trained a 3D face model called FLAME [13] from 4D scans. Because it is learned from large-scale, expressive, data of real people, it is more realistic than previous models. FLAME uses a linear shape space trained from 3800 scans of human heads and combines this with an articulated jaw, neck, eyeballs, pose-dependent corrective blendshapes, and additional global expression blendshapes. The pose and expression dependent articulations are learned from 4D face sequences to which we accurately register a template mesh. In total the model is trained from over 33,000 scans.

While expressive, it is difficult to capture the non-linear deformations of extreme expressions with FLAME's low-D linear subspace. While neural networks would be a natural choice for representing such deformations in a low-D latent space, existing convolutional neural networks do not generalize to 3D meshes in a straightforward way. To address this, we introduce a versatile encoder-decoder framework for meshes using spectral convolutions on a mesh surface [34]. Additionally, we introduce mesh up- and down-sampling operations that enable a hierarchical mesh representation that captures non-linear variations in shape at multiple scales. Our CoMA mesh convolution algorithm is generic and now widely used.

To capture, model, and understand facial expressions, we need to estimate the parameters of our face models from images and videos. Training a neural network to regress model parameters from image pixels is difficult because we lack paired training data of images and the true 3D face. To address this we learn this mapping using only 2D image features. The key is to leverage multiple images of a person with a novel loss that encourages the face shape to be similar when the identity is the same and different for different people. FLAME enables the network to factor out changes in expression so that it can exploit this shape constancy.

More information: https://ps.is.mpg.de/project/human-face-analysis

Hands in Action

Dimitris Tzionas, Javier Romero, Michael Black, Gul Varol, Cordelia Schmid, Yana Hasson, Igor Kalevatykh, Ivan Laptev

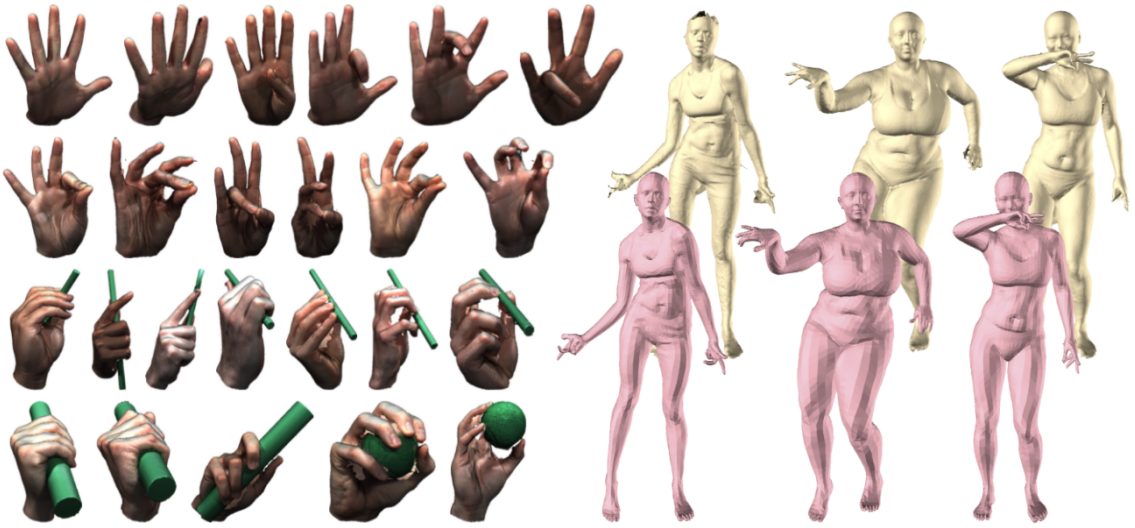


Figure 1.21: We collect 3D scans of human hands (left) from multiple people and model pose and shape variation across people and poses by learning a statistical model of the human hand, called MANO. We then combine the MANO hand model with our SMPL body model to build a holistic model called SMPL+H. The figure (right) shows example 3D scans (white) from our 4D sequences and corresponding fits of SMPL+H to these scans (pink). SMPL+H is able to capture natural motions even under challenging conditions, such as severe missing data due to fast motion, occlusion, finger-webbing, or noise.

Hands are important to humans for signaling and communication, as well as for interacting with the physical world. Capturing the motion of hands is a very challenging computer vision problem that is also highly relevant for other areas like computer graphics, human-computer interfaces, and robotics.

We focus on building an accurate and realistic model of the human hand [12] that captures the pose and shape variation across a human population. For this we collect many examples of human hands with our 3D scanner, following a systematic grasp taxonomy [22]. We then combine the hand model with our SMPL body model to build a seamless model of the body together with hands, called SMPL+H. This allows us to naturally capture the motion of people with expressive body and hand motion using our 4D scanner.

A strong hand model can be used to regularize fitting to noisy input data to reconstruct hands and/or objects [89]. We focus on hands that interact with other hands or known objects [23], using either a single RGB-D camera or multiple synchronized RGB cameras. Interaction cues can also reveal information that helps to reconstruct unknown properties of the object, like the kinematic skeleton [67].

Our current work focuses on estimating hands performing tasks from a single image or video. We use our hand model to generate synthetic training data of hand-object interaction and use deep learning to reconstruct hand-object configurations jointly from a single RGB image. By estimating the 3D hand and object shape together, we are able to reason about interactions such as proximity, contact, grasp stability, and forces while preventing interpenetration.

In collaboration with the Haptic Intelligence department we are extending our hand capture and modeling to account for the soft tissue deformation of the hand during contact and manipulation [0]. This is critical for realistic physical reasoning about grasp.

More information: https://ps.is.mpg.de/project/hands-in-action

Clothing Capture and Animation

Gerard Pons-Moll, Sergi Pujades, Christoph Lassner, Peter Vincent Gehler, Michael Black, Sonny Hu, Chao Zhang

Figure 1.22: (Left, top to bottom) Image-based generative model of people in clothing (ClothNet) [50]; estimating 3D human body shape under clothing (BUFF) [62]; 4D clothing capture for garment modeling, retargeting, and virtual try-on (ClothCap) [15]. (Right) Virtual try-on: a) Scan of a subject wearing clothes; b) image with a new body shape; c) estimated new body shape and pose; d) new body wearing the captured clothes.

While our detailed models of the body capture important aspects of body shape, they are missing something important – clothing. Most people appear in images in clothing and to analyze this we seek models of the body and clothing. Modeling clothing is hard, however, because of the variety of garments, varied topology of clothing, varied appearance, and the complex physical properties of cloth.

Standard methods for clothing 3D bodies rely on 2D patterns and physics simulation. These require expert knowlege, do not work for all types of clothing, and can be difficult to apply to arbitrary body shapes and poses. Consequently, we take a data-driven approach to learn the shape, movement, and appearance of clothing.

ClothNet [50] is a conditional generative model that is directly learned from images of people in clothing. Given a body silhouette, the model produces different people with similar pose and shape in different clothing styles by using a variational autoencoder, followed by an image-to-image translation network that generates the texture of the outfit.

To dress people in 3D, the minimally-clothed body shape is needed. To estimate this from clothed bodies, our BUFF method [62] estimates body shape under clothing from a sequence of 3D scans. In a scan sequence, different poses will make the clothing tight on body in different regions. All frames in a sequence are brought into an unposed canonical space and fused into a single point cloud. We optimize for the body shape to robustly fit the inside of this fused point cloud. This produces a remarkably accurate personalized body shape.

Given the underlying body shape, we can then model how clothing deviates from the body by capturing 3D and 4D scans of clothed people. ClothCap [15] is a pipeline that captures dynamic clothing on humans from 4D scans, segments the clothing from the body, segments it into pieces, and models how the clothing deviates from the body. ClothCap can then retarget the captured clothing to different body shapes paving the way towards virtual clothing try-on.

More information: https://ps.is.mpg.de/project/clothing

Physics of Body Shape and Motion

Michael Black, Sergi Pujades, Meekyoung Kim, Gerard Pons-Moll, Javier Romero, Ludovic Righetti

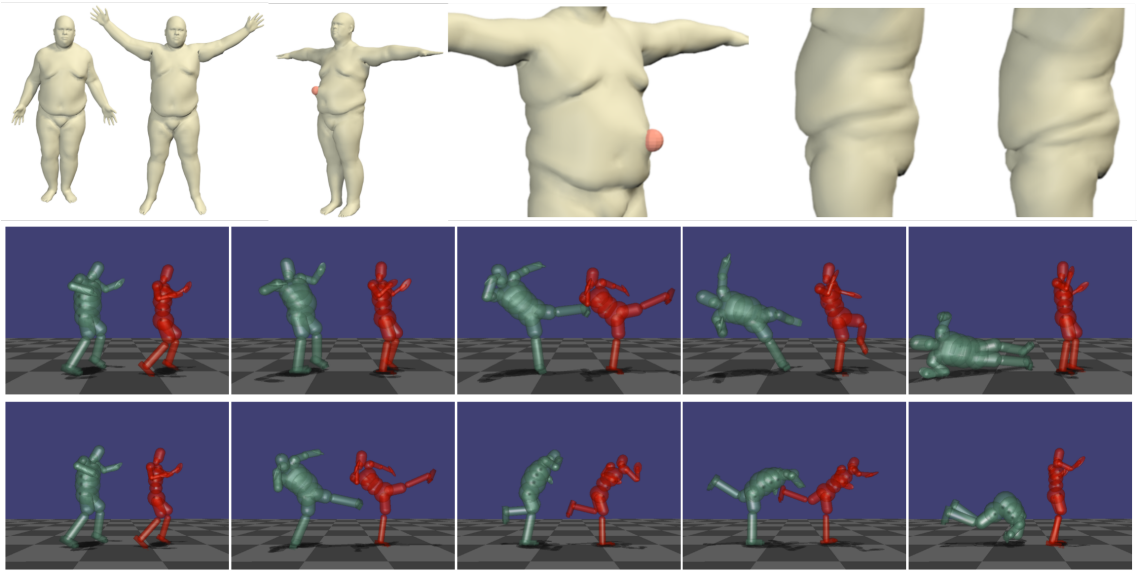


Figure 1.23: Top: We create avatars reproducing the soft tissue motions seen on real humans. We show how such avatars generalize to external forces (applied with the red sphere), and how they deform if gravity is 7 times higher (far right). Bottom: The red avatar performs a motion. A green heavier avatar tries to reproduce the same motion, struggling to raise the leg and loosing balance afterwards.

Humans live in a real world governed by the laws of physics; that is, we apply and exploit forces, such as gravity, in our daily interactions with the world. In this project we allow virtual humans to interact with a virtual world subject to the laws of physics. How would one's body shape deform in case of a collision with an object? How would our walk pattern look like if we weighed a few kilos more?

To model soft-tissue dynamics, we learn a layered volumetric body model from data [16]. To enable this we extend the triangulated mesh of the SMPL body model with a volumetric tetrahedral model called VSMPL. VSMPL contains an inner "rigid" layer and an outer soft-tissue layer. Given 4D sequences of people in motion, we learn the physical properties (Young's modulus) of the outer tetrahedra. We do this such that, when simulated in motion using a finite element method, the surface motion of the VSMPL model resembles the observations. The learned model is a realistic full-body avatar that generalizes to novel motions and external forces.

We also address the problem of retargeting the captured motion of one person onto a different person with a different body shape and physical properties (e.g. taller, heavier, thinner) such that the new morphology is taken into account [6]. We obtain visually plausible simulations using a simplified representation of human body shape that we animate using physically-based retargeting. We develop a novel spacetime optimization approach that learns and robustly adapts physical controllers to new bodies and constraints. The method automatically adapts the motion to a subject with the novel target body shape, respecting the physical properties, and producing an appropriate movement. This makes it easy to create a varied set of motions from a single mocap sequence by simply varying the characters.

More information: https://ps.is.mpg.de/project/physics-shape-motion

Capturing Animal Shape

Silvia Zuffi, Angjoo Kanazawa, Michael Black

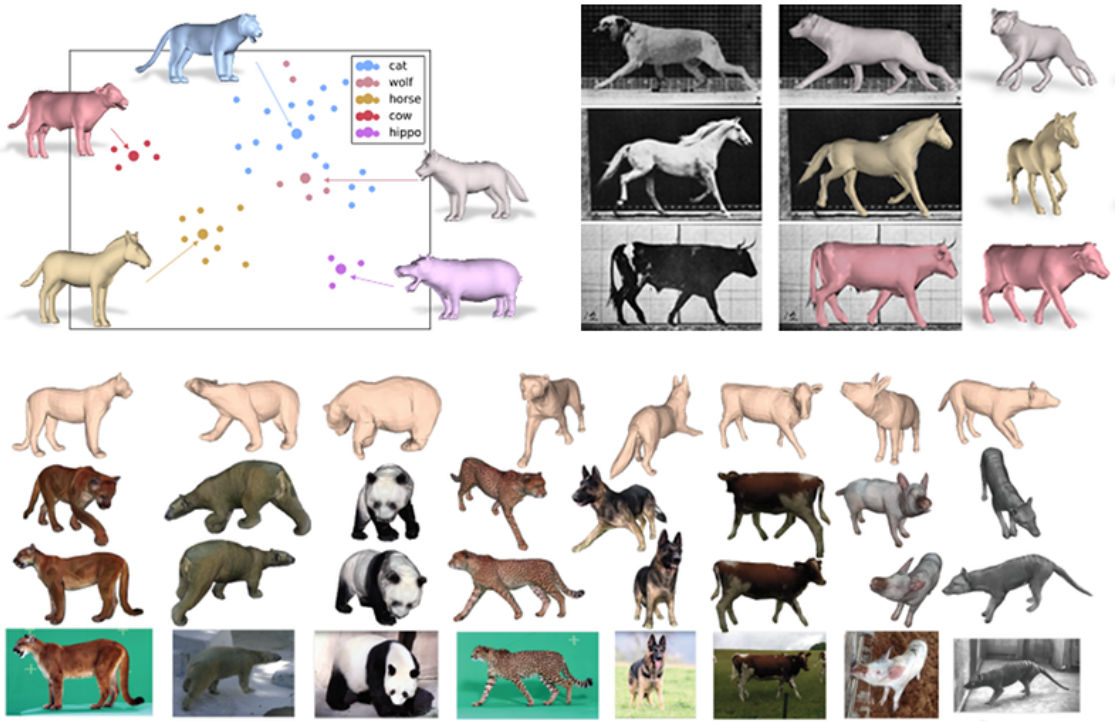


Figure 1.24: We learn an articulated, 3D, statistical shape model of animals (SMAL) that can represent quadrupeds of different species using very little training data (top). We fit SMAL to a set of uncalibrated images estimating pose, shape and vertex displacements and recover 3D textured meshes for a wide range of species (bottom).

In the past 15 years impressive advances have been made in capturing, modeling and tracking the human body. Animals have received much less attention, despite many applications in biomechanics, biology, neuroscience, robotics, and entertainment. The main reason for the lack of 3D animal models is that the methods for modeling the human body cannot be easily applied to animals: animals are not cooperative, cannot be brought to the lab in large numbers, and current scanners cannot be taken into the wild. Additionally they vary significantly in shape and even in the type of body parts they have.

In this project we develop methods to learn 3D articulated statistical shape models that can represent a wide variety of species in the animal kingdom, allowing intra- and inter-species analysis of 3D shape and the automatic and non-invasive assessment of animal shape from images.

From scans of toy animals, we learn the SMAL (Skinned Multi Animal Linear) model [57], a 3D articulated statistical shape model able to represent animal shapes for different species: big cats, dogs, cows, horses, zebras, and hippos. To capture animals outside the SMAL space, we developed SMALR (SMAL with Refinement) [57]. SMALR estimates a detailed 3D textured mesh using a small set of uncalibrated, non-simultaneous images of the animal.

Today animal motion is mostly captured indoors for domestic species with marker-based systems. To address this we are exploiting our 3D articulated animal shape models to develop a markerless motion capture system that will capture the shape and articulated motion of wild animals in their natural environment.

More information: https://ps.is.mpg.de/project/capturing-animal-shape

3D Body Shape and Pose from Images

Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Vincent Gehler, Javier Romero, Michael Black, Gerard Pons-Moll, Martin Kiefel, Yinghao Huang, Ijaz Akhter

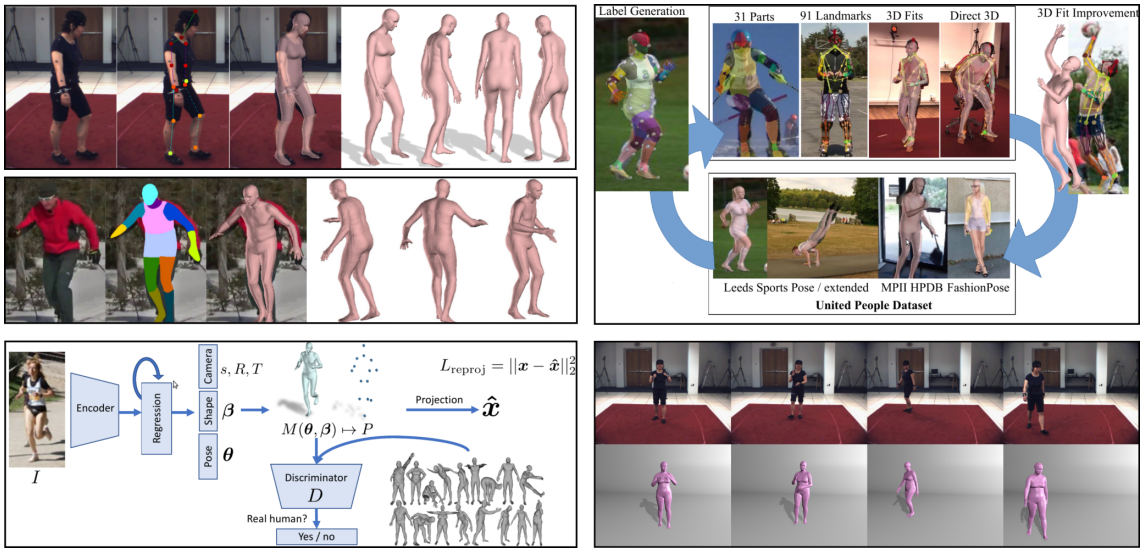


Figure 1.25: Overview of the methods and sample results. Clockwise from upper left: SMPLify [69], Unite the People [51], Multi-view SMPLify (MuVS) [43], Human Mesh Recovery (HMR) [40], Neural Body Fitting (NBF) [30]

Much of the field has focused on estimating 2D joints, 3D joints, or the skeleton of the body. We focus on estimating the full 3D shape and pose. This is crucial for reasoning about interactions. Having the ability to do so from RGB images enables markerless motion capture and provides the foundation for human behavior analysis. We explore two strategies: classical top-down model fitting and feed-forward regression.

SMPLify [69] combines bottom-up 2D feature detection with top-down 3D model fitting. The shape and 3D pose of a person are estimated by minimizing the error between the projected 3D joints of the SMPL model and 2D detected landmarks. Unite the People [51] adds a new loss term, creates a pseudo ground-truth dataset and trains discriminative models for detailed 2D landmark detection and 3D pose estimation. The whole process is repeated multiple times to refine the results and increase the quantity and quality of available data. In [43] the optimization pipeline of SMPLify is extended to handle multi-view imagery and video. Using temporal information helps resolve left/right ambiguities while giving better estimates of global orientation and body shape.

Human Mesh Recovery [40] learns to regresses the shape and the 3D pose directly from a single RGB image, by minimizing the reprojection error of 3D SMPL keypoints during training. This is not sufficient though, so we add an adversarial loss that forces the model to produce SMPL parameters that the discriminator is unable to distinguish from real ones drawn from a database of 3D human meshes. An advantage of this approach is that it can be trained without any expensive paired 2D-to-3D data.

In Neural Body Fitting [30] the shape and 3D pose parameters of SMPL are regressed from body part segmentations given by an intermediate network. Since the whole pipeline is differentiable, different types of supervision can be used, depending on the available information. Extensive experiments show that the body part segmentation is a good intermediate representation for lifting to 3D, as well as that competitive performance can be achieved with limited paired 2D-to-3D data.

More information: https://ps.is.mpg.de/project/3d-pose-and-shape-from-images

Groups and Crowds

Siyu Tang, Michael Black, Peter Vincent Gehler

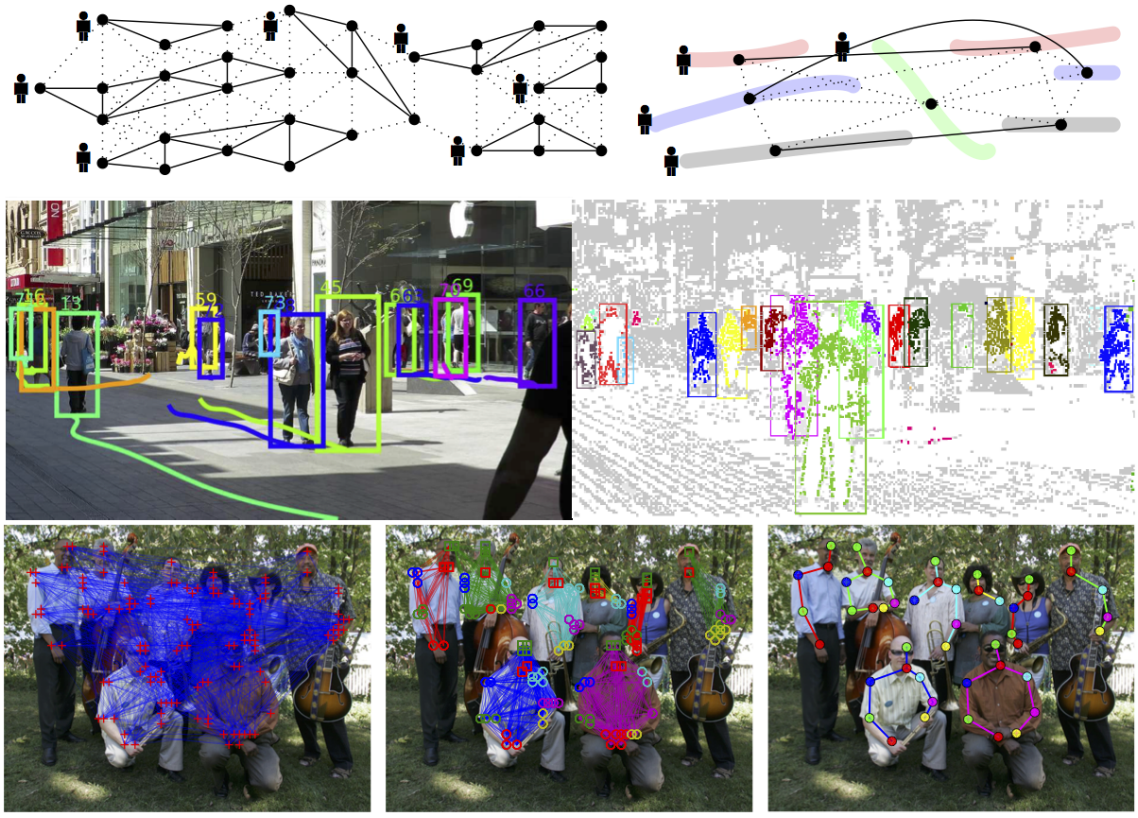


Figure 1.26: Top row illustrates the hierarchical correlation clustering formulation for multi-person tracking [27]. A dotted line indicates that the edge is a cut. The detection graph is partitioned into 7 components, indicating 7 people (top left), which are associated by the global clustering, resulting in 4 persons (top right). Middle row shows qualitative results of tracking and segmentation on the MOT16 benchmark. The solid line under each bounding box indicates the lifetime of the track. Bottom row illustrates the Deepcut model [74] for multi-person pose estimation. Initial detections (bottom left) and pairwise terms between all detections are jointly clustered and each part is labeled corresponding to its part class. Bottom right shows the predicted pose sticks.

People are often a central element of visual scenes. It has been a long-standing goal in computer vision to develop computational models that enable machines to detect crowds of people, analyze their motion and poses, infer their actions and reason about the consequences. Our research addresses a wide range of challenges in visual understanding of people in real-world crowded scenes. These include multi-person tracking [27] [45], multi-person pose estimation [74], segmentation [3] and person re-identification [39].

For multi-target tracking, our work [27] proposed to link, cluster and track targets jointly across space and time. We defined a novel mathematical abstraction for tracking in the form of a minimum cost multicut problem. In order to avoid that distinct but similar looking targets are assigned to the same track, we formulated tracking as a minimum cost lifted multicut problem [45].

Our work [39] presented a novel method to re-identify people in different images, where a second-pooling method is utilized to fuse the feature maps from the pose and the appearance estimator. The method significantly advanced the state-of-the-art on many challenging public benchmarks.

This work forms a foundation for our ongoing work on estimating detailed 3D motions of people in crowded scenes.

More information: https://ps.is.mpg.de/project/groups-and-crowds

AirCap: 3D Motion Capture

Aamir Ahmad, Eric Price, Nitin Saini, Guilherme Lawless, Roman Ludwig, Igor Martinovic, Michael Black

Figure 1.27: Two of our self-designed aerial robots cooperatively detecting and tracking a person on-board in real time (left). Cropped region of interests (ROIs) of images from both MAVs (right). The SMPL mesh with shape and pose estimated using our method is overlaid on all the images in a motion sequence.

Our goal is markless, unconstrained, human and animal motion capture outdoors. To that end, we are developing a flying mocap system using a team of aerial vehicles (MAVs) with only on-board, monocular RGB cameras. To realize such an outdoor motion capture system we need to address research challenges in both control and perception. In a separate ongoing project we solve the control-related challenges, with perception problem in the loop.

The perception functionality of AirCap is split into two phases, namely, i) online data acquisition, and ii) offline pose and shape estimation.

During the online data acquisition phase, the MAVs detect and track the 3D position of a subject while following them. To this end, they perform online and on-board detection using a deep neural network (DNN)-based detector. DNNs often fail at detecting small-scale objects or those that are far away from the camera, which are typical in scenarios with aerial robots. In our solution [7], the mutual world knowledge about the tracked person is jointly acquired by our multi-MAV system during cooperative person tracking. Leveraging this, our method actively selects the relevant region of interest (ROI) in images from each MAV that supplies the highest information content. Our method not only reduces the information loss incurred by down-sampling the high-res images, but also increases the chance of the tracked person being completely in the field of view (FOV) of all MAVs. The data acquired in the online data acquisition phase consists of images captured by all MAVs (see, for example, the left image above) and their estimated camera extrinsic and intrinsic parameters.

In the second phase, which is offline, human pose and shape as a function of time are estimated using only the acquired RGB images and the MAV's self-localization (the camera extrinsics). Using state-of-the-art methods like VNect and HMR, one obtains only a noisy 3D estimate of the human pose. Our approach is to exploit multiple noisy 2D body joint detectors and noisy camera pose information. We then optimize for body shape, body pose, and camera extrinsics by fitting the SMPL body model to the 2D observations. This approach uses a strong body model to take low-level uncertainty into account and results in the first fully autonomous flying mocap system.

More information: https://ps.is.mpg.de/project/aircap

AirCap: Perception-Based Control

Aamir Ahmad, Rahul Tallamraju, Eric Price, Roman Ludwig, Michael Black

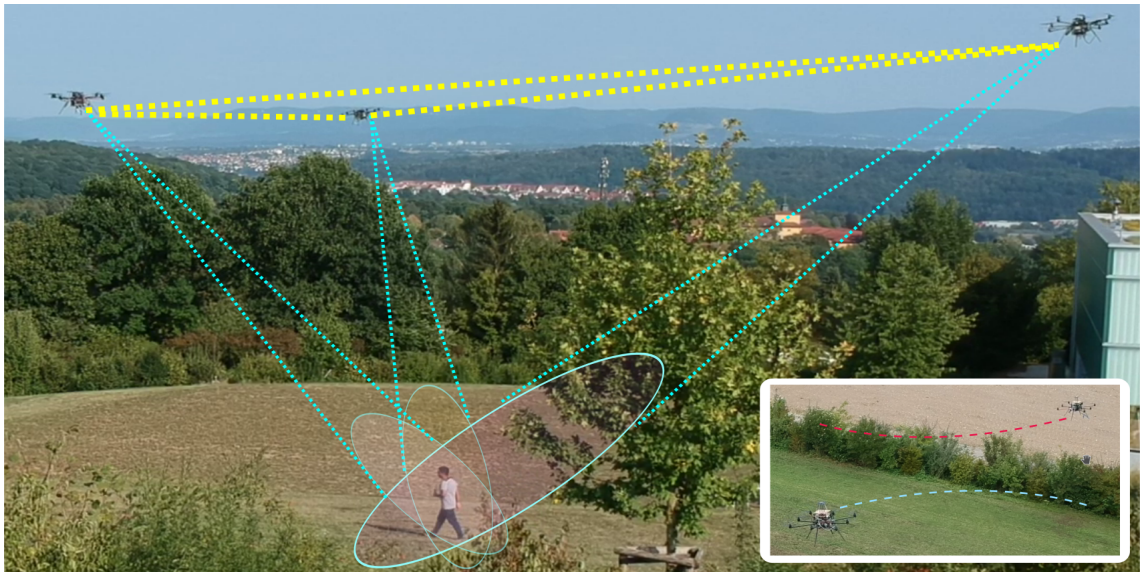


Figure 1.28: Perception-driven formation control of aerial robots tracking a person. The jointly estimated uncertainty in the person's 3D position estimate is minimized while avoiding inter-robot collisions.

Autonomous MoCap systems, like AirCap, rely on robots with on-board cameras that can localize and navigate autonomously. More importantly, these robots must detect, track and follow the subject (human or animal) in real time. Thus, a key component of such a system is motion planning and control of multiple robots that ensures optimal perception of the subject while obeying other constraints, e.g., inter-robot and static obstacle collision avoidance.

Our approach to this formation control problem is based on model predictive control (MPC). An important challenge is to handle collision avoidance as the constraint itself is non-convex and leads to local minima that are not easily identifiable. A possible approach is to treat it as a separate planning module that modifies the MPC-generated optimization trajectory using potential fields. This leads to sub-optimal trajectories and field local minima. In our work [37] we provide a holistic solution to this problem. Instead of directly using repulsive potential field functions to avoid obstacles, we replace them by their exact value at every iteration of the MPC and treat them as external input forces in the system dynamics. Thus, the problem remains convex at every time step. As long as a feasible solution exists for the optimization, obstacle avoidance is guaranteed. Even though field local minima issues remain, they become easier to identify and resolve. To this end, we propose and validate multiple strategies.

In ongoing work we address the complete problem of perception-driven formation control of multiple aerial robots for tracking a human using multiple aerial vehicles. For this, a decentralized convex MPC is developed that generates collision free formation motion plans while minimizing the jointly estimated uncertainty in the tracked person's position estimate. This estimation is performed using a cooperative approach[3] similar to the one developed in our recent work [14]. We validated the real-time efficacy of the proposed algorithm through several field experiments (see image above) with 3 self-designed octocopters and simulation experiments in a realistic outdoor environmental setting with up to 16 robots.

More information: https://ps.is.mpg.de/project/autonomous-mocap

[3] A. Ahmad, E. Ruff, H. Bülthoff. Dynamic baseline stereo vision-based cooperative target tracking. In *19th International Conference on Information Fusion*, pages 1728–1734, 2016.

IMU-based Human Motion Capture Systems

Gerard Pons-Moll, Michael Black, Yinghao Huang, Timo von Marcard, Roberto Henschel, Bodo Rosenhahn, Manuel Kaufmann, Emre Aksan, Otmar Hilliges

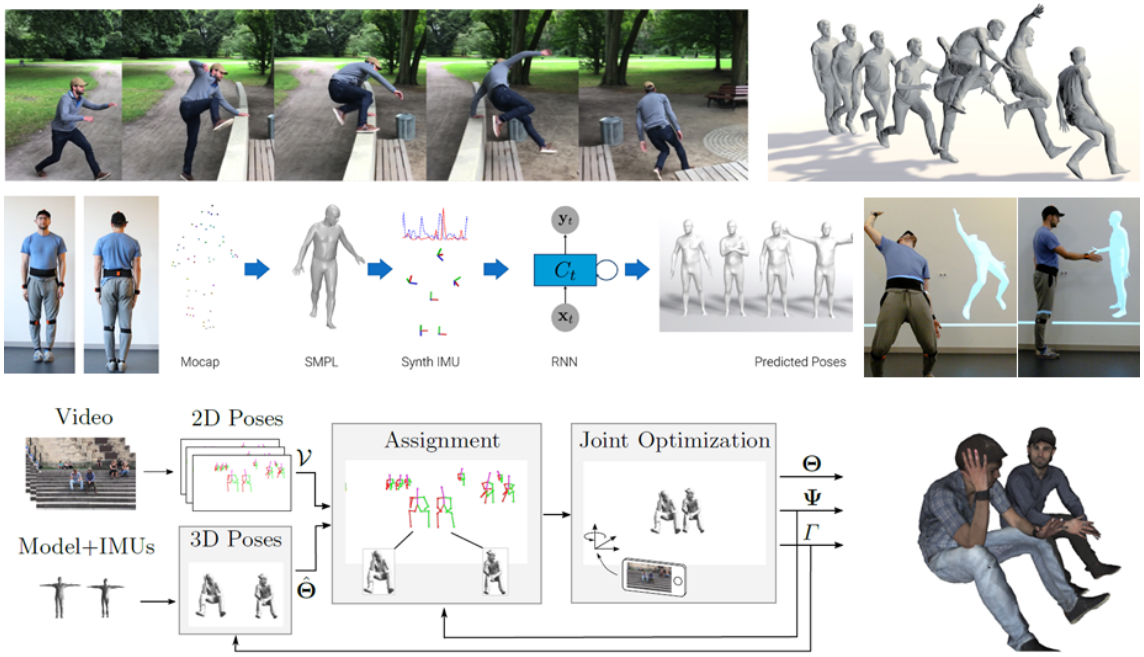


Figure 1.29: Overview. Top row: Unconstrained human motion capture using SIP. Mid row: In DIP, we synthesize an IMU dataset, and leverage that to train an RNN regressor, improving SIP both in accuracy and runtime. Bottom row: Using VIP, we combine videos with sparse IMUs to collect 3DPW, a new dataset of accurate 3D human poses in natural scenes, containing variations in person identity, activity and clothing.

Marker-based optical motion capture (mocap) systems are intrusive and restrict motions to controlled laboratory spaces. Therefore, simple daily activities like biking, or having coffee with friends cannot be recorded with such systems. To address this, and record human motion in everyday natural situations, we develop novel systems based on Inertial Measurement Units (IMUs), that can track the human pose without cameras, making them more suitable for outdoor recordings.

Existing commercial IMU systems require a considerable number of sensors, worn on the body or attached to a suit. These are cumbersome and expensive. To make full-body IMU capture more practical, we developed Sparse Inertial Poser (SIP) [18], which recovers the full 3D human pose from orientation and acceleration measrued by only 6 IMUs attached to the wrists, lower legs, waist and head. This setup is a minimally intrusive solution to capture human activities.

SIP gives an offline, non-intrusive, mocap system that can be used in unconstrained settings of daily life, but the method does not run in real time. In Deep Inertial Poser (DIP) [5], we go beyond the accuracy of SIP and further make it real time. To this end, we synthesize a large IMU dataset from motion capture data and leverage that to learn a deep recurrent regressor that produces SMPL pose parameters in real time from 6 IMU sensor recordings.

While portable, IMU systems are prone to drift. To address this we combine IMUs with a moving camera and current 2D pose-detection methods. Our VIP system [32] solves for the body movements that match the IMU data and project into the image to match 2D joints. Using VIP, we collected the 3DPW dataset, that includes videos of humans in challenging scenes with accurate 3D parameters that will provide the means to quantitatively evaluate monocular methods in difficult scenes.

More information: https://ps.is.mpg.de/project/imu-mocap

Modeling Human Movement

Julieta Martinez, Judith Bütepage, Javier Romero, Hedvig Kjellström, Michael Black, Ludovic Righetti, Partha Ghosh, Sergey Prokudin

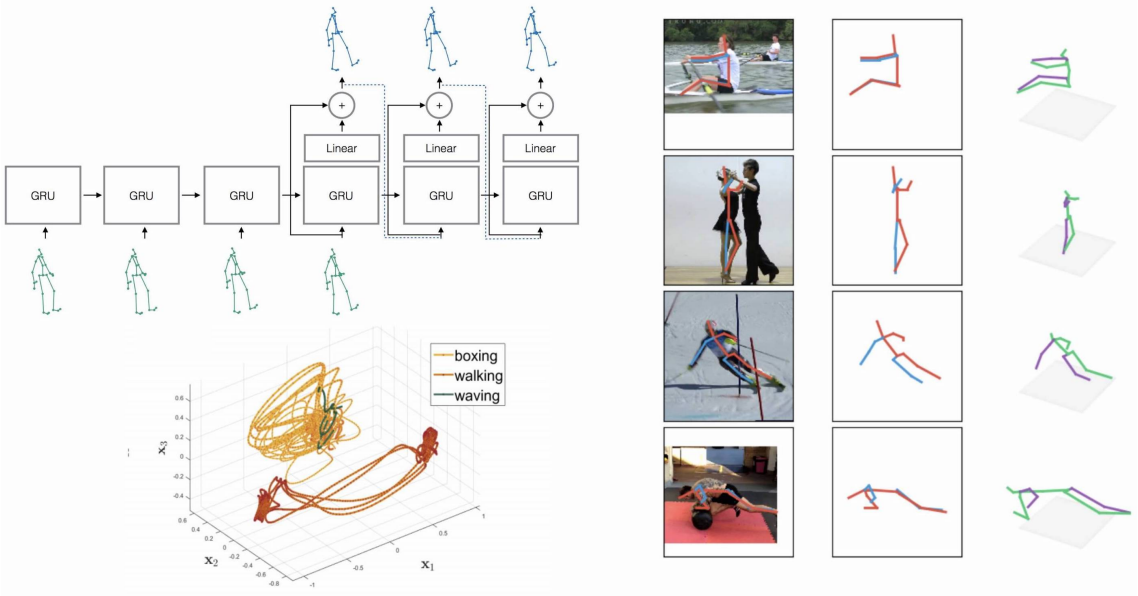


Figure 1.30: Clockwise from upper left: Recurrent neural network for motion prediction [47]; predicting 3D human pose from 2D images [48]; visualization of learned representations of motion sequences [53].

An expressive model of human motion is essential for action classification, motion prediction and synthesis. To that end, we are exploring several deep network architectures to predict human movement.

Current methods for motion prediction typically do not work for a wide range of actions and suffer from "regression to the mean". We show that, surprisingly, state-of-the-art performance can be achieved by a simple baseline that does not model motion at all. We investigate this and propose three changes to the standard RNN models typically used for human motion, which result in a simple and scalable RNN architecture that obtains state-of-the-art performance on human motion prediction [47].

We have also shown that a simple encoder/decoder architecture that takes a set of past poses and predicts a set of future poses works well and is simpler than RNN models. By forcing the encoding through a bottleneck, the approach learns features of human movement that are useful for action recognition. Our feed-forward networks outperform recurrent approaches for short- and long-term predictions and generalize to novel subjects and actions [53].

We have worked on several methods to estimate 3D pose from 2D joints. We show that this can actually be solved with a very simple network that outperforms previous, more complex, methods by a substantial margin. This suggests that "lifting" from 2D to 3D is not the really hard problem but, rather, that extracting the relevant information from the 2D image is the key [48].

Neural networks, however, may not generalize to scenarios that they have never seen – imagine someone floating in zero gravity. Hence we also explore physics-based controllers of human movement [6]. We envision a future that combines the best of both approaches with learned models of behavior combined with physical constraints coming from environmental interaction.

More information: https://ps.is.mpg.de/project/modeling-human-movement

Action and Behavior

Siyu Tang, Yan Zhang

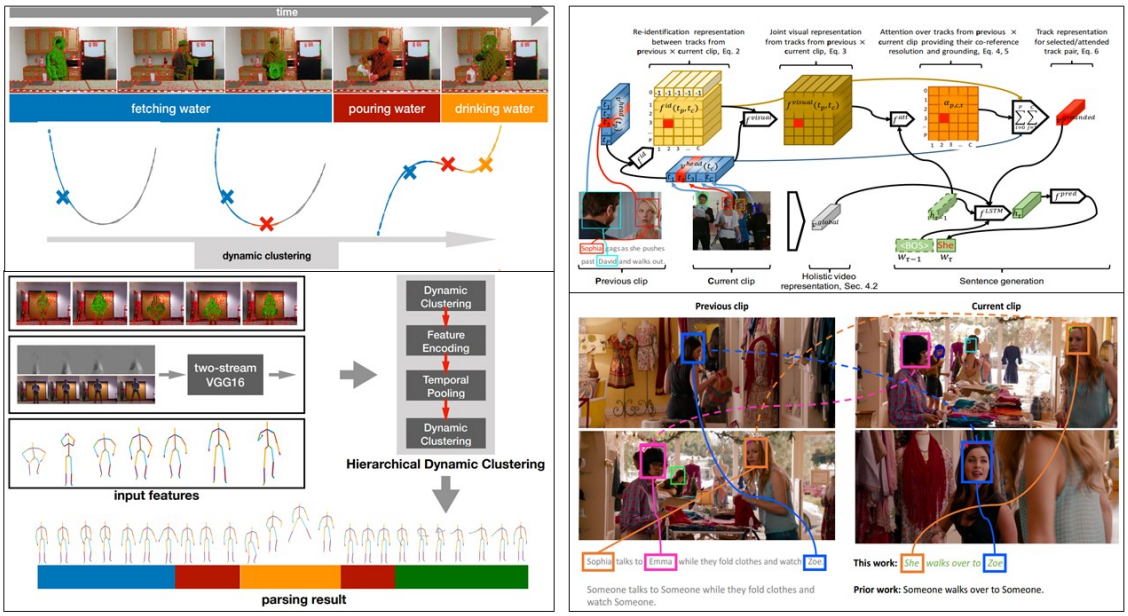


Figure 1.31: Left: A dynamic clustering method for low-level action understanding enables unsupervised human motion parsing [31]. Right: Our method and a result of generated descriptions with grounded and co-referenced people, linking scripts and people in a movie [58].

Human behavior can be described at multiple levels. At the lowest level, we observe the 3D pose of the body over time. Poses can be organized into primitives that capture coordinated activity of different body parts. These further form more complex "actions" or "behaviors". Finally, underlying all of the above are the goals, motives, and emotions of the person; that is, the *cause* of the movement. Our ultimate goal is to extract this high-level causal information from video.

Low-level understanding. Humans can readily differentiate biological motion from non-biological motion. They can do this from sparse visual cues like moving dots and without any explicit supervision. In this spirit, we perform behavior analysis at a low-level using a novel dynamic clustering algorithm that groups actions in an online fashion [31]. As a building block, dynamic clustering is employed in a computational pipeline, where low-level visual cues are aggregated to high-level action patterns via temporal pooling. Our experiments show that this hierarchical dynamic clustering scheme is reliable, generic for diverse input features and fast.

High-level understanding. Here we relate low-level behavior to high-level concepts by identifying individual actors and synthesizing natural-language descriptions of their actions and interactions. We do so using weak supervision provided by scripts associated with the video. As a first attempt, we generate descriptions with grounded and co-referenced people [58]. Specifically, we first learn to localize characters by relating their visual appearance to mentions in the descriptions via a semi-supervised approach. We then provide this (noisy) supervision to a description model, which greatly improves its performance. Our proposed description model improves over prior work w.r.t. generated description quality and additionally provides grounding and local co-reference resolution.

Ongoing work leverages richer models of the human body and its motion as well as richer models of the scene and the objects in it.

More information: https://ps.is.mpg.de/project/action-and-behavior

Learning Optical Flow

Michael Black, Andreas Geiger, Anurag Ranjan, Jonas Wulff, Deqing Sun, Varun Jampani, Laura Sevilla, Joel Janai, Fatma Güney

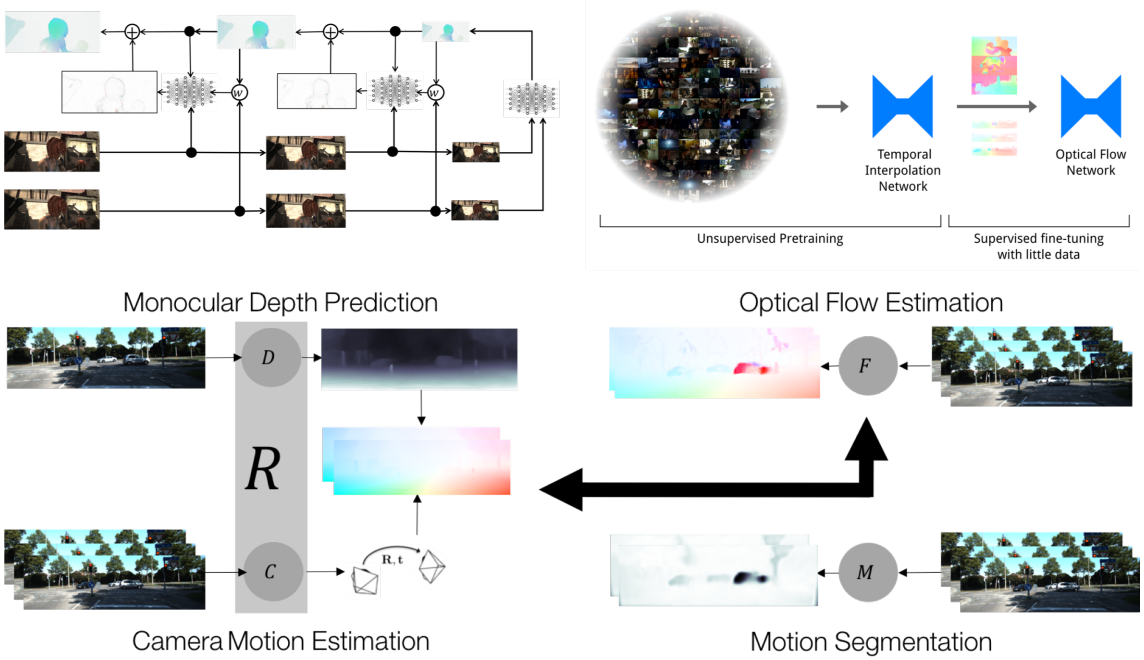


Figure 1.32: Top Left: A Spatial Pyramid Network used for optical flow estimation. Top Right: Optical flow estimates improve when the network is pretrained for temporal interpolation. Bottom: A fully unsupervised approach to solving depth, camera motion, optical flow and motion segmentation.

We view optical flow as the projection of the 3D motion field into the image plane. Until recently, optical flow algorithms were designed by hand and incorporated various heuristics. Deep learning methods provide an opportunity to move away from hand-crafted models but have several limitations. The key one is that they require significant amounts of training data and there are no sensors that give ground truth optical flow for real image sequences.

To deal with large image motions in a compact network, we developed the Spatial Pyramid Networks (SpyNet) [49], which computes optical flow by combining a classical coarse-to-fine flow approach with deep learning. At each level of a spatial pyramid, the deep network computes and update to the current flow estimate. SpyNet is 96% smaller than FlowNet, is very fast, and can be trained end-to-end, making it easy to incorporate into other networks for tasks like action recognition [28].

Synthetic data, used for training most deep flow methods is currently far from realistic. Consequently, we train our IPFlow [29] method on a temporal frame interpolation task using a movie database such that it is encouraged to learn about image motion in complex scenes. We show that this netwok can then be easily fine tuned to compute flow using a small amount of ground truth data.

We go further to address the problem of unsupervised learning. To make this feasible, we build in known geometric information about optical flow in rigid scenes. We introduce the Competitive Collaboration framework [8] and use it to train four different networks that estimate monocular depth, camera pose, optical flow and non-rigid motion segmentation. All of these models compete and collaborate to explain the image sequence. This produces the most accurate unsupervised flow results to date.

Additionally, occlusion boundaries give important information about scene structure and we have worked on learning to detect these [73]. Furthermore, we model occlusions and multiple frames in a video sequence for unsupervised learning of optical flow [33].

More information: https://ps.is.mpg.de/project/learning-optical-flow

Optical Flow and Human Action

Anurag Ranjan, Laura Sevilla, Javier Romero, Yiyi Liao, Fatma Güney, Varun Jampani, Andreas Geiger, Michael Black

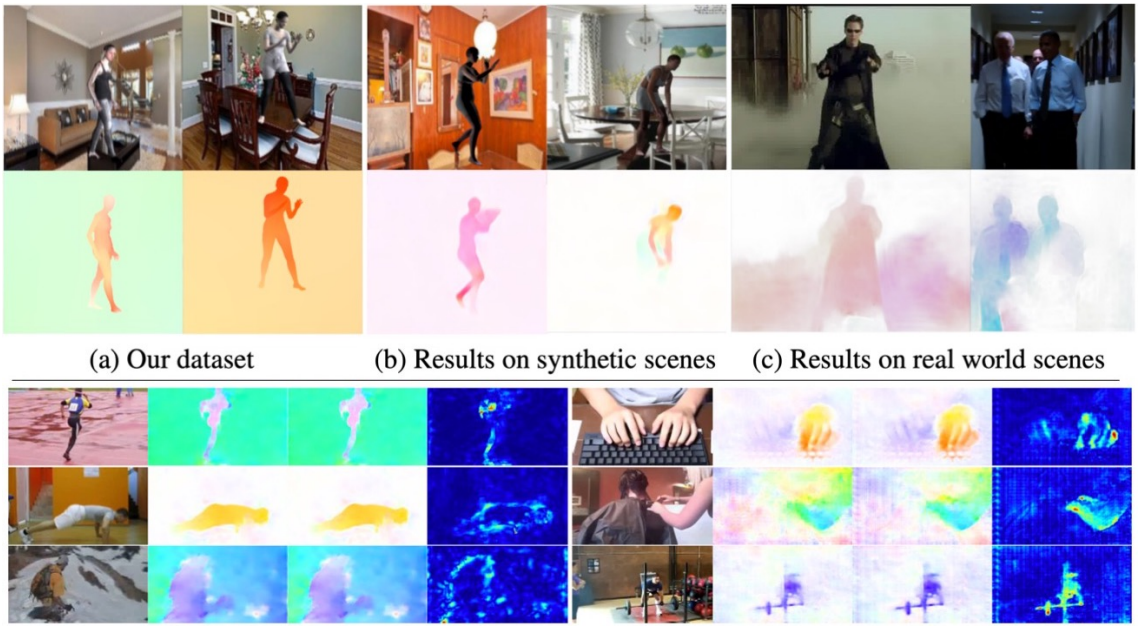


Figure 1.33: Top: We learn human flow from synthetically generated flow fields and find that this generalizes to real videos of human movement. Bottom: We fine tune an optical flow algorithm to produce flow that improves action recognition. Left columns: SpyNet. Right columns: FlowNet. In each set, left to right: first image in sequence, original flow, flow when trained on action recognition, differences in the flow are focused on the human action.

Understanding human action requires modeling and understanding human movement. While we mostly focus on 3D human movement, what is directly observable in videos is the 2D optical flow. Previous work has shown that flow is useful for action recognition and, consequently, we explore how to better estimate human flow and improve action recognition.

Specifically, we trained a neural network to compute human optical flow [36]. To enable this we created a new synthetic training database of image sequences with ground truth human flow. For this we use the 3D SMPL body model and motion capture data to synthesize realistic flow fields; this effectively extends the SURREAL dataset [52]. We then train a convolutional neural network (SpyNet [49] with some modifications) to estimate human flow from pairs of images. The new network is more accurate than a wide range of top methods on held-out test data and generalizes well to real image sequences. When combined with a person detector/tracker, the approach provides a full solution to the problem of 2D human flow estimation.

Most of the top performing action recognition methods use optical flow as a "black box" input. In [28], we take a deeper look at the combination of flow and action recognition, and investigate why optical flow is helpful, what makes a flow method good for action recognition, and how we can make it better. Specifically, we fine tune two neural-network flow methods end-to-end on the UCF101 action recognition dataset. Based on these experiments, we make the following five observations: 1) optical flow is useful for action recognition because it is invariant to appearance, 2) optical flow methods are optimized to minimize end-point-error (EPE), but the EPE of current methods is not well correlated with action recognition performance, 3) flow accuracy at boundaries and for small displacements is most correlated with action recognition performance, 4) training optical flow to minimize classification error instead of EPE improves recognition performance, and 5) optical flow learned for the task of action recognition mostly differs from traditional optical flow inside and at the boundary of the human body.

More information: https://ps.is.tuebingen.mpg.de/research_projects/optical-flow-and-human-action

Image Segmentation and Semantics

Raghudeep Gadde, Varun Jampani, Peter Vincent Gehler, Martin Kiefel, Daniel Kappler, Jun Xie

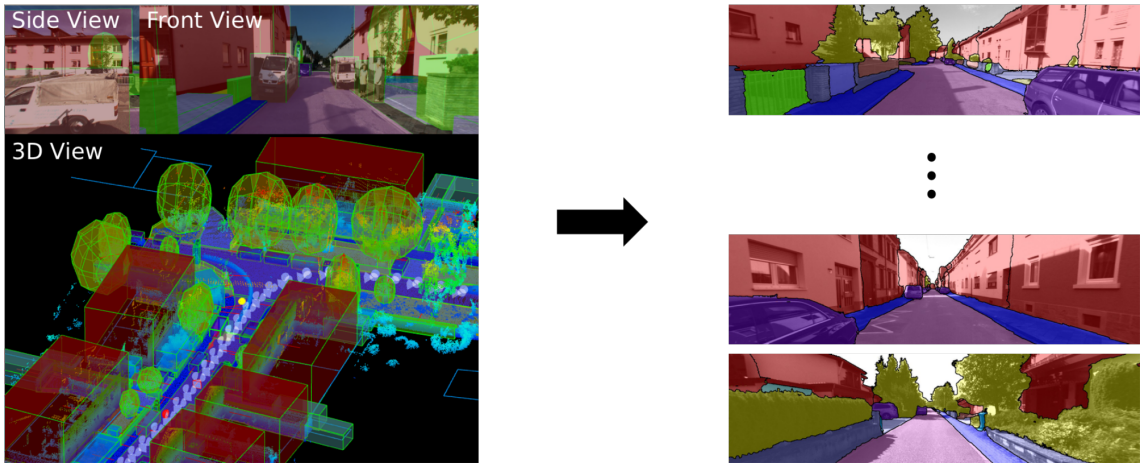


Figure 1.34: 3D to 2D label transfer.

Semantic segmentation is a fundamental problem of computer vision that requires answering what is where in a given image, video or 3D point cloud. The best performing recent techniques require human annotations to obtain ground truth used to train deep neural networks. Such annotation is costly and time consuming to obtain. Consequently, in this project, we address the following two questions:

- How to acquire accurate training data with minimal human cost [70]?
- How to build fast and efficient models for test time inference leveraging the collected data [17], [68]?

In [70], we developed a scalable technique to generate pixelwise annotations for images. For a given 3D reconstructed scene, we annotate static elements in a rough manner and transfer annotations into the image domain using a novel label propagation technique leveraging geometric constraints. We leverage our method to obtain 2D labels for a novel suburban video dataset that we have collected, resulting in 400k semantic and instance image annotations.

In [17], [68] we introduced fast and efficient techniques for semantic segmentation to propagate information using well established Auto-Context and Bilateral filter techniques.

Bilateral filters have wide spread use due to their edge-preserving properties. We generalize the approach to derive a gradient descent algorithm so the filter parameters can be learned from data [75]. This allows us to learn high dimensional linear filters that operate in sparsely populated feature spaces. We build on the permutohedral lattice construction for efficient filtering.

We further introduce a new "bilateral inception" module [68] that can be inserted in existing CNN architectures and performs bilateral filtering, at multiple feature-scales, between superpixels in an image. The feature spaces for bilateral filtering and other parameters of the module are learned end-to-end using standard backpropagation techniques. The bilateral inception module addresses two issues that arise with general CNN segmentation architectures. First, this module propagates information between (super) pixels while respecting image edges, thus using the structured information of the problem for improved results. Second, the layer recovers a full resolution segmentation result from the lower resolution solution of a CNN.

More information: https://ps.is.mpg.de/project/image-segmentation-and-semantics

Multi-View Stereo

Osman Ulusoy, Andreas Geiger, Michael Black

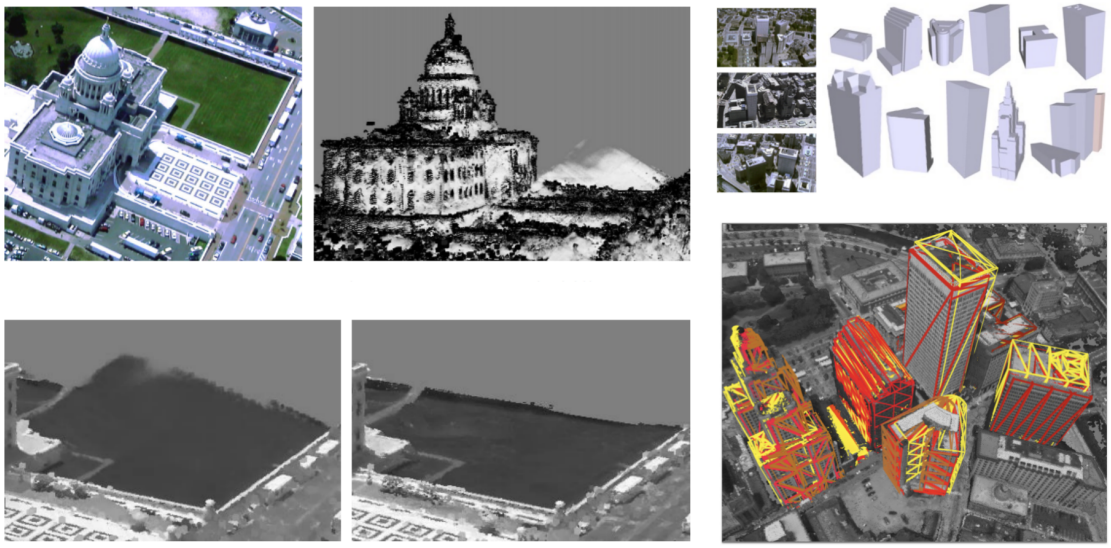


Figure 1.35: In the left: Patches, Planes and Probabilities [71] proposed a planarity prior that regularizes over large distances and helps reconstruct the correct surface. In the Right: Semantic Multi-view Stereo [56] jointly reconstructs a dense 3D model of the entire scene and solves for the existence and pose of each object model.

Dense 3D reconstruction from RGB images is a highly ill-posed problem due to occlusions, textureless or reflective surfaces, varied scene geometry, and spatial discontinuities. We propose algorithms that bring in various types of geometric information that imposes long-range, or semantic, knowlege to address these ambiguities

Our work on Patches, Planes and Probabilities [71] proposed a novel Markov random field model based on ray potentials and a non-local structured prior for volumetric multi-view 3D reconstruction. It was inspired by the planar nature of many elements in man-made environments, i.e., 3D range images of generic scenes can be approximated by piecewise smooth regions with discontinuities at object boundaries. The prior encourages planarity within image segments and regularizes over large voxel neighborhoods. The method was able to resolve reconstruction ambiguities of textureless and partially reflective surfaces and achieved state of-the-art results in reconstruction accuracy for highly challenging aerial datasets.

In our work on Semantic Multi-view Stereo [56], we address ambiguities in 3D reconstruction by presenting a probabilistic approach that integrates object-level shape priors with image-based 3D reconstruction. Our method can infer not only a dense 3D reconstruction of the scene but the existence and precise 3D pose of the objects in it as well. Thus our method not only yields an accurate mapping of the environment but also a semantic understanding in terms of the objects in the environment. The proposed prior allows for powerful regularization that can resolve large ambiguities common in 3D reconstruction. For instance, our shape prior can help reconstruct the back-side of an object even though it is occluded in the images.

More information: https://ps.is.mpg.de/project/multi-view-stereo

Scene Models for Optical Flow

Michael Black, Jonas Wulff, Anurag Ranjan, Laura Sevilla, Fatma Güney, Varun Jampani, Andreas Geiger, Deqing Sun

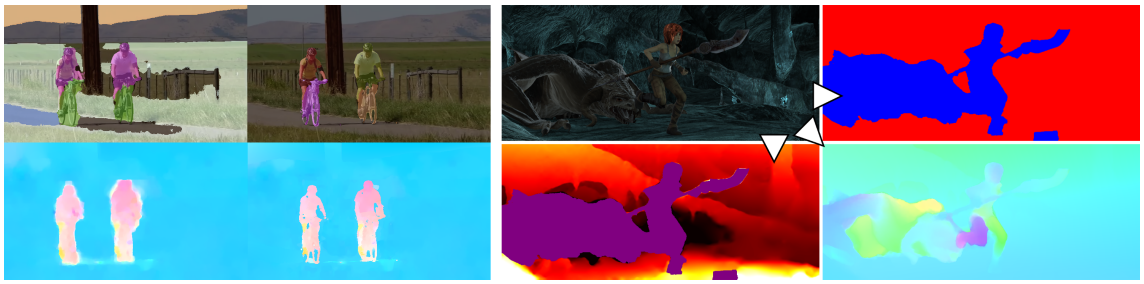


Figure 1.36: Reasoning about the structure of the scene improves optical flow estimation. Semantic segmentation helps to impose meaningful motion priors based on object identity (left). By segmenting the scene into a static background and moving objects an algorithm can use strong geometric constraints in the background region, simplifying the flow problem (right).

Historically, optical flow methods make generic, spatially homogeneous, assumptions about the spatial structure of the 2D image motion. In reality, optical flow varies across an image depending on object class. Simply put, different objects move differently. For rigid objects, the motion is related to the 3D object shape and relative motion. For articulated and non-rigid objects, the motion may be highly stereotyped. Consequently, we should be able to leverage knowledge about objects in the scene, their semantic category, and their geometry, to better estimate optical flow.

We proposed a method for semantic optical flow (SOF) [76] estimation that exploits recent advances in static semantic scene segmentation to segment the image into objects of different types. We define different models of image motion in these regions depending on the type of object. For example, we model the motion on roads with homographies, vegetation with spatially smooth flow, and independently moving objects like cars and planes with affine motion plus deviations. We then pose the flow estimation problem using a novel formulation of localized layers, which addresses limitations of traditional layered models for dealing with complex scene motion. At time of publication, SOF achieved the lowest error of any monocular method in the KITTI-2015 flow benchmark and produces qualitatively better flow and segmentation than recent top methods on a wide range of natural videos.

Furthermore, the optical flow of natural scenes is a combination of the motion of the observer and the independent motion of objects. Existing algorithms typically focus on either recovering motion and structure under the assumption of a purely static world or optical flow for general unconstrained scenes. We combine these approaches in an optical flow algorithm that estimates an explicit segmentation of moving objects using appearance and physical constraints. In static regions, we take advantage of strong constraints to jointly estimate the camera motion and the 3D structure of the scene over multiple frames. This allows us to also regularize the structure instead of the motion. Our formulation uses a Plane+Parallax framework, which works even under small baselines, and reduces the motion estimation to a one-dimensional search problem, resulting in more accurate estimation. In moving regions the flow is treated as unconstrained, and computed with an existing optical flow method. The resulting Mostly-Rigid Flow (MR-Flow) method [46] achieved state-of-the-art results on both the MPISintel and KITTI-2015 benchmarks.

These methods are optimization-based methods that tend to be slow. Furthermore, we manually define constraints, which are often strong simplifications of the real world. To overcome this, we present the Collaborative Competition framework [8], which reasons about the whole scene in a joint, data-driven fashion, and is able to learn to compute the segmentation and the geometry of the scene, and the motion of objects and the background, without explicit supervision.

More information: https://ps.is.mpg.de/project/scene-models-for-optical-flow

Video Segmentation

Varun Jampani, Raghudeep Gadde, Yi-Hsuan Tsai, Michael Black, Peter Vincent Gehler

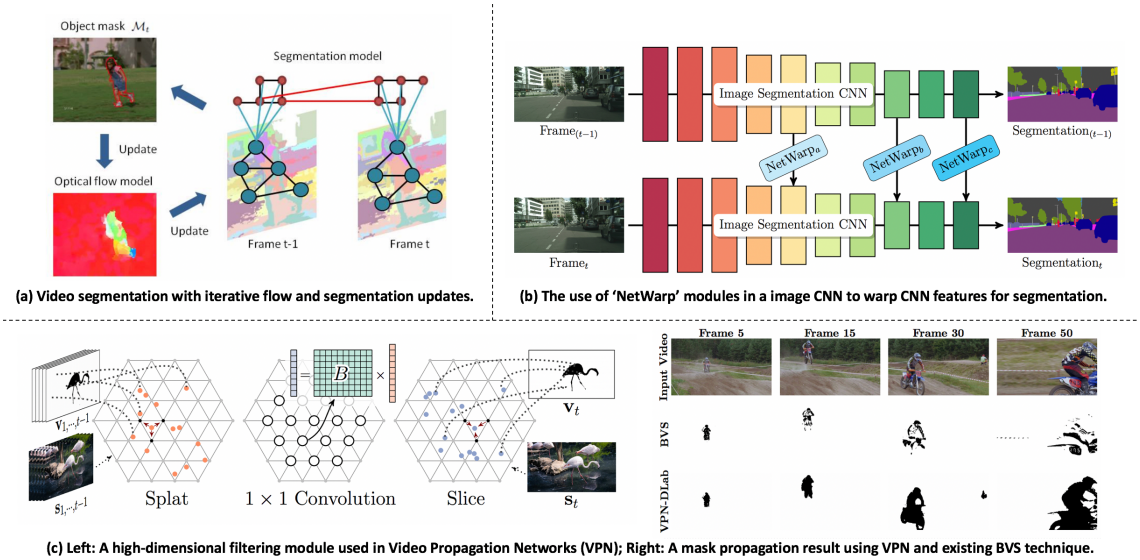


Figure 1.37: Illustration of different video segmentation and propagation techniques: (a) Object Flow [72]. (b) NetWarp [44]. (c) Video Propagation Networks [61].

Videos provide a much richer scene information compared to still images. Despite this, most existing techniques for video segmentation are dominated by per-frame techniques. Video segmentation is a challenging problem due to fast moving objects, deforming shapes and cluttered backgrounds. At Perceiving Systems, we study the use of motion information or pixel correlation that is present across video frames to overcome some of these challenges and obtain better video segmentations.

In [72], we propose an efficient algorithm that considers video segmentation and optical flow estimation simultaneously. We formulate a principled, multiscale, spatio-temporal objective function that uses optical flow to propagate information between frames. For optical flow estimation, we compute the flow independently in the segmented regions and recompose the results. We call the process "object flow" and demonstrate the effectiveness of jointly optimizing optical flow and video segmentation using an iterative scheme.

We also propose one of the first deep neural networks that can be used for general information propagation across video frames. In [61], we project video pixels into a six dimensional XYRGBT space and learn a deep network in this high-dimensional space thereby learning the efficient long-range information propagation across several video frames. Experiments on video object segmentation, video color propagation and semantic video segmentation demonstrate the generality and the effectiveness of our video propagation network.

More recently, we propose a fast and lightweight neural network module called "NetWarp" [44] that can learn to warp intermediate deep feature representations across video frames for better semantic segmentation. Introducing these NetWarp modules in already trained networks and then fine-tuning results in consistent improvements in segmentation accuracy.

More information: https://ps.is.mpg.de/project/video-segmentation

Learning from Synthetic Data

Javier Romero, Anurag Ranjan, Michael Black, Jonas Wulff, David Hoffmann, Dimitris Tzionas, Siyu Tang, Naureen Mahmood, Gül Varol, Xavier Martin, Igor Kalevatykh, Ivan Laptev, Cordelia Schmid

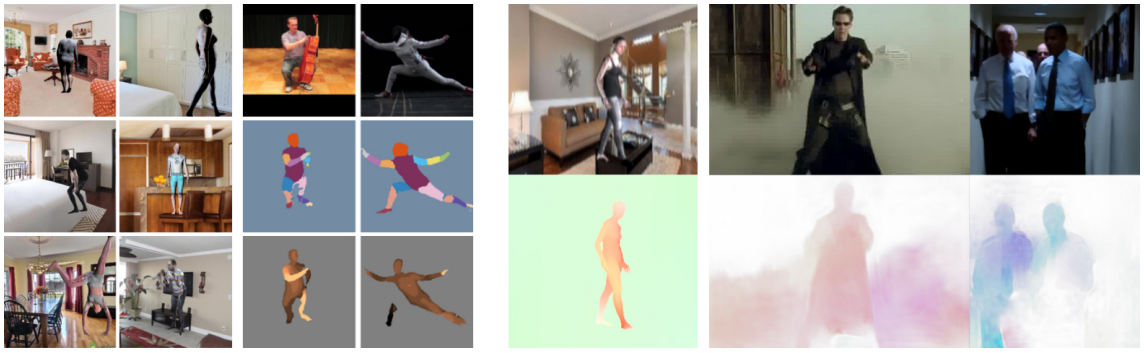


Figure 1.38: Left side: First two columns: SURREAL training data uses SMPL and MoSh to render textured humans in natural poses against random image backgrounds. Right two columns: Training on SURREAL enables the estimation of part segments, depth maps, and more from real images. Right side: Human Flow training data extends SURREAL to moving sequences, giving ground truth optical flow of people in movement (left column). Right two columns: a flow method trained on synthetic data generalizes to human flow in real sequences.

Deep learning has brought rapid progress for many computer vision problems but current methods require large training datasets with annotated ground truth. Human annotators tend to be reasonably efficient for tasks like sparse 2D joint estimation, however annotation for other tasks like dense optical flow estimation or 3D pose estimation is intractable.

To make progress on these tasks, we exploit our 3D body models to generate synthetic training data. In early work, we showed that synthetic data was useful for evaluating optical flow (Middlebury[4] and Sintel[5]). Progress in computer graphics has enabled rendering of synthetic scenes and people and, while not completely realistic, the trends are clear – the quality of such data will steadily improve. Synthetic rendering is appealing for creating training datasets, as it is easily scalable and automatically generates ground truth for a wide variety of problems such as 3D human joints, part segmentations, 3D pose, depth maps, optical flow, body shape, etc.

We focus on learning from synthetic data, using as realistic data as possible about humans, like their motion, body shapes, body textures and backgrounds. We create the SURREAL dataset (Synthetic hUmans foR REAL tasks) and learn deep models for depth estimation and body part segmentation for humans [52]. While not fully realistic, we show that pre-training on this data is valuable and reduces the amount of labeled real data that is needed.

We further create the Human Optical-Flow dataset [36] for learning optical flow of humans in motion. This uses motion capture sequences, processed by MoSh[2] , to produce realistic human optical flow.

Our current work focuses on extending synthetic rendering and inference to multiple people in a single image, for tasks like optical flow, 2D and 3D pose estimation. We further focus on rendering and reconstructing hand-object interactions with realistic hand shapes and poses, object shapes, textures, as well as realistic hand-object grasps. We then plan to extend synthetic data generation to more complex and realistic scenes to reduce the domain gap between real and synthetic data.

More information: https://ps.is.mpg.de/project/learning-from-synthetic-data

[4] S. Baker, D. Scharstein, J. Lewis, S. Roth, M. Black, et al. A Database and Evaluation Methodology for Optical Flow. In *Int. Conf. on Computer Vision, ICCV*, pages 1–8, 2007.

[5] D. J. Butler, J. Wulff, G. B. Stanley, M. J. Black. A naturalistic open source movie for optical flow evaluation. In *European Conf. on Computer Vision (ECCV)*. Part IV, LNCS 7577, pages 611–625, Oct. 2012.

Psychology and Body Shape

Simone Mölbert, Anne Thaler, Betty Mohler, Stephan Streuber, Javier Romero, Naureen Mahmood, Sergi Pujades, Alejandra Quiros-Ramirez, Silvia Zuffi, Joachim Tesch, Michael Black

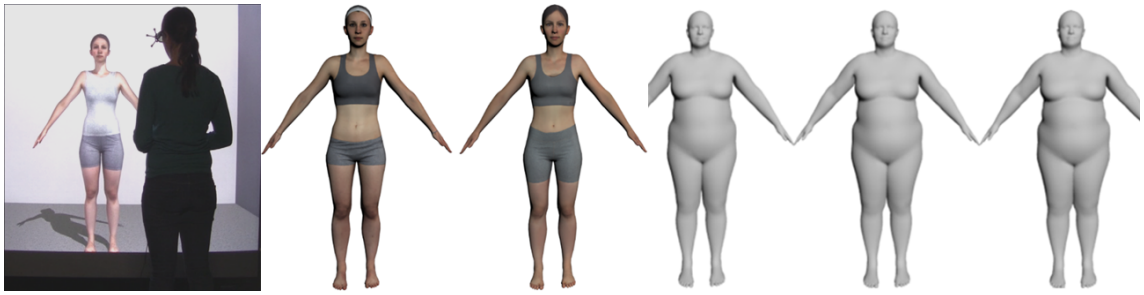


Figure 1.39: Illustration of the virtual reality mirror setup with weight manipulated individual avatars, the manipulation of avatar identity and our figure rating scale (partial view).

Body representation is an essential part of a person's self-concept and also shapes how we see the world. A disturbed body representation also plays a role in clinical conditions such as eating disorders and stroke. So far, a major hurdle for research was the lack of ecologically valid body stimuli. In this project, we cooperate with partners from the Max-Planck-Institute for Biological Cybernetics and the University Hospital Tübingen to develop ecologically valid methods for the assessment of body representation.

For example, we explore how body shape is perceived by people with anorexia nervosa (AN). It was thought that AN patients might perceive their bodies in a distorted way. Using virtual reality and body scans of AN patients we explored this by varying the shape of personal and other avatars to test their perception. We found that anorexics perceive body shape veridically but prefer unhealthy weights.

Based on a 3D body scan and a statistical body model learned from the CAESAR dataset, we generate individual avatars of the participants that can be distorted in terms of weight. Through texture manipulations, we are able to vary the identity of the displayed person. As a major improvement to the existing artist-generated figural drawing scales, we also created a biometric figure rating scale [11] and a desktop tool. In different projects, we assessed >100 participants from the general population as well as >30 women with anorexia nervosa.

Our results in [9] show that in the general population, the accuracy of own body size estimation is predicted by personal BMI, such that participants with lower BMI underestimated their body size and participants with higher BMI overestimated their body size. Critically, these biases suggest that people tend to perceive their weight in an exaggerated way, while there was no hint of a general denial in underweight or overweight persons. The same underestimation bias also occurred in women with anorexia nervosa. Further, we consistently observed that women with anorexia nervosa favored a much thinner body as ideal weight than healthy women. This observation has major clinical implications, because it questions the common idea that misperception of body dimensions may be a maintaining mechanism of this eating disorder. Rather, it suggests that treatment should support patients in accepting a healthy body weight for their own.

Through this work, we have developed a range of body shape modeling, animation, and VR technologies that can be clinically deployed. To study body shape perception we created a virtual reality mirror scenario [10] and our *virtual caliper* allows subjects to create a realistic 3D human avatar using only the controllers of a VR game system. We also studied how we see body shape and describe it with language [19]. We had subjects rate 3D bodies along many dimensions and then built a statistical model relating words and shape. With this "Body Talk" system, we could recover 3D shape from the descriptions of people [21]. This opened up research on body shape and subjective judgements.

More information: https://ps.is.mpg.de/project/anorexia-and-body-shape

Medical Diagnosis

Michael Black, Sergi Pujades, Javier Romero, Nikolas Hesse

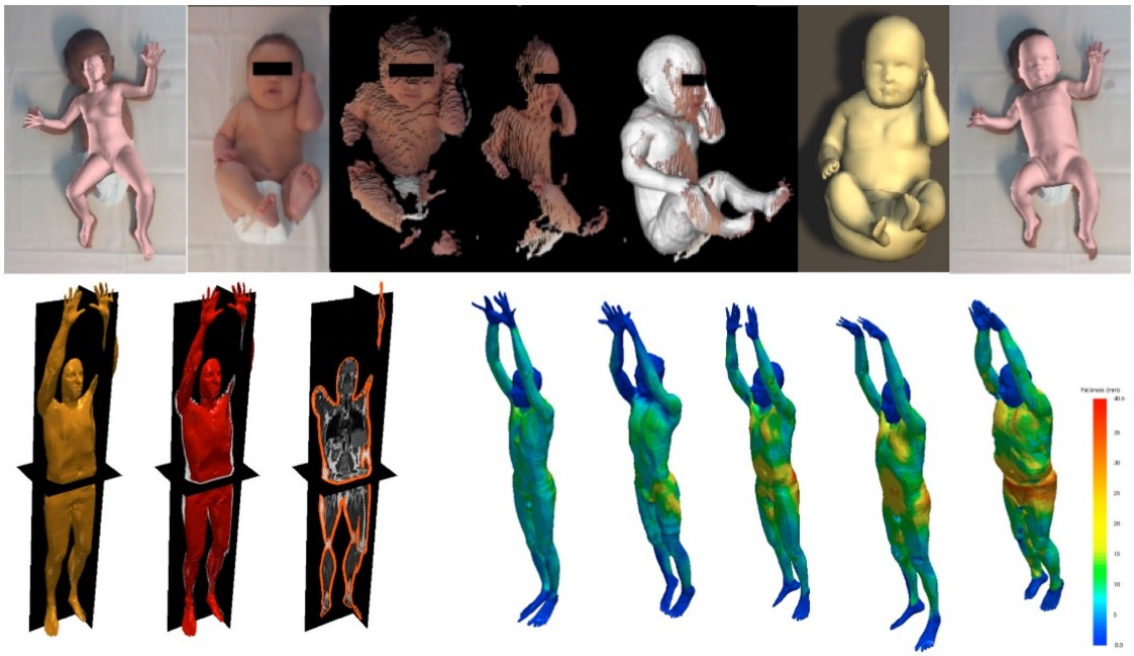


Figure 1.40: Top: Existing adult body models are not suitable for infants, as body proportions differ. We learn an infant body model from RGB-D sequences and use this to recover the shape and pose of freely moving infants. Bottom: We use an adult body model to guide the segmentation of the subcutaneous adipose tissue in a full body MRI (left). The thickness of the adipose tissue is illustrated on the bodies (right).

Body shape and movement are related to human health. For example, our shape tells us about our body fat and our movement tells us something about the health of our motor system. Using our 3D models of body shape we analyze movement and shape to create non-invasive and deployable methods of analyzing human health.

For example, if Cerebral Palsy (CP) is detected early, there are effective therapies to minimize the impact in later life. CP can be diagnosed in infants based on their spontaneous, undirected movements. Unfortunately, this requires expert training that is not widely available. If we can automatically track infant movement, we may automate the early detection of CP. The vision community has made great progress on 3D tracking of adults. Infants have a very different body shape from adults (see figure), which makes it difficult to directly extend prior work to infants. To address this, we learn a model of infant body shape [38] and use it to track 3D movement in RGB-D sequences. Previous models of 3D humans[ref:ps:smpl2015] were learned from thousands of high quality 3D scans, which is not practical with infants. Consequently we developed a novel method that learns infant body shape directly from low quality, incomplete, RGB-D scan sequences and deployed this in hospitals where we scanned over 30 infants.

Another example involves the distribution of adipose tissue in the body. Not all fat is the same. Visceral adipose tissue (around the organs) is highly correlated with diabetes and cardiovascular disease. In contrast, sub-cutaneous adipose tissue (fat under the skin) is relatively benign. Today an analysis of this fat distribution requires an MRI scan to reveal where fat is stored. We are developing methods to estimate this fat distribution purely from the surface shape of the body. To that end, we fit our 3D body models to full-body MRI scans [20] to model both the external surface and the subcutaneous fat layer. We are collecting a dataset of matched MRI data and 3D surface shape and our ongoing work is focused on predicting what is inside solely from the surface.

More information: https://ps.is.mpg.de/project/medical-diagnosis

1.3 Awards & Honors

2018

Michael J. Black, Alumni Research Award, Department of Computer Science, University of British Columbia, 2018

Siyu Tang, DAGM MVTec 2018 Dissertation Award at the German Conference on Pattern Recognition (GCPR) for her thesis "People Detection and Tracking in Crowded Scenes".

Christoph Lassner and **Gerard Pons-Moll**, Best Student Paper at 3DV 2018 for the paper "Neural Body Fitting: Unifying Deep Learning and Model-Based Human Pose and Shape Estimation."

Varun Jampani, Christoph Lassner, Juergen Gall, Yi-Hsuan Tsai, Julietta Martinez, Fatma Güney, Lars Mascheder, Gernot Riegler, Deqing Sun, Outstanding Reviewer Awards (active members and alumni), CVPR 2018.

Alejandra Quiros-Ramirez, Betty Mohler, Michael Black, Simone Mölbert and collaborators, Best Poster Award, Deutsche Gesellschaft für Essstörungen (DGESS), 2018, for the paper "Körper Sprache: Sprachliche Repräsentation von Körpern bei Patientinnen und Patienten mit Essstörungen.

1.3.1 2017

Andreas Geiger and collaborators, Best Student Paper at 3DV for the paper "Sparsity Invariant CNNs."

Siyu Tang, Early Career Research Grant from the University of Tuebingen to start a research group.

Siyu Tang, winner of the CVPR 2017 Multi-Object Tracking Challenge.

Varun Jampani, Osman Ulusoy, and Silvia Zuffi, Outstanding Reviewer Award, CVPR 2017.

Gerard Pons-Moll and Michael Black and collaborators, Best Paper Award, Eurographics 2017, for the paper "Sparse Inertial Poser: Automatic 3D Human Pose Estimation from Sparse IMUs."

1.3.2 2016

Federica Bogo, Javier Romero, Matthew Loper, and Michael Black Dataset Award at the Eurographics Symposium on Geometry Processing 2016. The award encourages and recognizes the importance of the distribution of high-quality datasets on which geometry processing algorithms are tested.

1.3.3 Faculty Appointments

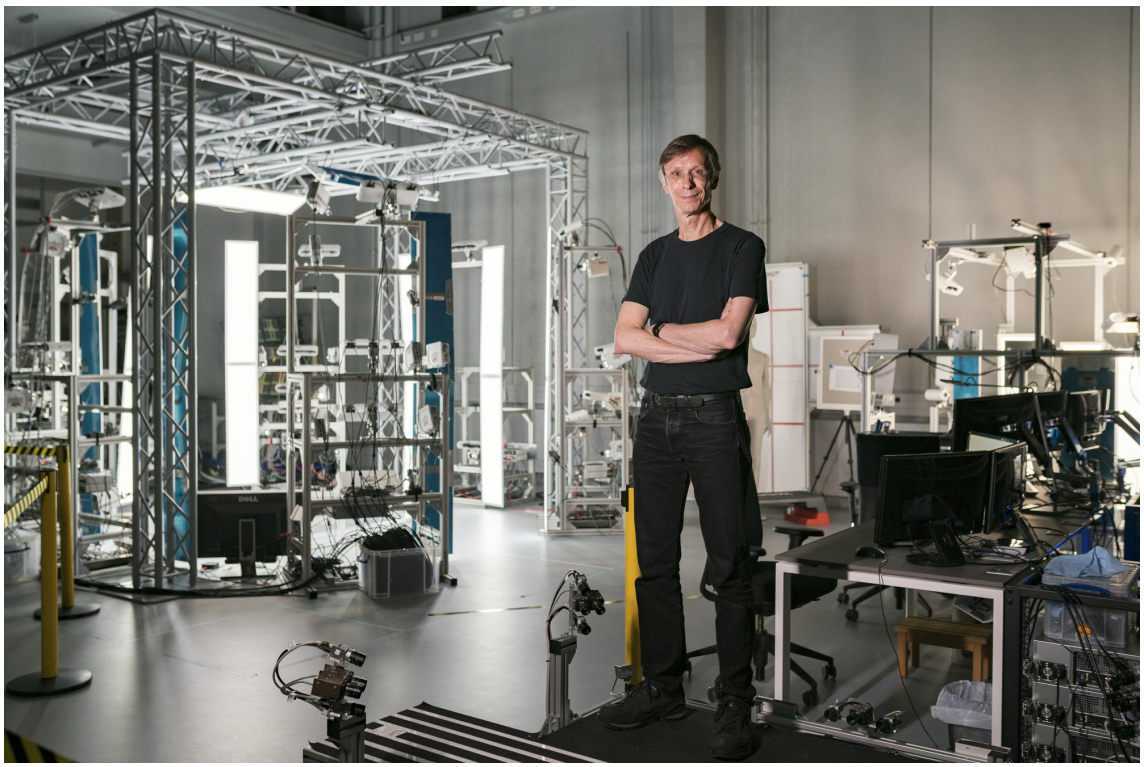
Laura Sevilla appointed Lecturer at the University of Edinburgh, Scotland.

Sergi Pujades appointed Associate Professor at Université Grenoble Alpes, France.

Stefan Streuber appointed Assistant Professor at the University of Konstanz, Germany.

Gerard Pons-Moll appointed independent group leader at the MPI for Informatics, Saarbrücken.

1.4 Director profile: Michael J. Black



Michael J. Black received his B.Sc. in Honours Computer Science from the University of British Columbia (1985), his M.S. in Computer Science from Stanford University (1989), and his Ph.D. in Computer Science from Yale University (1992). As a graduate student he performed research at the NASA Ames Research Center, Aerospace Human Factors Research Division. After one year as an assistant professor at the University of Toronto, he joined the Xerox Palo Alto Research Center in 1993 as a member of research staff. He went on to managed the Image Understanding Area and found the Digital Video Analysis Area. In 2000 he joined the faculty of Brown University in the Department of Computer Science as an Associate Professor with tenure. He was promoted to Full Professor in 2004.

In 2011 he joined the Max Planck Society as a Scientific Member and one of the founding directors of the Max Planck Institute for Intelligent Systems in Tübingen, Germany. Since 2017 he is also an Distinguished Amazon Scholar.

Dr. Black's research spans computer vision, computer graphics, and machine learning. He is most known for his work on optical flow, robust statistical methods, human motion capture and analysis, 3D body shape modeling, neural prosthetics, and motor-cortical decoding.

Dr. Black is a foreign member of the Royal Swedish Academy of Sciences. He is a recipient of the 2010 Koenderink Prize for Fundamental Contributions in Computer Vision and the 2013 Helmholtz Prize for work that has stood the test of time. His work has won several paper awards including the IEEE Computer Society Outstanding Paper Award (CVPR'91) and Honorable Mention for the Marr Prize in 1999 and 2005. His early work on optical flow has been widely used in Hollywood films including for the Academy-Award-winning effects in "What Dreams May Come" and "The Matrix Reloaded." He has contributed to several influential datasets including the Middlebury Flow dataset, HumanEva, and the Sintel dataset. Black has coauthored over 200 peer-reviewed scientific publications.

Dr. Black was a co-founder and member of the board of directors of Body Labs Inc., which commercialized his team's research on 3D human body shape. Body Labs was acquired by Amazon in 2017.

Dr. Michael J. Black

Appointments (2016-2018)

01/2011 – present	Director at the Max Planck Institute for Intelligent Systems
03/2018 – 11/2018	Managing Director of the MPI for Intelligent Systems, Stuttgart and Tübingen
09/2017 – present	Distinguished Amazon Scholar
05/2012 – present	Honorary Professor, Department for Computer Science, University of Tübingen
04/2014 – 04/2016	Visiting Professor, Dept. of Inf. Tech. and Electrical Eng., ETH Zurich
01/2011 – present	Adjunct Professor, Dept. of Computer Science, Brown University

Awards & Honors (Selected)

2018	Alumni Research Award, Dept. Computer Science, Univ. British Columbia
2017	Best Paper Award, Eurographics 2017
2015	Elected foreign member of the Royal Swedish Academy of Sciences
2013	Helmholtz Prize for work that has stood the test of time
2010	Koenderink Prize for Fundamental Contributions in Computer Vision
1999 & 2005	Marr Prize, Honorable Mention, Int. Conf. on Computer Vision, ICCV
1991	IEEE Computer Society, Outstanding Paper Award, CVPR

Selected Organization and Community Service (2016-2018)

2017	Co-organizer, Scenes from Video (SfV) Workshop, III, Lago di Garda, Italy
2016	Chair, Scientific Advisory Board, Computer Science Department, *École Normale Supérieure*, Paris
2016	SIGGRAPH Course, Co-organizer, "Learning human body shapes in motion," Anaheim, CA, 2016
2016	PAMI Young Investigator Award Committee

Selected Memberships (2016–2018)

Royal Swedish Academy of Science, since 2015
European Association for Computer Graphics, since 2017
Intel Network on Intelligent Systems (NIS), since 2017
Association of Computing Machinery (ACM), member since 2014
MPI-ETH Center for Learning Systems, Member since 2015
Werner Reichardt Center for Integrative Neuroscience (CIN), Tübingen University, member since 2011
Institute for Electrical and Electronics Engineers (IEEE): Senior Member since 2008

Startup Activity and Board Memberships (2016 – 2018)

Meshcapade GmbH, Tübingen, angel investor, Nov. 2018
Body Labs Inc., New York, NY, Co-founder, Member of the Board, 2013 – 2017

Selected Keynote, Conference, and Public Talks (2016-2018)

"Estimating Human Motion: Past, Present, and Future." *40 Years DAGM - Invited Talks*, GCPR 2018, Stuttgart, Oct. 2018.
"The Digital Body: Capturing, Modelling and Animating Realistic 3D Humans," *Public Lecture Series on 'What Beings are We?'*, Institute for Art and Architecture, IKA, Vienna, Austria, May 2018.
"Building digital humans by scanning real ones," Keynote, *13th European Conference on Visual Media Production (CVMP)* London, Dec. 12–13, 2016.
"Human body shape modeling: A tutorial," *Invited Tutorial: European Conference on Computer Vision and the ACM Multimedia Conference*, Amsterdam, Oct. 2016.
"The future of generative models: A case study of human bodies in motion," *Int. Computer Vision Summer School, ICVSS*, Sicily, July 2016.
"On building digital humans," *Shape Analysis and Learning by Geometry and Machine*, Inst. for Pure and Applied Mathematics (IPAM), UCLA, Feb. 2016.

1.5 Publications

1.5.1 Journal Articles

2019

[1] S. Pujades, B. Mohler, A. Thaler, J. Tesch, N. Mahmood, N. Hesse, H. H. Bülthoff, M. J. Black. The Virtual Caliper: Rapid Creation of Metrically Accurate Avatars from 3D Measurements. *IEEE Transactions on Visualization and Computer Graphics*, 2019.

2018

[2] A. Thaler, I. Piryankova, J. K. Stefanucci, S. Pujades, S. de la Rosa, S. Streuber, J. Romero, M. J. Black, B. J. Mohler. Visual Perception and Evaluation of Photo-Realistic Self-Avatars From 3D Body Scans in Males and Females. *Frontiers in ICT* **5**: 1–14, 2018.

[3] M. Keuper, S. Tang, B. Andres, T. Brox, B. Schiele. Motion Segmentation & Multiple Object Tracking by Correlation Co-Clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018 (cited on page 20).

[4] Y. Hu, C. J. Parde, M. Q. Hill, N. Mahmood, A. J. O'Toole. First Impressions of Personality Traits From Body Shapes. *Psychological Science* **29** (12): 1969–1983, 2018.

[5] Y. Huang, M. Kaufmann, E. Aksan, M. J. Black, O. Hilliges, G. Pons-Moll. Deep Inertial Poser: Learning to Reconstruct Human Pose from Sparse Inertial Measurements in Real Time. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)* **37**. Two first authors contributed equally: 185:1–185:15, 2018 (cited on pages 7, 23).

[6] M. A. Borno, L. Righetti, M. J. Black, S. L. Delp, E. Fiume, J. Romero. Robust Physics-based Motion Retargeting with Realistic Body Shapes. *Computer Graphics Forum* **37**: 6:1–12, 2018 (cited on pages 17, 24).

[7] E. Price, G. Lawless, R. Ludwig, I. Martinovic, H. H. Buelthoff, M. J. Black, A. Ahmad. Deep Neural Network-based Cooperative Visual Tracking through Multiple Micro Aerial Vehicles. *IEEE Robotics and Automation Letters* **3** (4). Also accepted and presented in the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).: 3193–3200, 2018 (cited on pages 7, 21).

[8] A. Ranjan, V. Jampani, K. Kim, D. Sun, J. Wulff, M. J. Black. Adversarial Collaboration: Joint Unsupervised Learning of Depth, Camera Motion, Optical Flow and Motion Segmentation, 2018 (cited on pages 6, 26, 30).

[9] A. Thaler, M. N. Geuss, S. C. Mölbert, K. E. Giel, S. Streuber, J. Romero, M. J. Black, B. J. Mohler. Body size estimation of self and others in females varying in BMI. *PLoS ONE* **13** (2), 2018 (cited on pages 6, 33).

[10] S. C. Mölbert, A. Thaler, B. J. Mohler, S. Streuber, J. Romero, M. J. Black, S. Zipfel, H.-O. Karnath, K. E. Giel. Assessing body image in anorexia nervosa using biometric self-avatars in virtual reality: Attitudinal components rather than visual body size estimation are distorted. *Psychological Medicine* **48** (4): 642–653, 2018 (cited on pages 6, 33).

2017

[11] S. C. Mölbert, A. Thaler, S. Streuber, M. J. Black, H. Karnath, S. Zipfel, B. Mohler, K. E. Giel. Investigating Body Image Disturbance in Anorexia Nervosa Using Novel Biometric Figure Rating Scales: A Pilot Study. *European Eating Disorders Review* **25** (6): 607–612, 2017 (cited on pages 6, 33).

[12] J. Romero, D. Tzionas, M. J. Black. Embodied Hands: Modeling and Capturing Hands and Bodies Together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*. 245:1–245:17 **36** (6): 245:1–245:17, 2017 (cited on pages 3, 7, 13, 15).

[13] T. Li, T. Bolkart, M. J. Black, H. Li, J. Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics* **36** (6). Two first authors contributed equally: 194:1–194:17, 2017 (cited on pages 3, 7, 13, 14).

[14] A. Ahmad, G. Lawless, P. Lima. An Online Scalable Approach to Unified Multirobot Cooperative Localization and Object Tracking. *IEEE Transactions on Robotics (T-RO)* **33**: 1184–1199, 2017 (cited on pages 7, 22).

[15] G. Pons-Moll, S. Pujades, S. Hu, M. Black. ClothCap: Seamless 4D Clothing Capture and Retargeting. *ACM Transactions on Graphics, (Proc. SIGGRAPH)* **36** (4). Two first authors contributed equally: 73:1–73:15, 2017 (cited on pages 5, 13, 16).

[16] M. Kim, G. Pons-Moll, S. Pujades, S. Bang, J. Kim, M. J. Black, S.-H. Lee. Data-Driven Physics for Human Soft Tissue Animation. *ACM Transactions on Graphics, (Proc. SIGGRAPH)* **36** (4): 54:1–54:12, 2017 (cited on pages 5, 17).

[17] R. Gadde, V. Jampani, R. Marlet, P. Gehler. Efficient 2D and 3D Facade Segmentation using Auto-Context. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017 (cited on page 28).

[18] T. von Marcard, B. Rosenhahn, M. Black, G. Pons-Moll. Sparse Inertial Poser: Automatic 3D Human Pose Estimation from Sparse IMUs. *Computer Graphics Forum 36(2), Proceedings of the 38th Annual Conference of the European Association for Computer Graphics (Eurographics)*: 349–360, 2017 (cited on page 23).

2016

[19] M. Q. Hill, S. Streuber, C. A. Hahn, M. J. Black, A. J. O'Toole. Creating body shapes from verbal descriptions by linking similarity spaces. *Psychological Science* **27** (11): 1486–1497, 2016 (cited on page 33).

[20] S. Y. Yeo, J. Romero, M. Loper, J. Machann, M. Black. Shape estimation of subcutaneous adipose tissue using an articulated statistical shape model. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*: 1–8, 2016. eprint: http://dx.doi.org/10.1080/21681163.2016.1163508 (cited on pages 6, 34).

[21] S. Streuber, M. A. Quiros-Ramirez, M. Q. Hill, C. A. Hahn, S. Zuffi, A. O'Toole, M. J. Black. Body Talk: Crowdshaping Realistic 3D Avatars with Words. *ACM Trans. Graph. (Proc. SIGGRAPH)* **35** (4): 54:1–54:14, 2016 (cited on page 33).

[22] T. Feix, J. Romero, H.-B. Schmiedmayer, A. Dollar, D. Kragic. The GRASP Taxonomy of Human Grasp Types. *Human-Machine Systems, IEEE Transactions on* **46** (1): 66–77, 2016 (cited on page 15).

[23] D. Tzionas, L. Ballan, A. Srikantha, P. Aponte, M. Pollefeys, J. Gall. Capturing Hands in Action using Discriminative Salient Points and Physics Simulation. *International Journal of Computer Vision (IJCV)* **118** (2): 172–193, 2016 (cited on page 15).

[24] T. von Marcard, G. Pons-Moll, B. Rosenhahn. Human Pose Estimation from Video and IMUs. *Transactions on Pattern Analysis and Machine Intelligence PAMI* **38** (8): 1533–1547, Jan. 2016.

[25] M. A. Brubaker, A. Geiger, R. Urtasun. Map-Based Probabilistic Visual Self-Localization. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 2016.

1.5.2 Conference Papers

2019

[26] P. Ghosh, A. Losalka, M. J. Black. Resisting Adversarial Attacks using Gaussian Mixture Variational Autoencoders. In *Proc. AAAI*, 2019.

2018

[27] L. Ma, S. Tang, M. J. Black, L. V. Gool. Customized Multi-Person Tracker. In *Computer Vision – ACCV 2018*, 2018 (cited on page 20).

[28] L. Sevilla-Lara, Y. Liao, F. Güney, V. Jampani, A. Geiger, M. J. Black. On the Integration of Optical Flow and Action Recognition. In *German Conference on Pattern Recognition (GCPR)*. Vol. LNCS 11269, pages 281–297, Oct. 2018 (cited on pages 26, 27).

[29] J. Wulff, M. J. Black. Temporal Interpolation as an Unsupervised Pretraining Task for Optical Flow Estimation. In *German Conference on Pattern Recognition (GCPR)*. Vol. LNCS 11269, pages 567–582, Oct. 2018 (cited on page 26).

[30] M. Omran, C. Lassner, G. Pons-Moll, P. V. Gehler, B. Schiele. Neural Body Fitting: Unifying Deep Learning and Model-Based Human Pose and Shape Estimation. In *3DV*, 2018 (cited on pages 3, 19).

[31] Y. Zhang, S. Tang, H. Sun, H. Neumann. Human Motion Parsing by Hierarchical Dynamic Clustering. In *Proceedings of the British Machine Vision Conference (BMVC)*, page 269, 2018 (cited on page 25).

[32] T. von Marcard, R. Henschel, M. J. Black, B. Rosenhahn, G. Pons-Moll. Recovering Accurate 3D Human Pose in The Wild Using IMUs and a Moving Camera. In *European Conference on Computer Vision (ECCV)*. Vol. Lecture Notes in Computer Science, vol 11214, pages 614–631, 2018 (cited on pages 7, 23).

[33] J. Janai, F. Güney, A. Ranjan, M. J. Black, A. Geiger. Unsupervised Learning of Multi-Frame Optical Flow with Occlusions. In *European Conference on Computer Vision (ECCV)*. Vol. Lecture Notes in Computer Science, vol 11220, pages 713–731, 2018 (cited on page 26).

[34] A. Ranjan, T. Bolkart, S. Sanyal, M. J. Black. Generating 3D Faces using Convolutional Mesh Autoencoders. In *European Conference on Computer Vision (ECCV)*. Vol. Lecture Notes in Computer Science, vol 11207, pages 725–741, 2018 (cited on pages 7, 14).

[35] S. Prokudin, P. Gehler, S. Nowozin. Deep Directional Statistics: Pose Estimation with Uncertainty Quantification. In *European Conference on Computer Vision (ECCV)*, 2018.

[36] A. Ranjan, J. Romero, M. J. Black. Learning Human Optical Flow. In *29th British Machine Vision Conference*, 2018 (cited on pages 3, 27, 32).

[37] R. Tallamraju, S. Rajappa, M. J. Black, K. Karlapalem, A. Ahmad. Decentralized MPC based Obstacle Avoidance for Multi-Robot Target Tracking Scenarios. In *2018 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*, pages 1–8, 2018 (cited on page 22).

[38] N. Hesse, S. Pujades, J. Romero, M. J. Black, C. Bodensteiner, M. Arens, U. G. Hofmann, U. Tacke, M. Hadders-Algra, R. Weinberger, W. Muller-Felber, A. S. Schroeder. Learning an Infant Body Model from RGB-D Data for Accurate Full Body Motion Analysis. In *Int. Conf. on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2018 (cited on pages 6, 7, 34).

[39] Y. Suh, J. Wang, S. Tang, T. Mei, K. M. Lee. Part-Aligned Bilinear Representations for Person Re-identification. In *European Conference on Computer Vision (ECCV)*. Vol. 11218, pages 418–437, 2018 (cited on page 20).

[40] A. Kanazawa, M. J. Black, D. W. Jacobs, J. Malik. End-to-end Recovery of Human Shape and Pose. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018 (cited on pages 3, 19).

[41] S. Zuffi, A. Kanazawa, M. J. Black. Lions and Tigers and Bears: Capturing Non-Rigid, 3D, Articulated Shape from Images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018 (cited on page 7).

2017

[42] S. Kenny, N. Mahmood, C. Honda, M. J. Black, N. F. Troje. Effects of animation retargeting on perceived action outcomes. In *Proceedings of the ACM Symposium on Applied Perception (SAP'17)*, 2:1–2:7, 2017.

[43] Y. Huang, F. Bogo, C. Lassner, A. Kanazawa, P. V. Gehler, J. Romero, I. Akhter, M. J. Black. Towards Accurate Marker-less Human Shape and Pose Estimation over Time. In *International Conference on 3D Vision (3DV)*, pages 421–430, 2017 (cited on page 19).

[44] R. Gadde, V. Jampani, P. V. Gehler. Semantic Video CNNs through Representation Warping. In *Proceedings IEEE International Conference on Computer Vision (ICCV)*, 2017 (cited on page 31).

[45] S. Tang, M. Andriluka, B. Andres, B. Schiele. Multiple People Tracking by Lifted Multicut and Person Re-identification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3701–3710, 2017 (cited on page 20).

[46] J. Wulff, L. Sevilla-Lara, M. J. Black. Optical Flow in Mostly Rigid Scenes. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017*, pages 6911–6920, 2017 (cited on pages 5, 30).

[47] J. Martinez, M. J. Black, J. Romero. On human motion prediction using recurrent neural networks. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017*, 2017 (cited on page 24).

[48] J. Martinez, R. Hossain, J. Romero, J. J. Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017 (cited on page 24).

[49] A. Ranjan, M. Black. Optical Flow Estimation using a Spatial Pyramid Network. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017*, 2017 (cited on pages 26, 27).

[50] C. Lassner, G. Pons-Moll, P. V. Gehler. A Generative Model of People in Clothing. In *Proceedings IEEE International Conference on Computer Vision (ICCV)*, 2017 (cited on page 16).

[51] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, P. V. Gehler. Unite the People: Closing the Loop Between 3D and 2D Human Representations. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017*, 2017 (cited on pages 3, 7, 19).

[52] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, C. Schmid. Learning from Synthetic Humans. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017*, 2017 (cited on pages 3, 7, 27, 32).

[53] J. Bütepage, M. Black, D. Kragic, H. Kjellström. Deep representation learning for human motion prediction and classification. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017*, 2017 (cited on page 24).

[54] J. Janai, F. Güney, J. Wulff, M. Black, A. Geiger. Slow Flow: Exploiting High-Speed Cameras for Accurate and Diverse Optical Flow Reference Data. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017*, pages 1406–1416, 2017 (cited on page 7).

[55] F. Bogo, J. Romero, G. Pons-Moll, M. J. Black. Dynamic FAUST: Registering Human Bodies in Motion. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017*, 2017 (cited on pages 7, 13).

[56] A. O. Ulusoy, M. J. Black, A. Geiger. Semantic Multi-view Stereo: Jointly Estimating Objects and Voxels. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017*, 2017 (cited on page 29).

[57] S. Zuffi, A. Kanazawa, D. Jacobs, M. J. Black. 3D Menagerie: Modeling the 3D Shape and Pose of Animals. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017*, pages 5524–5532, 2017 (cited on pages 7, 18).

[58] A. Rohrbach, M. Rohrbach, S. Tang, S. J. Oh, B. Schiele. Generating Descriptions with Grounded and Co-Referenced People. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4196–4206, 2017 (cited on page 25).

[59] E. Levinkov, J. Uhrig, S. Tang, M. Omran, E. Insafutdinov, A. Kirillov, C. Rother, T. Brox, B. Schiele, B. Andres. Joint Graph Decomposition and Node Labeling by Local Search. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1904–1912, 2017.

[60] T. Nestmeyer, P. V. Gehler. Reflectance Adaptive Filtering Improves Intrinsic Image Estimation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1771–1780, 2017.

[61] V. Jampani, R. Gadde, P. V. Gehler. Video Propagation Networks. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017*, 2017 (cited on page 31).

[62] C. Zhang, S. Pujades, M. Black, G. Pons-Moll. Detailed, accurate, human shape estimation from clothed 3D scan sequences. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Spotlight, 2017 (cited on pages 7, 13, 16).

[63] G. Riegler, O. Ulusoy, A. Geiger. OctNet: Learning Deep 3D Representations at High Resolutions. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017*, 2017.

[64] L. Balles, J. Romero, P. Hennig. Coupling Adaptive Batch Sizes with Learning Rates. In *Proceedings Conference on Uncertainty in Artificial Intelligence (UAI) 2017*, pages 410–419, 2017.

2016

[65] C. Lassner, D. Kappler, M. Kiefel, P. Gehler. Barrista - Caffe Well-Served. In *ACM Multimedia Open Source Software Competition*, 2016.

[66] F. Güney, A. Geiger. Deep Discrete Flow. In *Asian Conference on Computer Vision (ACCV)*, 2016.

[67] D. Tzionas, J. Gall. Reconstructing Articulated Rigged Models from RGB-D Videos. In *European Conference on Computer Vision Workshops 2016 (ECCVW'16) - Workshop on Recovering 6D Object Pose (R6D'16)*, pages 620–633, 2016 (cited on page 15).

[68] R. Gadde, V. Jampani, M. Kiefel, D. Kappler, P. Gehler. Superpixel Convolutional Networks using Bilateral Inceptions. In *European Conference on Computer Vision (ECCV)*. Lecture Notes in Computer Science, 2016 (cited on page 28).

[69] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, M. J. Black. Keep it SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image. In *Computer Vision – ECCV 2016*. Lecture Notes in Computer Science, pages 561–578, 2016 (cited on pages 3, 7, 19).

[70] J. Xie, M. Kiefel, M.-T. Sun, A. Geiger. Semantic Instance Annotation of Street Scenes by 3D to 2D Label Transfer. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016 (cited on page 28).

[71] A. O. Ulusoy, M. J. Black, A. Geiger. Patches, Planes and Probabilities: A Non-local Prior for Volumetric 3D Reconstruction. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016 (cited on page 29).

[72] Y.-H. Tsai, M.-H. Yang, M. J. Black. Video segmentation via object flow. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016 (cited on page 31).

[73] H. Fu, C. Wang, D. Tao, M. J. Black. Occlusion boundary detection via deep exploration of context. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016 (cited on page 26).

[74] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. Gehler, B. Schiele. DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4929–4937, 2016 (cited on page 20).

[75] V. Jampani, M. Kiefel, P. V. Gehler. Learning Sparse High Dimensional Filters: Image Filtering, Dense CRFs and Bilateral Neural Networks. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 4452–4461, 2016 (cited on page 28).

[76] L. Sevilla-Lara, D. Sun, V. Jampani, M. J. Black. Optical Flow with Semantic Segmentation and Localized Layers. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 3889–3898, 2016 (cited on pages 5, 30).

[77] R. Fleming, B. Mohler, J. Romero, M. J. Black, M. Breidt. Appealing female avatars from 3D body scans: Perceptual effects of stylization. In *11th Int. Conf. on Computer Graphics Theory and Applications (GRAPP)*, 2016.

1.5.3 Patents

2018

[78] M. Black, D. Hirshberg, M. Loper, E. Rachlin, A. Weiss. Co-Registration – Simultaneous Alignment and Modeling of Articulated 3D Shapes. U.S. Patent 9,898,848. Feb. 2018.

[79] M. J. Black, A. Balan, A. Weiss, L. Sigal, M. Loper, T. St Clair. Method and Apparatus for Estimating Body Shape. U.S. Patent 10,002,460. June 2018.

2017

[80] M. Loper, N. Mahmood, M. Black. Method for providing a three dimensional body model. U.S. Patent 9,710,964 B2. July 2017.

[81] M. J. Black, P. Guan. System and method for simulating realistic clothing. U.S. Patent 9,679,409 B2. June 2017.

[82] M. J. Black, O. Freifeld, A. Weiss, M. Loper, P. Guan. Parameterized Model of 2D Articulated Human Shape. U.S. Patent 9,761,060. Sept. 2017.

1.5.4 Book Chapters

2017

[83] T. Nestmeyer, P. Robuffo Giordano, H. H. Bülthoff, A. Franchi. Decentralized Simultaneous Multi-target Exploration using a Connected Network of Multiple Robots. In. Autonomous Robots, pages 989–1011, 2017.

[84] R. Fleming, B. J. Mohler, J. Romero, M. J. Black, M. Breidt. Appealing Avatars from 3D Body Scans: Perceptual Effects of Stylization. In. *Computer Vision, Imaging and Computer Graphics Theory and Applications: 11th International Joint Conference, VISIGRAPP 2016, Rome, Italy, February 27 – 29, 2016, Revised Selected Papers*. Springer International Publishing, pages 175–196, 2017.

[85] S. Prokudin, D. Kappler, S. Nowozin, P. Gehler. Learning to Filter Object Detections. In. *Pattern Recognition: 39th German Conference, GCPR 2017, Basel, Switzerland, September 12–15, 2017, Proceedings*. Springer International Publishing, Cham, pages 52–62, 2017.

1.5.5 Theses

PhD Theses

[86] J. Wulff. Model-based Optical Flow: Layers, Learning, and Geometry. PhD thesis. Tuebingen University, 2018.

[87] V. Jampani. Learning Inference Models for Computer Vision. PhD thesis. MPI for Intelligent Systems and University of Tübingen, 2017.

[88] M. M. Loper. Human Shape Estimation using Statistical Body Models. University of Tübingen, 2017.

[89] D. Tzionas. Capturing Hand-Object Interaction and Reconstruction of Manipulated Objects. PhD thesis. University of Bonn, 2017 (cited on page 15).

[90] A. Lehrmann. Non-parametric Models for Structured Data and Applications to Human Bodies and Natural Scenes. PhD thesis. ETH Zurich, 2016.