RESEARCH AND STATUS REPORT

Max Planck Institute for Intelligent Systems

Perceiving Systems Department

January 2011 - December 2015

Excerpt from Scientific Advisory Board Report

April 2016

MAX-PLANCK-GESELLSCHAFT

Contents

1 PERCEIVING SYSTEMS

1.1 Research Overview

Perceiving Systems is focused on understanding the 3D world and its motion as captured by images. Our goal is to formulate models of the world, refine these models with machine learning, and then relate these to how the world appears in images, enabling detection, recognition, tracking, and analysis. Given rich representations of the 3D world, image formation is the easy part; inverting the imaging process to produce descriptions from pixels is the challenge. We seek representations and algorithms that facilitate this reasoning and provide a foundation for vision systems that understand and interact with the time-varying 3D world.

Our fundamental assumption is that the parts of vision that involve the 3D world, surfaces, and light are "easy" in the sense that they relate to physical properties that can be precisely modeled. Such physical models are powerful because they require little (or no) training data and generalize widely. While the physical models may be simple, using them to analyze images – to solve inverse graphics – has proven hard. We have made significant progress in this regard in terms of new inference methods, our ability to fit 3D graphics models to image data, and our ability to extract basic physical primitives (intrinsic images) from images and video.

There are many aspects of the world that are not so easily described by physics – these we must learn. For example, the shapes of objects, the patterns of textures, the motion of animals are all more complex and we resort to learning their statistics. We have developed new statistical models of shape, new tools for transfer learning, new methods for object detection and recognition, and new robust methods for dimensionality reduction.

Like most computer vision groups today, we are making extensive use of deep learning. Un-

like some groups, however, these tools have not replaced modeling and optimization but have augmented them. We are using convolutional neural networks (CNNs) where they excel, on problems like 2D human pose estimation and semantic segmentation. For example, we leverage CNNs to improve results for estimating 3D human pose and for semantic optical flow estimation. The CNNs provide important cues that are combined with generative models to achieve novel results.

Following the approach of combining modeling and learning, Perceiving Systems has developed world-leading models and methods in optical flow estimation, human pose estimation, motion capture, body shape modeling, and that combine high-level scene understanding with low-level vision. We are also at the forefront of motor cortical decoding for neural prosthetics. These projects are described below.

Context. Computer vision is changing rapidly due to several simultaneous technological revolutions and, while deep learning is the most visible, others may be even more important. Maybe more significant are big datasets of images and platforms that enable human labeling on a large scale; these are key enablers of deep learning. Less noticed, but no less important, is the revolution taking place in computer graphics with open source gaming engines bringing high-quality rendering to everyone. For the first time, generative models can be rendered in real time with high quality for generating training data as well as for performing inference. Additionally, we have only just seen the beginnings of 3D scanning with devices like Kinect. This technology is poised to become widespread in consumer devices in the next few years and will result in a new data deluge, but one that we are less prepared to deal with. The tools, technologies, data structures, and algorithms for dealing with noisy, incomplete, 3D scans of the world are far less developed than techniques for image processing. Likewise, 3D printing and virtual reality applications suggest that databases of 3D CAD models are poised to expand.

The great leaps in deep learning have resulted from category-level labeling, which is relatively easy for humans. The labeling of metric properties of 3D scenes and objects will prove much harder. This argues for the sensible use of generative models of the 3D world that can be fit to relatively small amounts of training data and yet have strong generalization ability. A key lesson of current deep learning methods is that simple models, that are easy to train, are often preferable to more powerful models that are hard to train. Future generative models will exploit this insight to enable end-to-end training. The future will also likely combine discriminative, bottom up, pattern recognition methods with generative, top down, models. This promises a return to the early roots of computer vision but with new tools. The successful vision systems of the next decade are likely to build on, and embrace, all of these trends rather than focus on any one of them.

This generative approach is at an inflection point. New sensors and methods allow the capture of 3D objects, full 3D scenes, materials, and even 4D shape (3D shape over time). Rendering engines are better, more realistic, and more open than ever. Large datasets enable learning of object and scene statistics. Deep networks give new modeling tools to capture non-linear properties of the world. The combination of generative models, data, and learning offers a path to solving hard vision problems. Our approach is highly interdisciplinary, integrating computer vision, machine learning, computer graphics, and computational neuroscience.

Overview. Our approach can be summarized as "model what you can and learn the rest." As an example, the distribution over the shapes of different cars is something that is hard to write down but can be learned. The projection of a car shape into the image, the motion of the car on the road, contact and interpenetration with other objects, and the appearance under different lighting conditions are all physical things that are relatively easy to model. We see this philosophy of learning and modeling throughout our work. Some examples:

Inverse rendering: A rendering engine takes 3D models, materials, and lighting and produces images of the scene. The goal of inverse rendering is to turn this around and infer the 3D scene that generated the image. To that end we have developed an approximate differentiable renderer that efficiently does this when one is close to the solution. We have also developed sampling methods to deal with more complex scenes and to represent distributions over solutions.

Human body shape and motion: Humans and animals have complex 3D shapes that vary across individuals, with pose, and with motion. Using 3D and 4D scans we learn the world's most accurate statistical models of detailed human body shape. We use inverse rendering to then estimate human shape and pose from a variety of sources including mocap markers, RGB-D sequences, images, and video.

Scene understanding: Scenes are composed of objects with a spatial layout. We expect different objects and different spatial relations in outdoor scenes, traffic scenes, homes, and offices. Our goal is to combine semantic information about scenes with 3D information about objects to infer what objects are present, their shape, 3D pose, and how they are moving.

Stereo and optical flow: Both stereo and optical flow give important information about the 3D structure of the scene and the location of surface boundaries. They are typically viewed as low-level problems that provide this structural information to higher-level processes. We take a different view. Knowing something about the scene and its objects can make stereo and flow estimation easier. Consequently we formulate these estimation problems jointly to describe images and sequences in terms of semantic primitives and to leverage semantic segmentation.

Intrinsic images: Between image pixels and the 3D world are intermediate representations that are registered with the image but relate to the physical world. Examples include depth, flow, albedo, shading, object contours, cast shadows, etc. Extracting these intermediate representations has long been a goal and is now becoming feasible. By taking an integrated approach to estimating these intrinsic properties over time we are able to extract fundamental physical properties of scenes from video.

History. Perceiving Systems began operations in January 2011 with one employee (Black). We have grown into a department with a steady-state size of around 30 people including support staff, technicians, students, and scientists at various career stages. Already there are about 30 alumni including six graduated Ph.D. students.

The department has several exceptional group leaders including Juergen Gall (now a professor in Bonn), Peter Gehler (Senior Research Scientist), and Andreas Geiger (Research Scientist); all of these are top young researchers in the field of computer vision. Group leaders receive department funding and raise external funds to support their research. These group leaders independently supervise Ph.D. students.

We have a highly active visitor program and lecture series. We have had over 100 invited speakers, including many of the leaders in the field. A full list is here

https://ps.is.tuebingen.mpg.de/talks

Sabbatical and long-term visitors include Cordelia Schmid (INRIA), Stan Sclaroff (Boston University), Niko Troje (Queen's University), and Garrett Stanley (Georgia Tech).

The department occupies temporary space, primarily located in the first floor of the Magnetic Resonance Center of the MPI for Biological Cybernetics. Significant effort has gone into planning lab space in the new building to support our research program.

Our main research themes are described below, followed by more detailed project descriptions that provide insight into how we translate these themes into algorithms and solutions. The work presented here is just a sampling of the research in the department over our first five years of operation. Our website provides all this information as well as many other projects and greater detail

http://ps.is.tue.mpg.de

In addition, many of our videos are available on the Youtube channel

https://www.youtube.com/user/BlackAtBrown

Finally 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/

Selected highlights (2011–2015):

2011. Middlebury Optical Flow benchmark paper published.
2011. Demonstrated human neural control of a cursor 1000 days after implantation.
2011. Demonstrated first decoding of point and click from human motor cortex.
2011. First method to estimate human shape from multiple Kinect RGB-D images.
2011. Best method for estimating intrinsic images using a global sparsity prior.

2012. Released the MPI-Sintel dataset for optical flow evaluation.
2012. World's first high-speed body scanner, capturing the full range of human poses.
2012. Deformable parts models with 3D object geometry for object recognition.
2012. Coregistration method enables learning of body shape from a corpus of scans.
2012. Department's first SIGGRAPH paper (on clothing shape and deformation).

2013. Spun out Body Labs Inc. with 2 million USD of angel funding.
2013. Release of JHMDB action recognition dataset.
2013. State of the art 2D human pose estimation with poselets.
2013. Demonstrated state-of-the-art optical flow estimation using layers.

2014. World's first 4D body scanner, capturing 3D meshes at 60 fps.
2014. First learned model of human shape change during breathing.
2014. MoSh estimates human shape and detailed motion from standard markers.
2014. FAUST dataset for 3D mesh registration released.
2014. Released first method for robust PCA that scales to big data.

2015. Body Labs receives $8 million in venture funding; licenses new IP.
2015. SMPL body model released; compatible with standard graphics packages.
2015. Released KITTI 2015 dataset with ground truth non-rigid motions.
2015. Discrete optical flow achieves top performance on benchmarks.
2015. Joint estimation of high-level object models and low-level scene properties.
2015: Dyna the first realistic model of dynamic human shape in motion.

1.1.1 Human Pose

Since humans are often the subject of photographs, detecting them and analyzing their pose is critical for image understanding. The photographic study of human pose and motion dates from the late 1800's with the work of Muybridge and Marey. Our research continues this tradition but with new capture technology, advanced graphics models of the body, new algorithms for pose and shape estimation, machine learning methods, and quantitative analysis of human motion and pose on ground-truth datasets.

In the first five years of Perceiving Systems we have made significant progress towards automatically estimating 2D human pose from images by leveraging training datasets and machine learning methods. We also leverage our expertise in optical flow estimation to extend 2D pose estimation over time, resulting in increased accuracy.

Beyond 2D pose, we are pushing the technology of "motion capture" in new directions. The goal is always to leverage what we know about bodies to get more from less - more accuracy and more shape detail from a small number of simple sensors. From 3D mocap markers, we recover detailed shape, pose, and soft tissue motion. Using a single RGB-D sensor and a parametric model of the human body, we are able to estimate human body shapes and poses from complex sequences of unconstrained motion.

Our current work is pushing the state of the art in monocular pose and motion capture to automatically go from 2D images or monocular video to 3D pose and shape of the human body. We are also expanding our research from tracking humans to tracking animals of many kinds.

More information: https://ps.is.tuebingen.mpg.de/field/human-pose

1.1.2 Human Shape

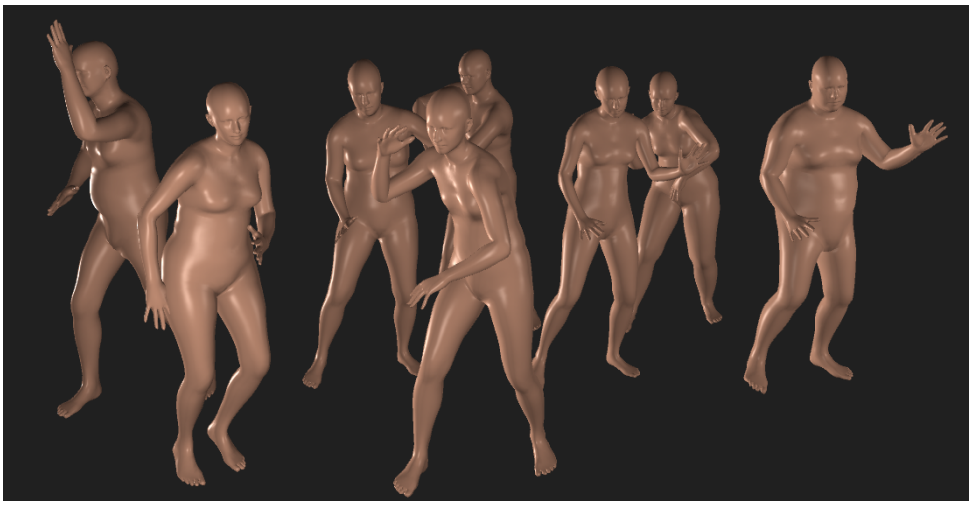


Figure 1.1: **Virtual Humans.** SMPL is an example of a statistical body model that is learned from thousands of scans of people. Unlike previous models, SMPL is compatible with existing game engines. Here many bodies are rendered in motion by a common game engine

The human body is special. Most images and videos depict humans, and understanding humans is important for many problems including human-computer interaction, video retrieval, activity recognition, special effects, sports medicine, etc. We take an approach that leverages strong models to interpret ambiguous sensor data. Such models express the statistics of body shape and pose and allow the robust integration of measurements from different sources. A model-based approach is particularly important for the analysis of complex, articulated, and non-rigid objects such as the body.

To that end we have built the most detailed and accurate statistical models of 3D human body shape to date. These models are learned from over 4000 3D body scans of different people and approximately 1800 scans capturing a wide range of poses by people of many body shapes.

Additionally we learn models of soft tissue dynamics using 40,000 scans captured by a unique full-body 4D body scanner that gives detailed 3D meshes at 60 fps.

Our latest SMPL body model is available for research purposes, makes it easy to create any human body shape, and allows the body to be animated in standard game engines and graphics software. The model is appropriate for use in animation and computer vision.

Some of our current work addresses: learning models of clothing in motion; modeling hands, faces and bodies together; learning compositional models of 3D shape; animal shape and motion; estimating 3D shape from monocular cues.

Impact. From its inception, Perceiving Systems has been working on commercializing body shape modeling technology. An engineering team from Brown formed the first group of employees in the department and they came to Germany to make our technology ready for the real world. After nearly two years of work, several significant papers, and patenting activity, the team spun off in 2013 into Body Labs Inc. The company licensed technology from Brown and Max Planck and in the fall of 2015 completed a second round of licensing from MPI.

Located in New York City, Body Labs began with angel funding and quickly built a base of paying customers. In 2015 it closed a Series A financing round with Intel Capital in the lead, bringing total funding to approximately $10 million dollars. At the same time Body Labs and Intel announced a partnership to bring body scanning and clothing sizing to consumers using Intel's RealSense depth sensors.

More information can be found at the Body

Labs website

http://www.bodylabs.com/

In addition to our commercialization efforts we have made several websites available to users for free. In particular we developed websites to help people better understand their body shape and how this shape relates to their body mass index, or BMI. We have two websites, using different visualization technology, that attract about one million users a year

http://www.bodyvisualizer.com
http://www.bmivisualizer.com

More information: https://ps.is.tuebingen.mpg.de/field/shape

1.1.3 Stereo and Optical Flow

A fundamental problem in computer vision is the reconstruction of the shape and motion of the 3D world. This has applications as varied as self-driving cars, 3D mapping, virtual reality, graphics, and robotics. We think that reasoning about the 3D world and its structure is at the heart of computer vision.

To help push the field in new directions, we have co-organized two workshops on *Scenes from Video* that bring together researchers working on video, flow, and structure from motion with researchers working on semantic scene analysis. The idea is that integration of these fields (metric and semantic) will lead to improvements in both.

Image Motion: Perceiving Systems is at the forefront of research on optical flow; it is one of our core competencies and our algorithms are regularly at the top of the optical flow benchmarks.

By optical flow we mean the projection of the 3D motion field onto the image plane of the camera. We focus on this (as opposed to apparent motion) because this flow is related to the structure of the 3D scene, the boundaries of objects, and the motion of the camera. Flow is an important mediating representation (an intrinsic image) that helps the analysis of scenes.

Optical flow has proven useful for problems throughout computer vision, graphics, medical imagining, robotics, and many application domains. and while there are many reasons to compute flow, the ones that interest us most are to

1. establish correspondence across time - this enables reasoning across time, establishes object permanence etc.;
2. to determine scene structure - what is rigid, what isn't, where the boundaries are, etc.

Open problems in the field include: dealing with fast motion of small objects, modeling motion with complex material properties, reflections and transparency, dealing with motion blur, accurately estimating flow at surface boundaries, segmenting scenes into regions, and improving accuracy and speed simultaneously.

Our current work is focused on combining the estimation of flow with higher level scene analysis, including combing flow with the estimation of 3D objects and their motion and estimating 3D scene flow. Our most recent work combines semantic scene segmentation with optical flow, achieving state-of-the-art accuracy. We also are using optical flow in many applications, including human shape and motion analysis.

Scene Structure: Beyond motion, we study the recovery and reconstruction of 3D structure from single images, RGB-D data, video sequences, stereo, and multi-view stereo.

Our major innovations lie in combining high-level and semantic cues with low-level features. We view the problem as the integration of model fitting with dense structure recovery. While much of our work has focused on object-specific models like people and cars, we are particularly interested in generic representations and compositional models of objects and scenes.

Increases in computing power, labeled training data, large databases of 3D CAD models, 3D sensors, and open-source rendering engines, are all opening new opportunities to model and infer 3D objects and scenes.

More information: https://ps.is.tuebingen.mpg.de/field/stereo-and-flow

1.1.4 Vision as Inverse Graphics

Computer vision as analysis by synthesis has a long tradition and remains central to a wide class of generative methods. In this top-down approach, vision is formulated as the search for parameters of a model that is rendered to produce an image (or features of an image), which is then compared with image pixels (or features). The model can take many forms of varying realism but, when the model and rendering process are designed to produce realistic images, this process is often called inverse graphics. In a sense, the approach tries to reverse-engineer the physical process that produced an image of the world.

Recent advances in graphics hardware, open source renderers, and probabilistic programming is making this approach viable. We are addressing inverse rendering in multiple projects that use autodifferentiation and stochastic sampling to solve different aspects of the problem. For example, the OpenDR framework is widely used in much of our research on human body modeling. It allows us to very quickly formulate a problem and prototype a solution.

We also approach the problem from the "bottom up"; that is, from images and videos we extract intrinsic images, which represent physical properties of the scene tied to the pixel grid. These provide a generative model of images (or video) and can be used as an intermediate representation between graphics models and images.

More information: https://ps.is.tuebingen.mpg.de/field/inverse-graphics

1.1.5 Learning & Inference

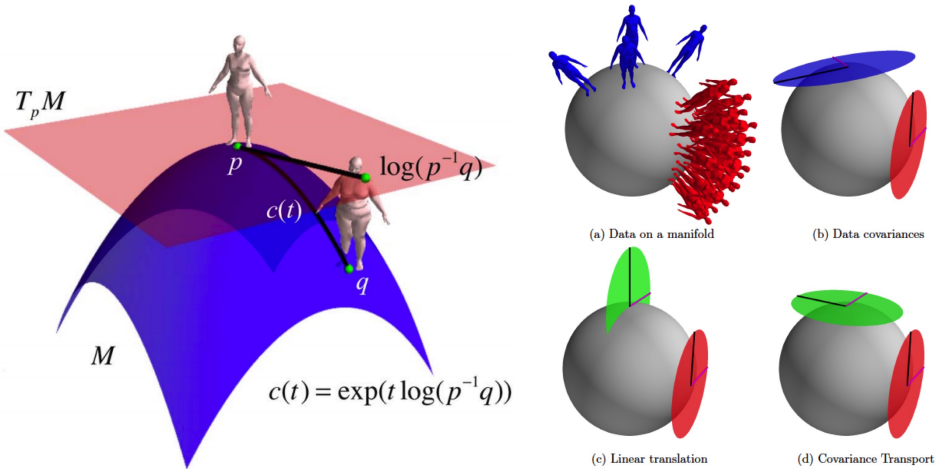


Figure 1.2: Statistics on manifolds. Left: body shape represented as deformations where the deformations lie on a manifold. Right: transporting statistics on a manifold. Here we use the statistics of female shape to regularize the shape of men using covariance transport.

Our work on learning is infused through our projects and is typically grounded in specific computer vision applications. We also work on more generic learning problems. For example, we have formulated the learning of 3D shapes represented by triangle deformations (Lie shapes) [108]. We show how such deformations live on a manifold and how to model the statistics on this manifold using principal geodesic analysis.

Often good data about 3D object shape is hard to come by. Consequently to learn 3D shape models, it is useful to be able to transfer shape knowledge from previously learned shapes. When these shapes live on a manifold, however, standard methods for domain transfer do not apply. We show how to transfer statistics on a manifold using parallel transport [79]. This provides an efficient way to exploit the manifold structure to improve learning of object shape when very little data is available.

Often the manifold structure of the data is not known a priori and we need to learn it. To address this, multi-metric learning techniques learn local metric tensors in different parts of a feature space. The learned distance measure is,

however, non-metric, which has prevented multi-metric learning from generalizing to tasks such as dimensionality reduction and regression in a principled way. We prove that, with appropriate changes, multi-metric learning corresponds to learning the structure of a Riemannian manifold [103]. We then show that this structure gives us a principled way to perform dimensionality reduction and regression according to the learned metrics. Algorithmically, we provide the first practical algorithm for computing geodesics according to the learned metrics, as well as algorithms for computing exponential and logarithmic maps on the Riemannian manifold. Together, these tools let many Euclidean algorithms take advantage of multi-metric learning.

Today learning is often applied to large datasets labeled by humans or generated by algorithms. In either case, we expect the number of outliers to increase with data size. While principal component analysis (PCA) can reduce data size, and scalable solutions exist, it is well-known that outliers can arbitrarily corrupt the results. Unfortunately, state-of-the-art approaches for robust PCA are not scalable. We note that in a zero-mean dataset, each observation spans a one-dimensional subspace, giving a point on the Grassmann manifold. We use this insight to compute PCA by computing average subspaces. Because averages can be efficiently computed, we immediately gain scalability. We exploit robust averaging to formulate the Robust Grassmann Average (RGA) as a form of robust PCA [7]. Our algorithm has linear computational complexity and minimal memory requirements.

Our work on inference is grounded in probability and exploits graphical models, belief propagation, and stochastic sampling to name just a few methods. In most problems that we care about, the variables of interest are continuous and high dimensional. Specifically, we developed a particle-based max-product algorithm which maintains a diverse set of posterior mode hypotheses, and is robust to initialization [75]. At each iteration, the set of hypotheses at each node is augmented via stochastic proposals, and then reduced via an efficient selection algorithm. The integer program underlying our optimization-based particle selection minimizes errors in subsequent max-product message updates. This objective automatically encourages diversity in the maintained hypotheses, without requiring tuning of application-specific distances among hypotheses. By avoiding the stochastic resampling steps underlying particle sum-product algorithms, we also avoid common degeneracies where particles collapse onto a single hypothesis. Our approach significantly outperforms previous particle-based algorithms in experiments focusing on the estimation of human pose from single images.

More information: https://ps.is.tuebingen.mpg.de/field/learning-inference

1.1.6 Understanding Objects and Scenes

Object and scene understanding involves figuring out, at the very least, what is in an image and where things are. Moreover we want to know information about the scene and how objects in it are spatially related. The dominant paradigms treat this as primarily a pattern recognition problem that involves learning some filter-based representation of images that makes the detection and classification problem easier. In contrast, our work on recognition often brings in 3D knowledge about objects in a variety of ways.

Our work addresses:

- object modeling
- object detection
- object recognition
- scene understanding
- scene segmentation
- humans interacting with objects
- machine learning methods
- statistical modeling of scene properties
- geometric models and reasoning.

Scene understanding, in contrast to object recognition, attempts to analyze objects in context with respect to the 3D structure of the scene, its layout, and the spatial, functional, and semantic relationships between objects. Our research in this area combines object detection/recognition with 3D reconstruction and spatial reasoning. We believe that the integrated analysis of low-

level image features, together with high-level semantic and 3D object models, will enable robust scene understanding in complex and ambiguous environments and will provide the foundation for further reasoning.

More information: https://ps.is.tuebingen.mpg.de/field/understanding-objects-and-scenes

1.1.7 Datasets & Evaluation

Datasets with ground truth have driven many of the recent advances in computer vision. They allow evaluation and comparison so the field knows what works. They also provide training data to machine learning methods that are hungry for data. Creating good datasets that are valuable to the community and have a reasonable lifespan is hard work. Key issues are the quality and quantity of the data, how well that data addresses a specific problem in the field, and whether it is well curated with a good evaluation.

We have played central roles in many influential datasets and evaluations in the field including

- Middlebury flow dataset
- MPI-Sintel datasets
- KITTI datasets
- HumanEva for human pose estimation
- FAUST for 3D mesh registration
- JHMDB for action recognition
- MPI-I human pose dataset.

Note that Perceiving Systems is involved in all three of the standard benchmarks in optical flow (Middlebury, KITTI, and Sintel).

We are committed to releasing data whenever possible including

- Dyna: 40,000 4D human body scans
- Motion capture of extreme human poses.

More information: https://ps.is.tuebingen.mpg.de/field/datasets-evaluation

1.1.8 Computational Neuroscience

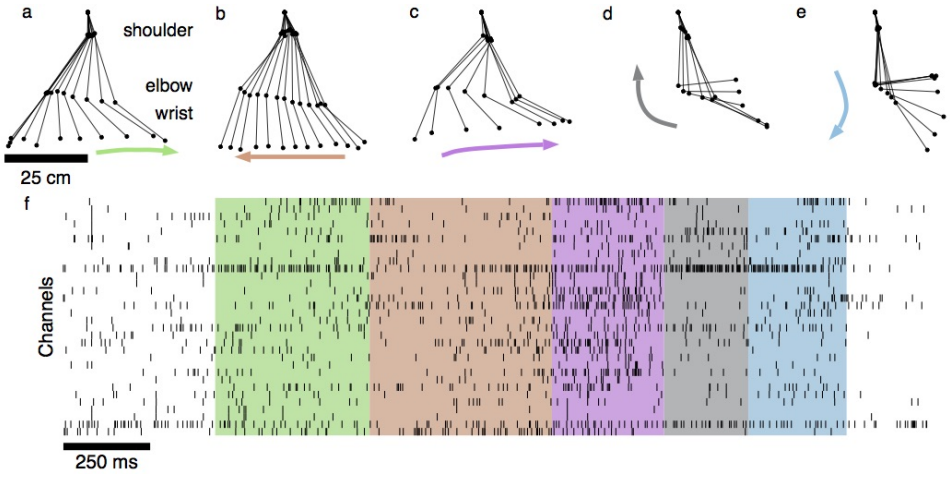


Figure 1.3: **Decoding walking.** The motion of the front leg of a monkey walking on a treadmill is shown [116]. The raster plot show the neural firing activity from a population of motor cortical neurons. The color coding corresponds to different phases of the walk cycle, which can be decoded from the neural firing rates.

While most of our work is on computer vision, we are also interested in how brains solve problems in vision and motor control. For example we have shown how the synchronous firing of cells in LGN with highly overlapping receptive fields can code information about optical flow [32].

Most of our work in computational neuroscience, however, is focused on movement. We do not just care about how people and their ac-

tions *appear* in image sequences – we care about *how* they move. We take a somewhat radical view, that to understand movement in images, it is useful to understand how motor systems produce movement. The majority of work in basic motor neurophysiology, however, focuses on motor control in highly constrained scenarios. We argue that such scenarios lead to overly simplistic models. To go beyond this, the science of motor control must be studied in natural settings. We take a novel approach that marries computer vision with motor neurophysiology.

We seek insight about how the brain controls natural behavior in natural environments. To that end we study

- motion capture of natural behavior
- modeling motor cortical activity during natural behavior
- developing new neural decoding algorithms
- applying our models to brain machine interfaces with implanted electrocortical arrays.

In collaboration with researchers at Brown University, we were the first to decode motor cortical signals from the human brain using an implanted microelectrode array and turn these signals into a viable computer control system. We demonstrated the first point and click system that could decode both intended cursor movement and a "click" signal from the same neural population. We also demonstrated the viability of such systems in humans by showing human neural cursor control at over 1000 days post implantation.

To go further we need a richer understanding of how the brain controls movement and for that we need to observe and model the neural control of animal movement in much more complex and natural settings than previously seen. To that end, our current work focuses on markerless animal motion capture.

We are just at the beginning of our ability to capture animal motion. Our current work is focused on learning models of animal shape and motion that are similar in quality to our models of humans. This is a challenging task. Unlike humans, it will be impossible to capture 3D scans of thousands of animals of varying shape in a wide variety of poses. To build such models we are using heterogenous sources of data of varying quality and are developing new models and algorithms to learn high quality 3D models from this data. While this presents many challenges, the ability to track animals (and groups of animals) in natural settings would revolutionize many fields of biology.

More information: https://ps.is.tuebingen.mpg.de/field/computational-neuroscience

1.2 Selected Research Projects

2D Pose from Images

Peter Gehler, Javier Romero, Silvia Zuffi, Martin Kiefel, Jürgen Gall, Michael Black

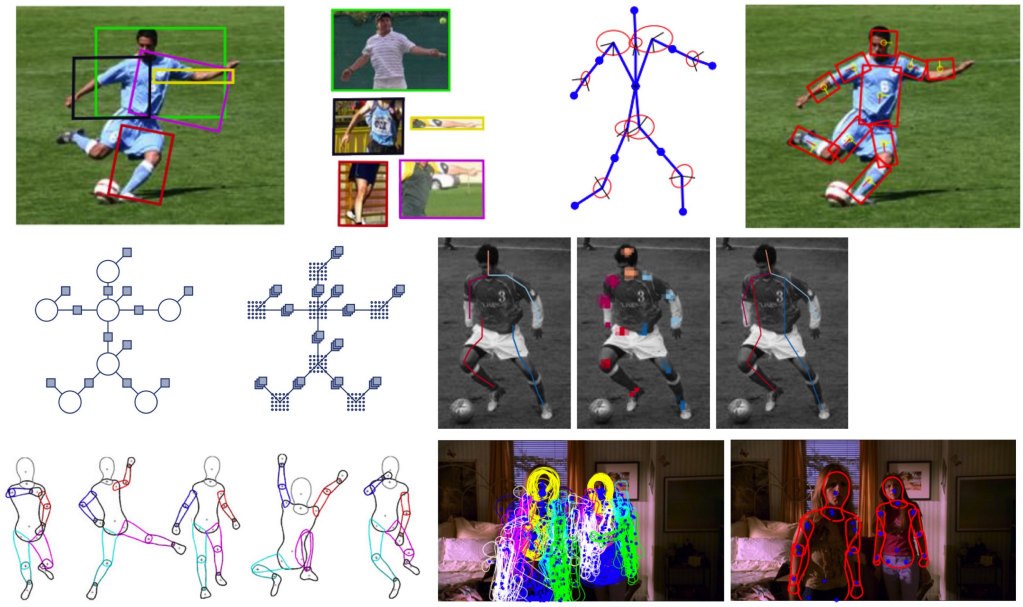


Figure 1.4: Top row: Poselets are used to condition a pictorial structures model, providing more contextual information, while maintaining efficient pose inference [90, 97]. Middle row: The Fields of Parts model reformulates pose using a binary variable for every possible part location, orientation and scale [70]. Bottom row: The deformable structures model [121] contains information about 2D body shape and how it varies with pose. Inference uses a new non-parametric belief propagation algorithm [75].

Estimating 2D human pose is hard because people appear in a wide range of poses and have varying body shape. They wear varied clothing and the articulation results in significant self occlusion. We have developed several state-of-the-art methods to address these problems.

Poselets [90, 97] capture how human motions and activities simultaneously constrain the positions of multiple body parts. Our model incorporates higher-order part dependencies while retaining efficient inference. We achieve this by defining a conditional model in which all body parts are connected a-priori, but which becomes a tractable tree-structured pictorial structure model given image observations. In order to derive a set of conditioning variables we exploit the poselet-based features that capture extended spatial information about pose.

Our Fields of Parts model [70] reformulates the problem as a binary Conditional Random Field that models local appearance and joint spatial configuration of the human body. Using a novel graph structure, we model the presence and absence of a body part at every possible position, orientation, and scale in an image with a binary random variable; this encodes the same appearance and spatial structure as Pictorial Structures. While the formulation results into a vast number of random variables, approximate inference is efficient. Fields of Parts can use evidence from the background, include local color information, and it is connected more densely than a kinematic chain structure.

Like pictorial structures, these models lack an explicit model of body shape. We learn a deformable structures body model that captures body shape and how it deforms with pose in 2D [121]. The DS image likelihoods explicitly model image information at the boundaries of body parts, simplifying learning. The model is not much more complex than previous models but results in improved accuracy. Inference uses a new non-parametric method for max-product belief propagation that preserves particle diversity, models uncertainty, and estimates the pose of multiple bodies simultaneously [75].

More information: https://ps.is.tuebingen.mpg.de/project/people-from-images

2D Pose from Optical Flow

Michael Black, Javier Romero, Matthew Loper, Silvia Zuffi, Cordelia Schmid, Hueihan Jhuang

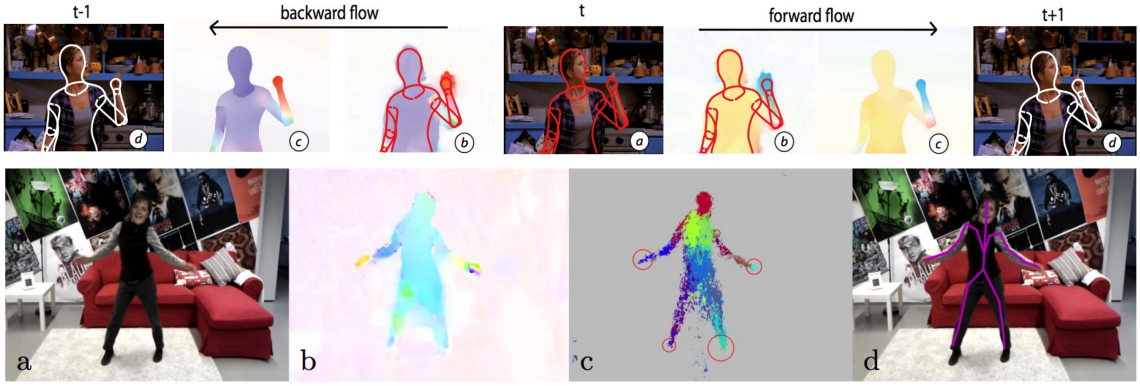


Figure 1.5: Top row: Flowing puppets. (a) Frame with a hypothesized human “puppet” model. (b) Dense flow between frame (a) and its neighboring frames. (c) The flow of the puppet is approximated by a part-based affine model. (d) Prediction of the puppet from (a) into the adjacent frames using the estimated flow. Bottom Row: FlowCap. (a) Example frame from a video sequence shot with a phone camera. (b) Optical flow. (c) Per-pixel part assignments based on flow with overlaid uncertainty ellipses (red). (d) Predicted 2D part centroids connected in a tree.

Much of the work on human pose estimation focuses on still images. We argue that there is much to be gained by looking at video sequences and, specifically, using optical flow. Flow tells us what goes with what over time. This allows the temporal propagation of information, which can reduce uncertainty in pose estimation. Flow also provides strong cues about objects in the scene, their boundaries, and how they move. We find that optical flow algorithms are now good enough to play an important role in human pose estimation.

Inferring pose over a video sequence is advantageous because poses of people in adjacent frames exhibit properties of smooth variation due to the nature of human and camera motion. Here we make a simple observation: Information about how a person moves from frame to frame is present in the optical flow field. We develop an approach for tracking articulated motions that "links" articulated shape models of people in adjacent frames trough the dense optical flow [91]. Key to this approach is a 2D shape model of the body [121] that we use to compute how the body moves over time. The resulting "flowing puppets" integrate image evidence across frames to improve pose inference.

Dense optical flow provides information about 2D body pose [48]. Like range data, flow is largely invariant to appearance but unlike depth it can be directly computed from monocular video. We demonstrate that body parts can be detected from dense flow alone using the same random forest approach used by the Microsoft Kinect. Unlike range data, when people stop moving, there is no optical flow and they effectively disappear. To address this, our FlowCap method uses a Kalman filter to propagate body part positions and velocities over time and a regression method to predict 2D body pose from part centers from only monocular video of people moving.

Finally in [87] we explore the importance of optical flow for human activity recognition. We create a novel dataset of complex video sequences with ground truth 2D pose and flow using our deformable structures model [121]. We find that optical flow can play an important role in human action recognition.

More information: https://ps.is.tuebingen.mpg.de/project/pose-from-flow

Beyond Motion Capture

Javier Romero, Naureen Mahmood, Matthew Loper, Federica Bogo, Alex Weiss, Michael Black

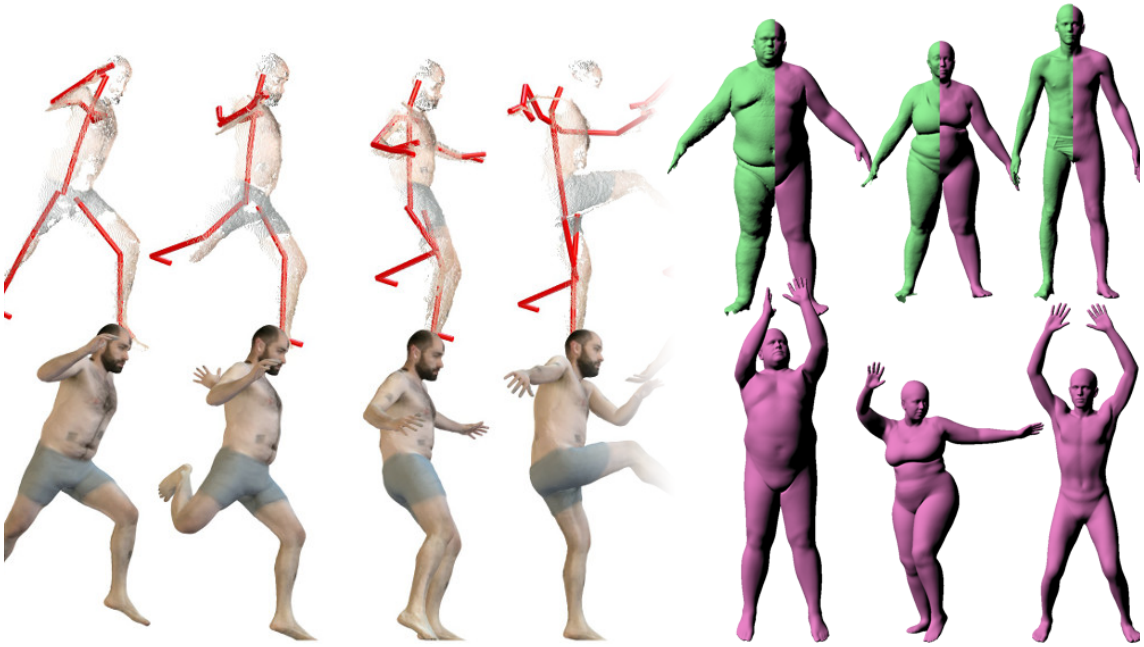


Figure 1.6: We use a parametric body model to estimate accurate body shape, pose and appearance, and even to extract soft-tissue deformations from incomplete, noisy 3D data. Left: we show a sequence of monocular RGB-D frames from Kinect (top row, Kinect skeleton in red) and our model, estimated from the frames (bottom row). Right: MoSh computes body shape and pose from standard mocap marker sets (green = 3D scan, purple = estimated body shape and pose).

Accurately capturing human body shape and motion is important for many applications in computer vision and graphics. Traditional motion capture (mocap) focuses on extracting a skeleton from a sparse set of markers. Our work pushes the boundaries of motion capture to use new sensors and to extract richer information about body shape and human movement.

Traditional mocap uses a set of sparse markers placed on the body to estimate skeleton motion. These markers are typically placed on parts of the body that move rigidly to try to minimize the effects of soft tissue motion. In this process nuanced information about surface motion is lost and animations using mocap often feel lifeless or eerie. MoSh (Motion and Shape capture) [15] addresses this problem by directly estimating a 3D parametric body model from 3D markers. Given a standard marker set, MoSh simultaneously estimates the marker locations on the 3D model and recovers body shape and pose. By allowing body shape to vary over time, MoSh can also capture the non-rigid motion of soft tissue. From a small set of markers MoSh is able to recover a remarkably accurate 3D model of the body. The motions can then be retargetting to new characters, resulting in realistic, lifelike, animations.

In comparison with mocap, consumer RGB-D devices provide denser observations of the body, but these scans are incomplete (taken from a single view) and noisy. By using RGB-D sequences of bodies in motion, we can extract more detailed information about body shape and motion [42]. To do so, we introduce a multi-resolution body model and exploit time continuity of human motion and RGB appearance to estimate accurate body shape, pose and appearance. The approach can track arbitrary challenging motions, and extracts highly realistic 3D textured avatars with an accuracy rivaling high-cost laser scanners.

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

Multi-Camera Capture

Michael Black, Jürgen Gall, Gerard Pons-Moll

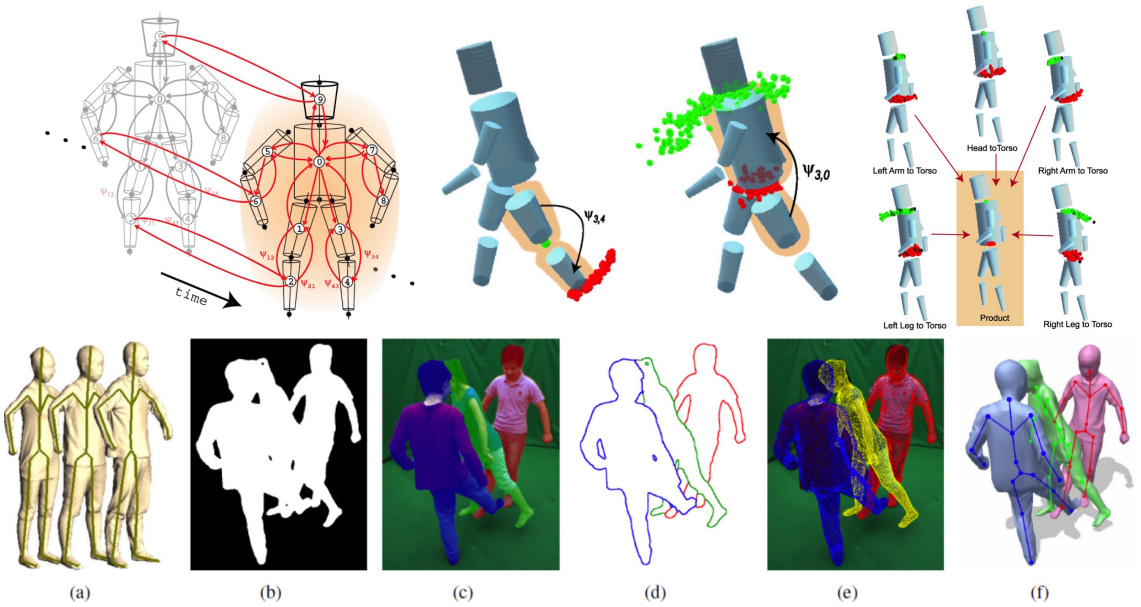


Figure 1.7: Top row: (left) In [35], bodies are represented by a part-based graphical model in space and time. (middle) Messages between parts are represented by particles. (right) Non-parametric belief propagation computes message products. Bottom row: In [29], we segment and fit bodies multi-camera images. (a) Articulated template models. (b) Input silhouettes. (c) Segmentation. (d) Contour labels assigned to each person. (e) Estimated surface. (f) Estimated 3D models with embedded skeletons.

While multi-camera video data facilitates markerless motion capture, many challenges remain.

We formulate the problem of 3D human pose estimation and tracking as inference in a graphical model [35]. The body is modeled as a collection of loosely-connected body-parts (a 3D pictorial structure) using an undirected graphical model in which nodes correspond to parts and edges to kinematic, penetration, and temporal constraints. These constraints are encoded using pair-wise statistical distributions, learned from mocap data. Human pose and motion are computed using Particle Message Passing, a form of non-parametric belief propagation that can be applied over graphical models with loops. The loose-limbed model and decentralized graph structure allow us to incorporate "bottom-up" visual cues, such as limb and head detectors into the inference process. These detectors enable automatic initialization and aid recovery from transient tracking failures.

Capturing the skeleton motion and detailed time-varying surface geometry of multiple, closely interacting persons is harder still, even in a multi-camera setup, due to frequent occlusions and ambiguities in feature-to-person assignments. To address this, we propose a framework that exploits multi-view image segmentation [29]. To this end, a probabilistic shape and appearance model is employed to segment the input images and to assign each pixel uniquely to one person. Given the articulated template models of each person and the labeled pixels, a combined optimization scheme, which splits the skeleton pose optimization problem into a local one and a lower dimensional global one, is applied one-by-one to each individual, followed by surface estimation to capture detailed non-rigid deformations. Our approach can capture the 3D motion of humans accurately even if they move rapidly, wear apparel, and engage in challenging multi-person motions.

More information: https://ps.is.tuebingen.mpg.de/project/multi-camera-capture

Pose and Motion Priors

Peter Gehler, Andreas Lehrmann, Ijaz Akhter, Gerard Pons-Moll, Søren Hauberg, Jürgen Gall, Michael Black

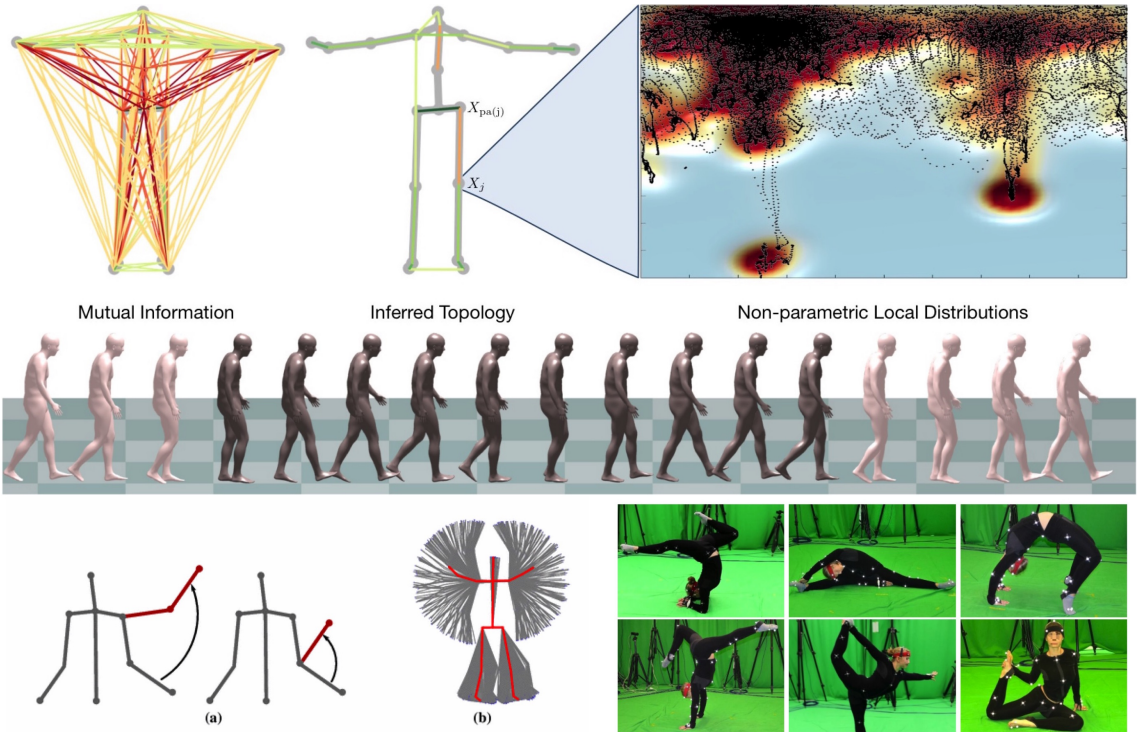


Figure 1.8: Top row: Representation of a human pose as a Bayesian network with optimal tree-structured topology and non-parametric local distributions [89]. Middle row: Markov model from [84] interpolates human pose realsitically. Bottom row: (left) Distance metric based on a manifold in joint space [104]. (right) New motion capture dataset of extreme poses [56].

A prior over human pose is important for many human tracking and pose estimation problems.

We introduce a sparse Bayesian network model of human pose that is non-parametric with respect to the estimation of both its graph structure and its local distributions [89]. Using an efficient sampling scheme, we tractably compute exact log-likelihoods. The model is compositional, representing poses not present in the training set. It remains useful for real-time inference despite being non-parametric.

Action recognition and pose estimation are closely related topics; information from one task can be leveraged to assist the other, yet the two are often treated separately. In [34] we develop a framework for coupled action recognition and pose estimation by formulating pose estimation as an optimization over a set of action-specific manifolds. The framework allows for integration of a 2D appearance-based action recognition system as a prior for 3D pose estimation and for refinement of the action labels using relational pose features based on the extracted 3D poses.

Modeling distributions over human poses requires a distance measure between human poses; this is often taken to be the Euclidean distance between joint angle vectors. In [104] we present an algorithm for computing geodesics in the Riemannian space of joint positions, as well as a fast approximation that allows for large-scale analysis. Articulated tracking systems can be improved by replacing the standard distance with the geodesic distance in the space of joint positions. This measure significantly outperforms the traditional measure in classification, clustering and dimensionality reduction tasks.

To better model human pose we collected a new motion capture dataset of extreme poses [56] that is available to the public.

More information: https://ps.is.tuebingen.mpg.de/project/pose-and-motion-priors

Hands in Action

Dimitris Tzionas, Jürgen Gall, Javier Romero, L. Ballan, A. Taneja, L. Van Gool, M. Pollefeys, Michael Black

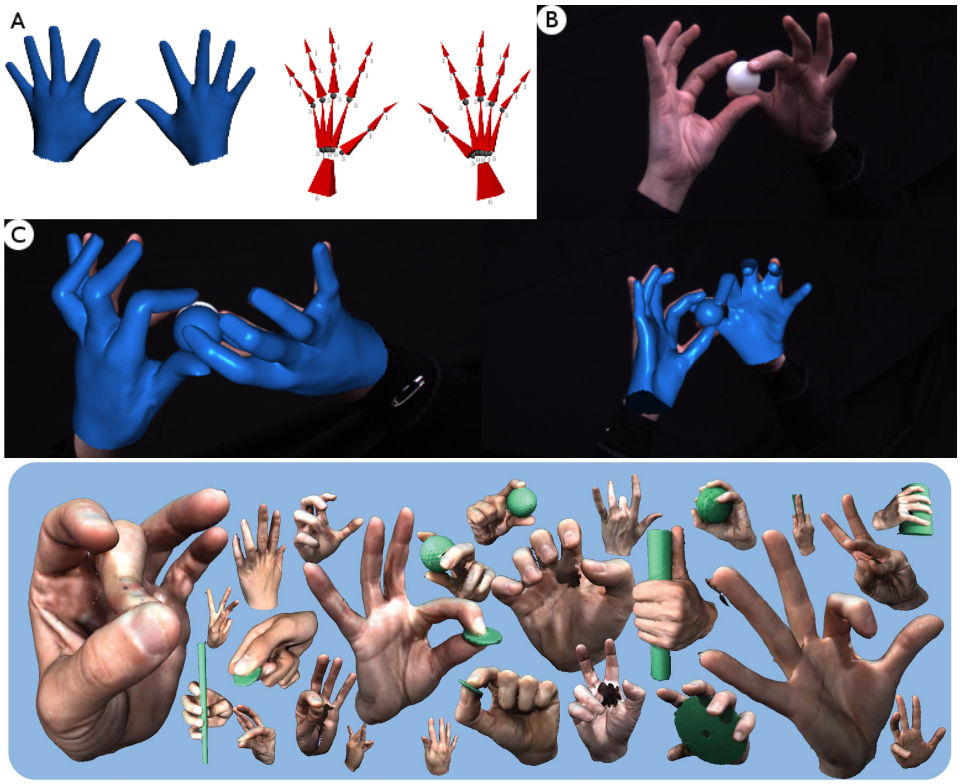


Figure 1.9: (top) Capturing the motion of two hands interacting with an object. A. Mesh models (blue) and underlying bone skeletons (red) are used to represent the hands in the scene. B. One frame of a sequence where two hands are interacting with a ball. This sequence consists of a total of 73 degrees of freedom and has been captured by 8 synchronized cameras. C. Estimated poses of the hands and the ball superimposed on two different camera views. (bottom) Examples from a database of 3D hand scans.

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, or robotics.

We focus on hands that interact with other hands or objects and develop a framework that successfully captures motion in such interaction scenarios for both rigid and articulated objects [112]. Our framework combines a generative model with discriminatively trained salient points and collision detection to achieve a low tracking error and physically plausible poses even in case of occlusions and missing visual data. Our approach captures hand motion using either a single RGB-D camera or multiple synchronized RGB cameras.

The captured hand motion can be used to improve the reconstruction of hand-sized objects that are manipulated in front of a RGB-D camera. Instead of discarding the hands, we developed a framework that uses the captured hand motion together with texture and geometric features for object reconstruction [41]. Since the hand motion provides additional information about the object motion, we can reconstruct even textureless and symmetric objects.

Our current work is focused on modeling hand shape and pose across many people, on estimating hands and their pose from low-resolution data, and on estimating hands and bodies together. To that end, we have collected a large dataset of hands of various subjects in a wide range of poses that include object interaction. From this we are building a statistical model of hand shape using the same approach as our SMPL body model.

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

3D Mesh Registration

Matthew Loper, Javier Romero, Federica Bogo, Aggeliki Tsoli, David Hirshberg, Eric Rachlin, Alex Weiss, Gerard Pons-Moll, Michael Black

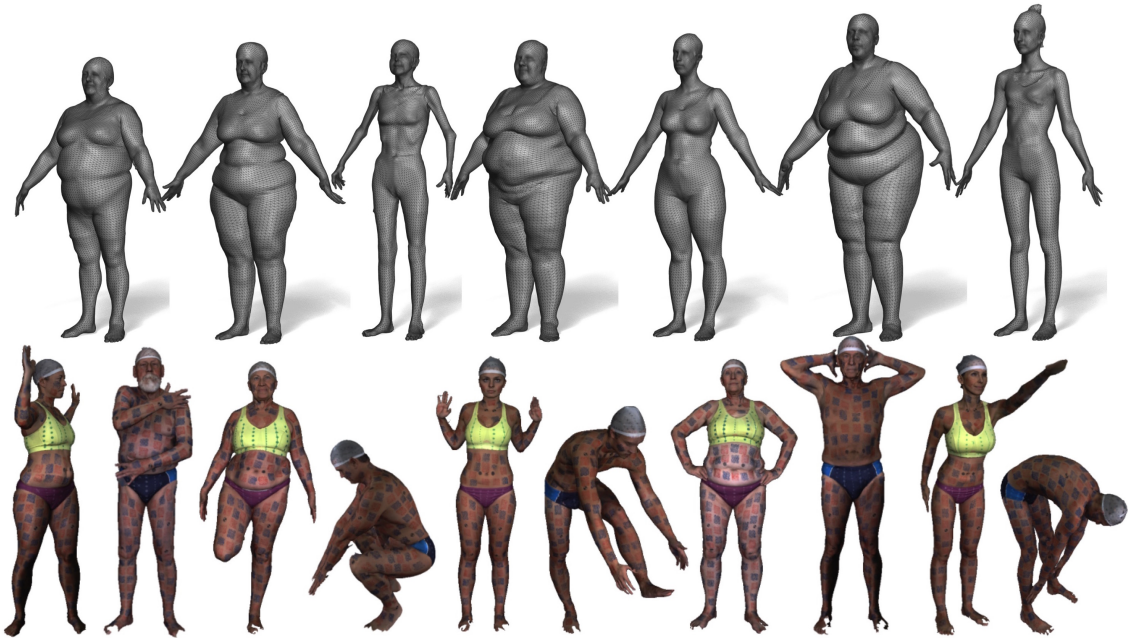


Figure 1.10: Registering a corpus of 3D body scans involves bringing a template mesh into alignment with each scan. (top) Example registrations from CAESAR of widely different body shapes. (bottom) Example of pose variation from FAUST with high-frequency texture pattern.

Statistical shape models enable the inference of object shape from incomplete, noisy, ambiguous 2D or 3D data. Training such models requires precisely registering a corpus of 3D scans with a common 3D template.

Registering a template mesh to 3D scans is challenging [123]. Scans may have noise, holes, and self contact while objects like the human body deform in complex non-rigid ways. Registration is ill-posed and solutions typically use generic regularizers to penalize implausible template deformations. Instead we wish to penalize deformations from a model of body shape. Constructing such a model, however, requires having a registered corpus of scans.

We solve this chicken-and-egg problem by performing modeling and registration together. Coregistration [102] registers a corpus of scans and simultaneously learns a parametric model of human body shape and pose by minimizing a single objective function. The model greatly improves robustness to noise and missing data. Since it explains a corpus, it captures how body shape varies across people and poses. This is the key to accurate body shape modeling.

Using coregistration we have registered a template mesh with 7000 vertices to the 4000 bodies in the CAESAR dataset. We think this is the most accurate and detailed registration of CAESAR to date. Using our model we pose-normalize the meshes enabling us to learn a vertex-based statistical model that is independent of pose variation in the dataset [9].

Geometry alone is ambiguous in smooth regions and results in registrations that slide along the surface. To address this, we extend coregistration to use both geometry and surface color [77]. Our approach estimates scene lighting and surface albedo to construct a high-resolution textured 3D model. This model is robustly brought into registration with multi-camera image data. We build our statistics shape models from about 1800 scans of 60 people. Some of this data is publicly available in the FAUST dataset.

More information: https://ps.is.tuebingen.mpg.de/project/3d-mesh-registration

4D Shape

Gerard Pons-Moll, Javier Romero, Naureen Mahmood, Matthew Loper, Aggeliki Tsoli, Michael Black

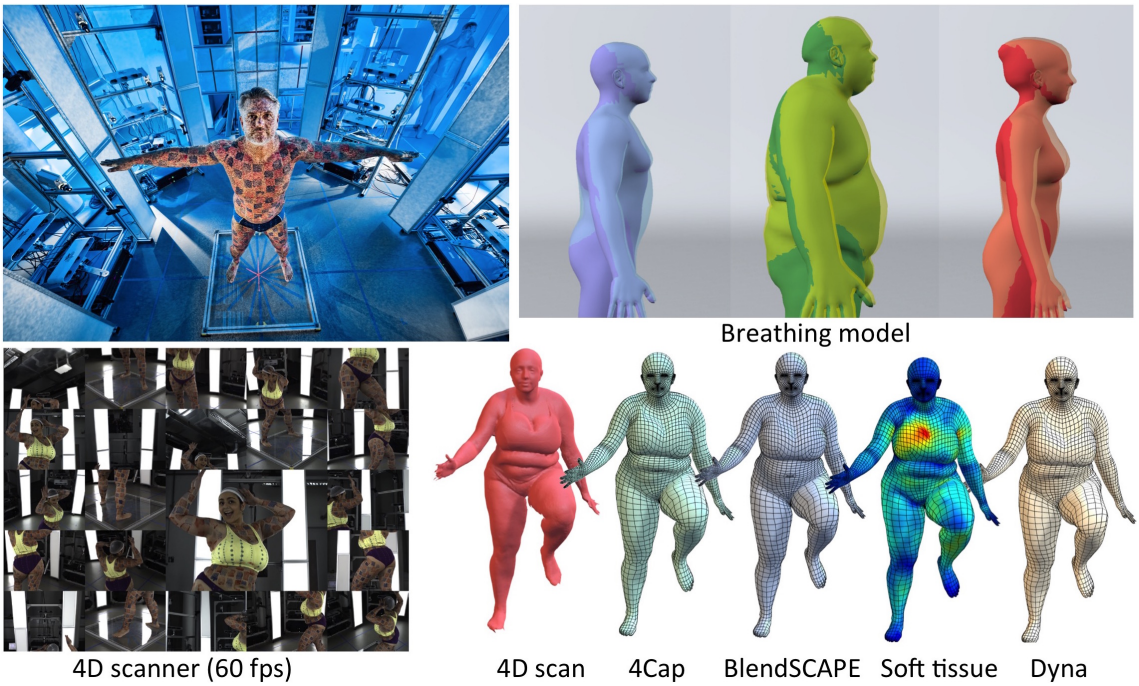


Figure 1.11: We use a novel 4D full-body scanner with 66 cameras to capture body shapes in motion. We register a template to these scans in a process called 4Cap (4D motion capture). Given non-rigid deformations from our standard body models, we learn statistics of soft tissue dynamics [12] or breathing deformations [18].

Human bodies are dynamic; they deform as they move, jiggle due to soft-tissue dynamics, and change shape with respiration. In [18] we learn a model of body shape deformations due to breathing for different breathing types and provide simple animation controls to render life-like breathing regardless of body shape. Using 3D scans of 58 human subjects, we augment a SCAPE model to include breathing shape change for different genders, body shapes, and breathing types.

Current 3D scanners capture only static bodies with high spatial resolution while mocap systems only capture a sparse set of 3D points at high temporal resolution. To better understand how people deform as they move, we need both high spatial and temporal resolution. To that end we commissioned the world's first 4D scanner that captures detailed full body shape at 60 frames per second. This 4D output, however, is simply a sequence of point clouds. To model the statistics of human shape in motion, we first register a common template mesh to each sequence in a process we call 4Cap [12].

Using over 40,000 registered meshes of ten subjects, we learn how soft- tissue motion causes mesh triangles to deform relative to a base 3D body model [12]. The resulting Dyna model uses a second-order auto-regressive model that predicts soft-tissue deformations based on previous deformations, the velocity and acceleration of the body, and the angular velocities and accelerations of the limbs. Dyna also models how deformations vary with a person's body mass index (BMI), producing different deformations for people with different shapes. We provide tools for animators to modify the deformations and apply them to new stylized characters. We have also ported this model to our vertex-based SMPL model [9].

More information: https://ps.is.tuebingen.mpg.de/project/4d-shape

Virtual Humans

Michael Black, Javier Romero, Matthew Loper, Gerard Pons-Moll, Naureen Mahmood, Federica Bogo

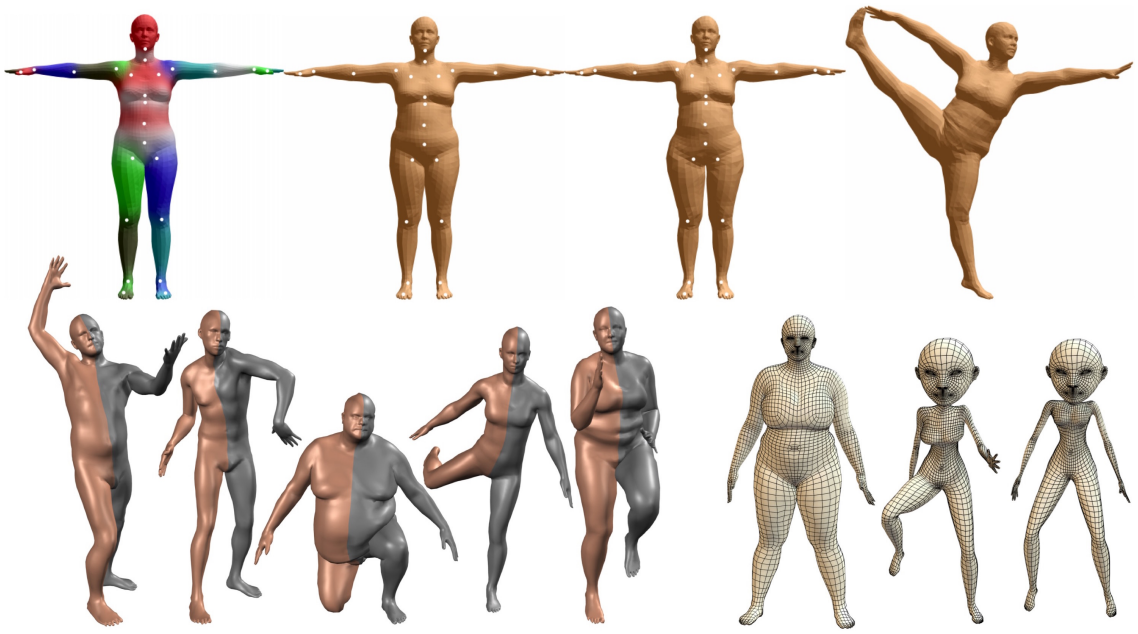


Figure 1.12: Top: The SMPL body model uses a template shape, blend weights, a function to predict joint locations from shape, shape blend shapes to change identity, pose blend shapes to correct for pose deformations, and dynamic blend shapes to capture soft-tissue dynamics. All these are learned from data and fit scan data more accurately than SCAPE (bottom left: scans in gray, SMPL in copper). (bottom right) We can retarget pose and soft-tissue dynamics to new characters.

The human body is certainly central to our lives and is commonly depicted in images and video. We are developing the world's most realistic models of the body by learning their shape and how they move from data. Our goal is to make 3D models of the body look and move in ways that make them indistinguishable from real humans. Such virtual humans can be used in special effects and will play an important role in emerging virtual reality systems. They can also be used in computer vision to generate training data for learning methods or can be fit directly to sensor data. What makes this hard is that the human body is highly articulated, deforms with kinematic changes, and exhibits large shape variability across subjects.

Over the last five years we have developed a series of 3D body models that can be used for both graphics and vision: BlendSCAPE [102], Delta [42], Dyna [12] and finally SMPL [9]. In particular, SMPL is a realistic human body model that is more accurate than previous SCAPE models yet is based on standard blend skinning and blend shapes. Pose blend shapes correct blend-skinning artifacts and are driven by elements of the body part rotation matrices. Shape blend shapes capture how body shape varies across people; these are computed using pose normalized 3D scans of 4000 people. Dynamic blend shapes capture how soft tissue deforms with motion.

The simplicity of our formulation means that SMPL can be trained from large amounts of data. It also means that it is compatible with current game engines and graphics software, running much faster than real time. The model is licensed commercially to Body Labs Inc. and is made freely available for research purposes.

More information: https://ps.is.tuebingen.mpg.de/project/virtual-humans

Part-based Body Models

Silvia Zuffi, Javier Romero, E. Sudderth, S. Ghosh, J. Pacheco, L. Sigal, Michael Black

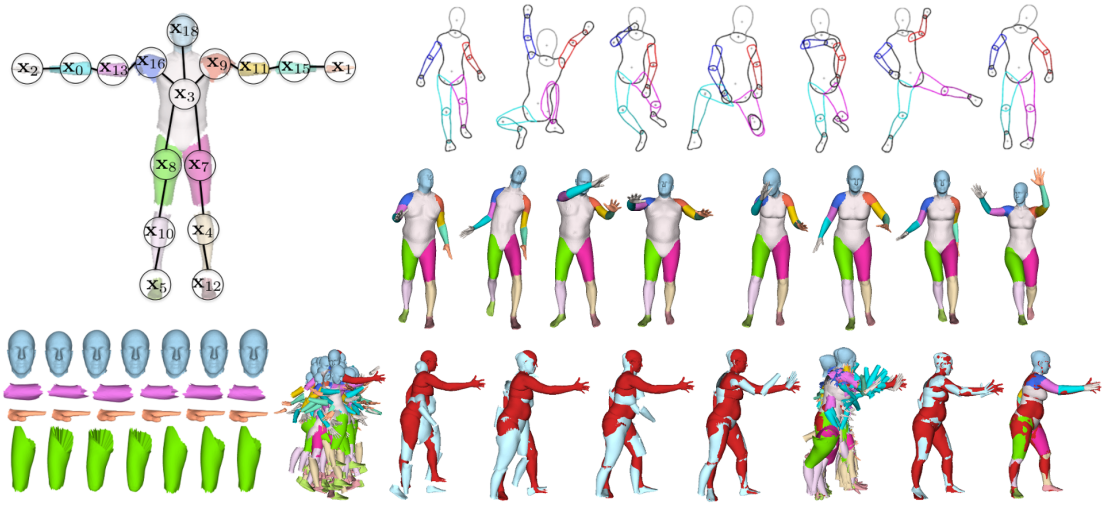


Figure 1.13: Left (top): Graphical structure of the body. (bottom) 3D parts. Right: 2D Deformable Structure model (top row), 3D Stitched Puppet model (middle row), model alignment to data exploiting the part-based representation (bottom row).

Human pose and shape estimation can be seen as a proxy for a wide range of problems in object representation and recognition. Humans are complex and articulated, appear in images in a variety of clothing, and come in a wide range of shapes. Teaching computers to understand people and their movements in images and videos is a great challenge of computer vision with manifold applications in entertainment, human-computer interaction, web search, medicine, and autonomous vehicles.

Most of the existing methods for human pose detection and tracking are based on part-based models, where the human body is represented as a set of "boxes" in two-dimensions (2D) or simple geometric primitives like cylinders or cones in three dimensions (3D) [35]. These models map to probabilistic generative models where each body part is represented with a node in a graph, and edges represent connections between parts. Efficient inference for the models' parameters given data can be performed with message passing algorithms.

Traditional part-based models cannot reach the level of realism of global models, as they do not represent body shape deformations with pose. Moreover they do not parameterize intrinsic body shape, and have been only used so far to estimate body pose.

We have introduced part-based models that are parameterized for body pose and shape. The Deformable Structures model (DS) [121] is a 2D model that is able to generate contours of human bodies with pose-dependent deformations. The Stitched Puppet model (SP) [54] is a 3D model that can generate body meshes with different pose and intrinsic shape, and realistic pose-dependent deformations. We have also learned the part segmenation from scans [101].

These models live in a higher dimensional space compared with models that do not represent shape. Furthermore, these shape parameters are represented by continuous random variables. To make inference practical in graphical models with high-dimensional continous parameters, we use a new particle-based belief propagation algorithm that mantains particle diversity [75].

More information: https://ps.is.tuebingen.mpg.de/project/part-based-body-models

Bodies from RGB-D

Federica Bogo, Javier Romero, Alex Weiss, David Hirshberg, Matthew Loper, Michael Black

Figure 1.14: We accurately estimate the 3D geometry and appearance of the human body from a monocular RGB-D sequence of a user moving freely in front of the sensor. Our approach proceeds in a coarse-to-fine manner. Given a monocular sequence (background), we estimate a low-dimensional parametric model of body shape (left), detailed 3D shape (middle), and a high-resolution texture map (right).

Accurate 3D body shape and appearance capture is useful for applications ranging from special effects, to fashion, to medicine. High-resolution scanners can capture human body shape and texture in great detail but these are bulky and expensive. In contrast, inexpensive RGB-D sensors are proliferating but are of much lower resolution. Scanning a full body from multiple partial views requires that the subject stands still or that the system precisely registers deforming point clouds captured from a non-rigid and articulated body.

We developed the first method to estimate human body shape from Kinect data [125, 129]. The approach fits a body model to depth and image silhouettes to estimate body shape and pose from static scans of a subject in one or more static poses. We have since improved this greatly and our latest method estimates body shape with the realism of a high-resolution body scanner by allowing a user to move freely in front of a single commodity RGB-D sensor [42]

To achieve this, we develop a new parametric 3D body model, Delta, that is based on SCAPE but contains several important innovations. First, we define a parametric shape model at multiple resolutions that enables the estimation of body shape and pose in a coarse-to-fine process. Second, we define a variable-detail shape model that models facial shape with higher detail than body shape; this is important for realistic avatars. Third, we combine a relatively-low polygon count mesh with a high-resolution displacement map to capture realistic shape details, and a high-resolution texture map estimated from the sequence.

We bring color and range data in each frame into alignment with our body model adopting a coarse-to-fine approach. The method exploits geometry and image texture over time to obtain accurate shape, pose, and appearance information despite unconstrained motion, partial views, varying resolution, occlusion, and soft tissue deformation.

Our recovered models are competitive with high-resolution scans from a professional 3D scanning system. Our system creates accurate 3D avatars from challenging motion sequences and even captures soft tissue dynamics.

More information: https://ps.is.tuebingen.mpg.de/project/bodies-from-rgbd

Body Perception

Betty Mohler, Stephan Streuber, Alejandra Quiros-Ramirez, Anne Thaler, Javier Romero, Michael Black

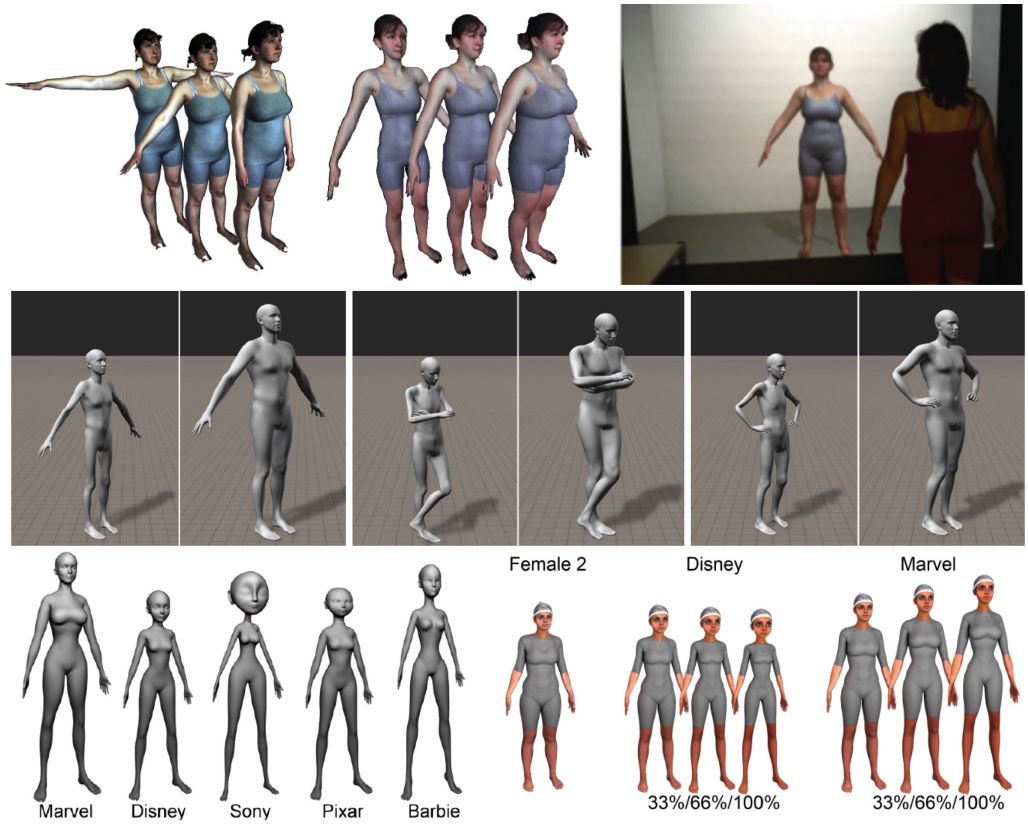


Figure 1.15: Top: Given scans of a person (left) we construct a 3D avatar and then change the shape of the avatar while keeping the identity fixed (middle). Subjects view their avatar in a virtual mirror and have to judge whether it is their body shape. Middle: We find that both pose and body shape effect the perception of social "power". Bottom: We study what makes avatars appealing by taking real bodies and making them more like cartoon characters. The most appealing bodies are neither fully real or fully cartoons.

We create virtual avatars from full body 3D scans and then manipulate body shape, pose, and appearance to create realistic stimuli for the study of the human perception of body shape.

We created personalized avatars and varied their weight to investigate the relative importance of visual cues (shape and texture) on the ability to accurately perceive own current body weight [17]. Participants perceived their body weight veridically when they saw their own photo-realistic texture and significantly underestimated their body weight when the avatar had a checkerboard texture.

Body shape and pose influence the perception of physical strength and social power of male virtual characters [46]. The perception of physical strength was mainly driven by the shape of the body, while the social attribute of power was influenced by an interaction between pose and shape. The effect of pose on power ratings was greater for weak body shapes; a character with a weak shape can be perceived as more powerful when in a high-power pose.

To study the "uncanny valley" we use cartoon body styles derived from popular characters and present a method to stylize the body shape and color of realistic avatars [39]. In perceptual studies we found that partially stylized body shapes result in increased perceived appeal. Avatars with high stylization or no stylization at all were rated to have the least appeal.

Our ongoing work is focused on body shape perception in patients with anorexia.

More information: https://ps.is.tuebingen.mpg.de/project/body-perception

Dense Optical Flow

Jonas Wulff, Laura Sevilla, Moritz Menze, D. Sun, Andreas Geiger, Michael Black

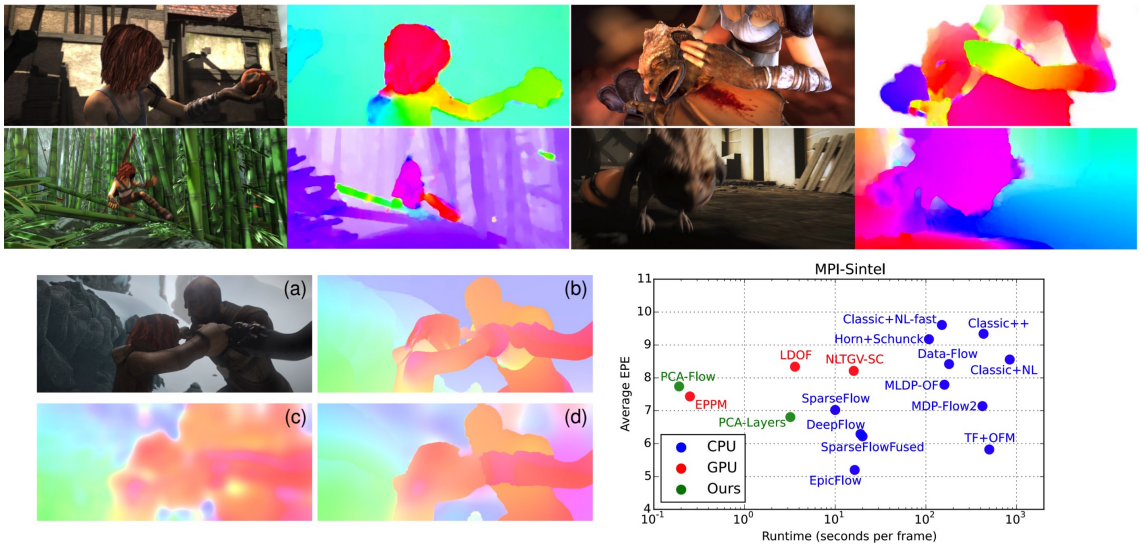


Figure 1.16: Top row: Results of discrete flow [49] on MPI-Sintel. Bottom row: PCA flow and PCA-layers [55] balance speed with accuracy, producing accurate flow efficiently without a GPU (right).

While the accuracy of optical flow estimation has increased markedly, a number of problems remain, most notably the treatment of motion and image boundaries, the tracking of fast but small/thin objects, and the computational complexity of current methods.

In [21], we comprehensively analyze the lessons the field has learned in the past decades. We systematically evaluate techniques to find out what really works. We observe that traditional formulations can achieve competitive performance when implmented using modern practices. Moreover we find that image-mediated spatial smoothing is critical to accuracy and formulate the prevailing ad hoc approach as a principled objective function.

In [64] we address the problem of small objects that move more than their own spatial extent. While large displacements are usually captured using an image pyramid, this blurs over small objects, making their motion untrackable. To address this we replace images with Distribution Fields, which allow the use of spatial pyramids to capture large motions while preserving high-frequency spatial detail, allowing us to track small objects over large displacements.

In [49] we use discrete optimization to estimate optical flow. As naive discretization of the 2D flow space is intractable, we investigate three different strategies, which are able to reduce computation and memory demands by several orders of magnitude. Their combination allows us to estimate large-displacement optical flow both accurately and efficiently, attaining state-of-the-art performance.

PCA-Flow [55] takes a non-standard approach to compute optical flow efficiently. Instead of modelling the motion of each pixel as a variable, we treat the full optical flow field as a datapoint in a 500-dimensional subspace, the structure of which is learned from 8 hours of movie data. Computing an optical flow field then becomes equivalent to finding a point in this subspace. This allows us to rapidly compute an approximate, smooth flow field, which can then serve as a building block for other applications, such as layered optical flow. PCA flow is the fastest non-GPU flow method obtaining peformance better than methods like Classic+NL and LDOF.

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

Layered Optical Flow

Jonas Wulff, D. Sun, Michael Black

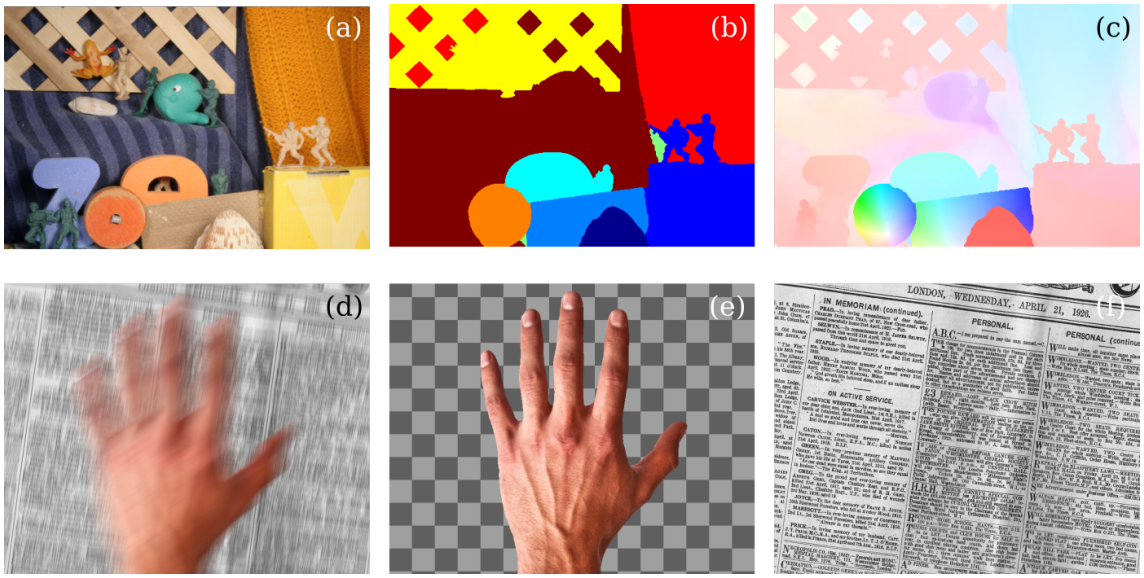


Figure 1.17: Top row: From a sequence of images (a), we extract the layer assignments (b) and compute highly accurate flow (c), especially at motion boundaries. Bottom row: Using a layered model, a motion-blurred sequence (d) can be decomposed into foreground (e) and background (f), which can then be separately deblurred.

Layered models allow scene segmentation and motion estimation to be formulated together and to inform one another. They separate the problem of enforcing spatial smoothness of motion within objects from the problem of estimating motion discontinuities at surface boundaries. Furthermore, layers define a depth ordering, allowing us to reason about occlusions.

In [120], we present an optical flow algorithm that segments the scene into layers, estimates the number of layers, and reason about their relative depth ordering using a novel discrete approximation of the continuous objective in terms of a sequence of depth-ordered MRFs and extended graph-cut optimization methods. We extend layer flow estimation over time, enforcing temporal coherence on the layer segmentation and show that this improves accuracy at motion boundaries.

In [98], we extend the layer segmentation algorithm using a densely connected Conditional Random Field. To segment the video, the CRF can use evidence from any location in the image, not just from the immediate surroundings of a pixel. Additionally, the CRF drastically reduces runtime of the segmentation step, while preserving the high fidelity at motion boundaries.

PCA-Layers [55] combines a layered approach with a fast, approximate optical flow algorithm. Within each layer, the optical flow is smooth and can be expressed using low spatial frequencies. Sharp discontinuities at surface boundaries, on the other hand, are captured by the layered formulation, and therefore do not need to be modeled in the spatial structure of the flow itself, allowing highly efficient layered flow computation.

We also use layered models in the treatment of motion blur [63]. In a dynamic scene, objects can move and occlude each other. Together with the nonzero shutter speed of the camera, this creates motion blur, which can be complex close to object boundaries; pixel values arise as a combination of foreground and background. Using a layered model allows us to separate overlapping layers from each other, making it possible to simultaneously segment the scene compute optical flow in the presence of motion blur, and deblur each layer independently.

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

High-level Priors

Andreas Geiger, Fatma Güney, Chaohui Wang

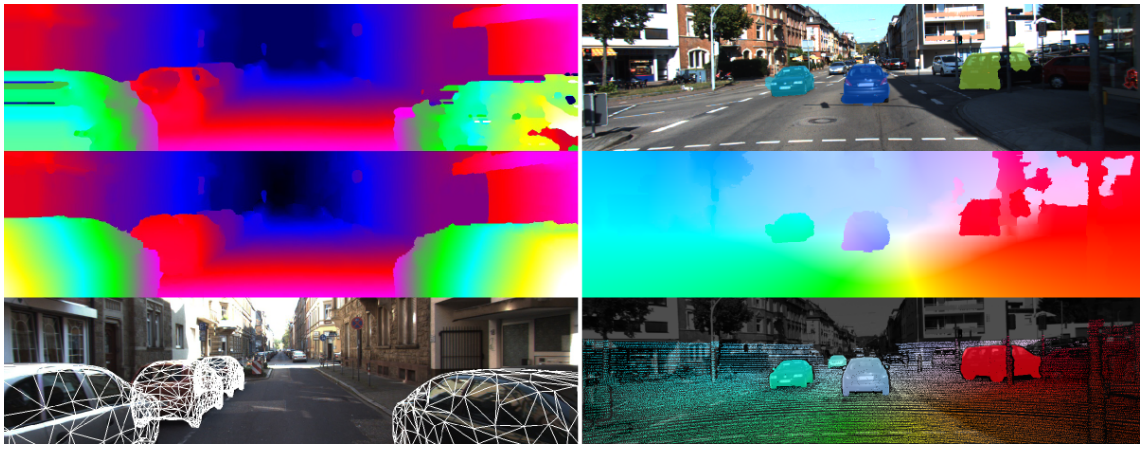


Figure 1.18: Left: Using object knowledge we are able to resolve stereo ambiguities, in particular at textureless and reflective surfaces. This leads to smoother and more accurate depth maps (middle) compared to using classical local regularizers (top). In additon we obtain highly detailed 3D object estimates (bottom) [58]. Right: Our 3D scene flow model decomposes the scene into its rigid components. This way, we simultaneously obtain a motion segmentation of the image (top) while implicitly regularizing the scene flow solution (middle) [52, 57].

While many computer vision problems are formulated as purely bottom-up processes, it is well known that top-down cues play an important role in human perception. But how can we integrate this high-level knowledge into current models? In this project, we investigate this question and propose models for stereo [58], scene flow [52, 57], and 3D scene understanding [50] which formulate our prior belief about the scene in terms of high-order random fields [25, 132]. We also tackle the aspect of tractable approximate inference which is particularly challenging for these kind of models.

Stereo techniques have witnessed tremendous progress over the last decades, yet some aspects of the problem remain challenging today. Striking examples are reflective and textureless surfaces which cannot easily be recovered using traditional local regularizers. In [58], we therefore propose to regularize over larger distances using object-category specific disparity proposals. Our model encodes the fact that objects of certain categories are not arbitrarily shaped but typically exhibit regular structures. We integrate this knowledge as non-local regularizer for the challenging object category "car" into a superpixel based CRF framework and demonstrate its benefits on the KITTI stereo evaluation.

In [52, 57], we propose a novel model and dataset for 3D scene flow estimation with an application to autonomous driving. Taking advantage of the fact that outdoor scenes often decompose into a small number of independently moving objects, we represent each element in the scene by its rigid motion parameters and each superpixel by a 3D plane as well as an index to the corresponding object. This minimal representation increases robustness and leads to a discrete-continuous CRF where the data term decomposes into pairwise potentials between superpixels and objects. Moreover, our model intrinsically segments the scene into its constituting dynamic components. We demonstrate the performance of our model on existing benchmarks as well as a novel realistic dataset with scene flow ground truth. We obtain this dataset by annotating 400 dynamic scenes from the KITTI raw data collection using detailed 3D CAD models for all vehicles in motion. Our experiments also reveal novel challenges which cannot be handled by existing methods.

More information: https://ps.is.tuebingen.mpg.de/project/high-level-priors

Intrinsic Properties of Scenes

Naejin Kong, Martin Kiefel, Peter Gehler, Michael Black

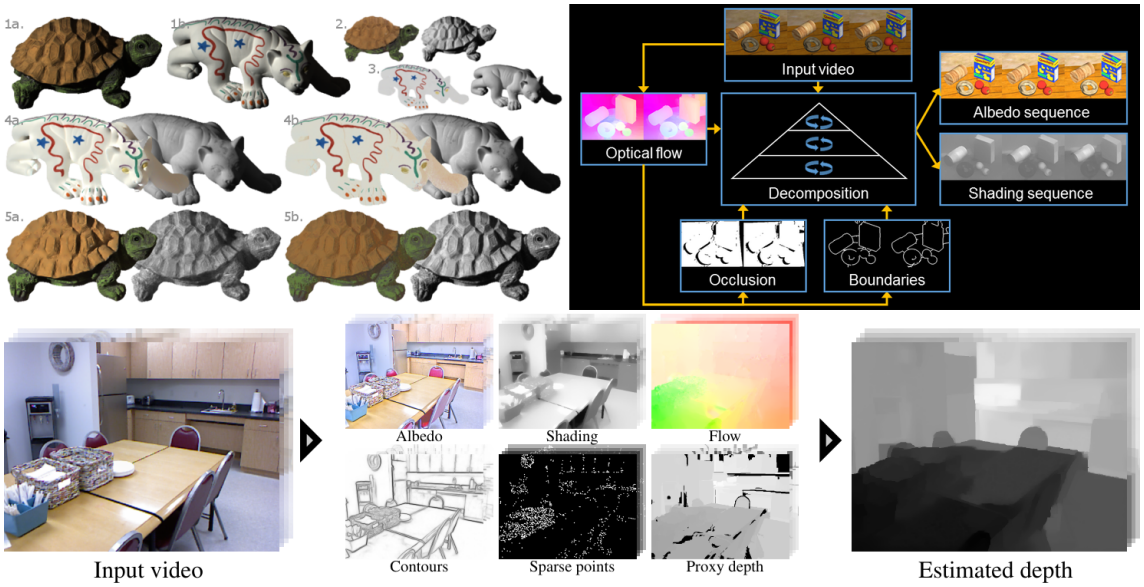


Figure 1.19: (Top left) Given a single image we decouple albedo and shading with a Global Sparsity prior on albedo. 1a-b: input images, 2,3: ground truth, 4a-b,5a-b: estimated albedo and shading with different settings a and b. (Top right) We extract temporally coherent albedo and shading sequences from video alone by exploiting physical properties derived from temporal variation in the video. (Bottom) We formulate the estimation of dense depth maps from video sequences as a problem of intrinsic image estimation.

Intrinsic images correspond to physical properties of the scene. It is a long-standing hypothesis that these fundamental scene properties provide a foundation for scene interpretation.

To decouple albedo and shading given a single image, we introduce a novel prior on albedo, that models albedo values as being drawn from a sparse set of basis colors [122]. This results in a Random Field model with global, latent variables (basis colors) and pixel-accurate output albedo values. We show that without edge information high-quality results can be achieved, that are on par with methods exploiting this source of information. Finally, we can improve on state-of-the-art results by integrating edge information into our model.

While today intrinsic images are typically taken to mean albedo and shading, the original meaning includes additional images related to object shape, such as surface boundaries, occluding regions, and depth. By using sequences of images, rather than static images, we extract a richer set of intrinsic images that include: albedo, shading, optical flow, occlusion regions, and motion boundaries. Intrinsic Video [67] estimates temporally coherent albedo and shading sequences from video by exploiting the fact that albedo is constant over time while shading changes slowly. The approach makes only weak assumptions about the scene and substantially outperforms existing single-frame intrinsic image methods on complex video sequences.

Intrinsic Depth [43] steps towards a more integrated treatment of intrinsic images. Our approach synergistically integrates the estimation of multiple intrinsic images including albedo, shading, optical flow, surface contours, and depth. We build upon an example-based framework for depth estimation that uses label transfer from a database of RGB and depth pairs. We also integrate sparse structure from motion to improve the metric accuracy of the estimated depth. We find that combining the estimation of multiple intrinsic images improves depth estimation relative to the baseline method.

More information: https://ps.is.tuebingen.mpg.de/project/scene-intrinsics-from-x

Inverse Graphics

Matthew Loper, Varun Jampani, Peter Gehler, Michael Black

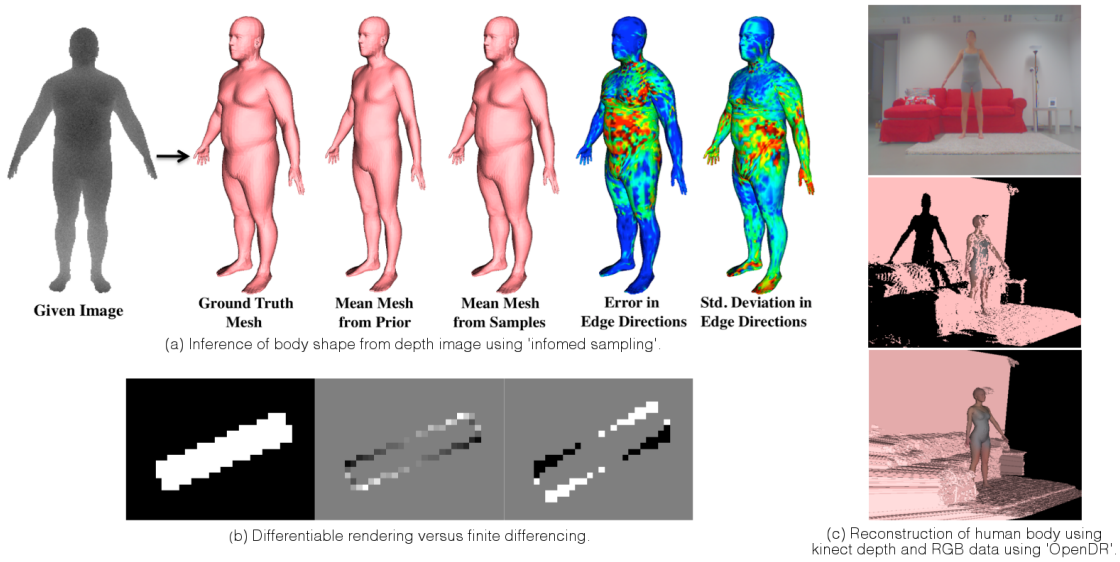


Figure 1.20: (a) 3D mesh reconstruction with the first 1000 samples obtained using 'informed sampling'. (b) Left: a rotating quadrilateral. Middle: OpenDR's predicted change in pixel values with respect to in-plane rotation. Right: corresponding finite differences. (c) Top: captured image. Middle: captured point cloud together with estimated body model. Bottom: estimated body shown on background point cloud.

A long standing and conceptually elegant view of computer vision is to use a generative model of the physical image formation process and posterior inference to infer or explain the image observations. A key problem in this inverse graphics view is the difficulty of posterior inference at run time. This difficulty stems from a number of causes: (1) high-dimensionality of the posterior, (2) complex and dynamic dependency between model parameters and (3) the forward graphics simulations being expensive. We address these issues in terms of local and global optimization.

For local optimization, we propose an approximate differentiable renderer (DR) [66] that explicitly models the relationship between changes in model parameters and image observations. The OpenDR framework makes it easy to express a forward graphics model and then automatically obtain derivatives with respect to the model parameters and to optimize over them. Built on a new auto-differentiation package and OpenGL, OpenDR provides a local optimization method that can be incorporated into probabilistic programming frameworks. We demonstrate the power and simplicity of programming with OpenDR by using it to solve the problem of estimating human body shape from Kinect depth and RGB data.

To address issues of more global optimization, we also propose the informed sampler [61] that leverages computer vision features and algorithms to make informed proposals for the state of latent variables. These proposals are accepted or rejected based on the generative graphics model. The informed sampler is simple and easy to implement, yet it enables inference in generative models that were out of reach for current uninformed samplers. We demonstrate this claim on challenging models that incorporate rendering engines, object occlusion, ill-posedness, and multi-modality.

More information: https://ps.is.tuebingen.mpg.de/project/inverse-graphics-proj

Scene Understanding

Andreas Geiger, Peter Gehler, Varun Jampani, Chaohui Wang

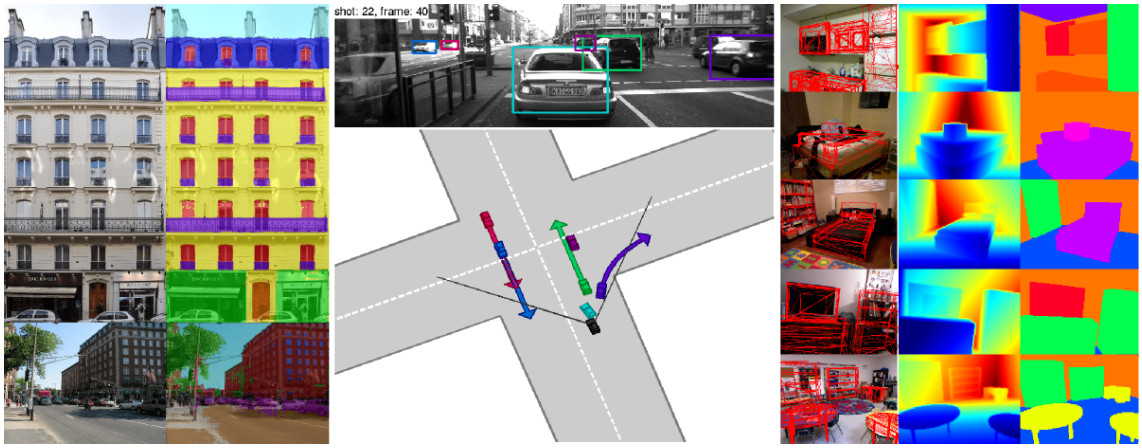


Figure 1.21: Left: Facade parsing using auto-context [60]. Center: Recovering 3D urban scene layout while estimating and associating all objects in the scene [20, 92]. Right: Indoor scene understanding from a single RGB-D image using 3D CAD priors [50].

Holistic scene understanding is an important prerequisite for many indoor and outdoor applications, including autonomous driving, navigation, indoor and outdoor mapping as well as localization. Given a high-dimension input (e.g., image or video stream), the task is to extract a rich but compact representation that is easily accessible to subsequent processing stages. Typical outputs comprise semantic information [60, 115] or 3D information about the shape and pose of objects and layout elements in the scene [20, 50, 92]

In [20, 92], we present novel probabilistic generative models for multi-object traffic scene understanding from movable platforms which reason jointly about the 3D scene layout as well as the location and orientation of objects in the scene. In particular, the scene topology, geometry and traffic activities are inferred from short video sequences. Inspired by human driving capabilities, our models do not rely on GPS, lidar or map knowledge. Instead, we take advantage of a diverse set of visual cues in the form of vehicle tracklets, vanishing points, semantic scene labels, scene flow and occupancy grids. Our approach successfully infers the correct layout in experiments on varied videos of 113 challenging intersections.

In [50], we propose a model which infers 3D objects and the layout of indoor scenes from a single RGB-D image captured with a Kinect camera. In contrast to existing holistic scene understanding approaches, our model leverages detailed 3D geometry using inverse graphics and explicitly enforces occlusion and visibility constraints for respecting scene properties and projective geometry. We cast the task as MAP inference in a high-order conditional random field which we solve efficiently using message passing. Our experiments demonstrate that the proposed method is able to infer scenes with a large degree of clutter and occlusions.

In [60], we propose a system for the problem of facade segmentation. Building facades are highly structured images and consequently most methods that have been proposed for this problem, aim to make use of this strong prior information. In this work, we propose a system which is almost domain independent and consists of standard segmentation methods. A sequence of boosted decision trees is stacked using auto-context features. We find that this, albeit standard, technique performs better, or equals, all previous published empirical results on all available facade benchmark datasets.

More information: https://ps.is.tuebingen.mpg.de/project/scene-understanding

3D Recognition

Peter Gehler, M. Stark, B. Schiele, B. Pepik

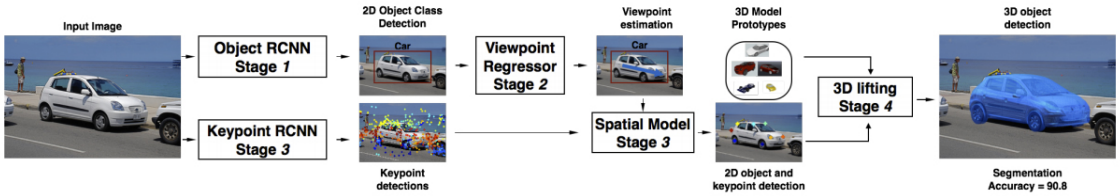


Figure 1.22: An overview of our 3D object detection model. This model localizes objects and infers their articulation and 3D shape from single static images. Objects of different categories can be detected in cluttered environments.

The ability to recognize and categorize objects in any type of visual scene is an integral part of scene recognition systems. While for constrained scenarios, like face detection, this problem has largely been solved, the general case of recognizing any kind of object in real world and cluttered environments remains an open research problem. Many different factors contribute to the complexity of this problem. A main complicating factor is in images one has access to 2D projections of objects, whereas they really are three dimensional physical objects. This projection leads to significant ambiguity in object appearance. Therefore the predominant paradigm today, is to largely ignore the 3D structure, and attack object class recognition using 2D feature-based models.

In contrast to this we investigate models that take into account the 3D structure of objects. We believe that building models that encode the three dimensional origin of real world objects has several benefits. It leads to more compact computational models that therefore will need less training data during a training phase. Further, the output of 3D object detection systems will enable richer reasoning about entire visual scenes. A simple bounding box around objects of interest may be sufficient for counting and coarse localizing of objects. However, we aim to recover articulations and the 3D extent of objects together with a precise localization and categorization. This richer information is needed in order to extract knowledge about object compositions in a scene as a whole.

In the last years we have made progress towards this goal. The work [8, 114, 119] proposes methods that perform 3D bounding box detection from 2D images. We extend a state-of-the-art model, namely the Deformable Parts Model (DPM) from Felzenszwalb et al., to a full 3D object model. The DPM is a mixture of star based CRF models that include deformation and appearance terms. We propose a CRF that models an object directly in 3D and that can be evaluated using any image projection. Then, for novel unseen images, the object identity, its localization and the projection from 3D to 2D is reasoned about. Since most benchmarks promote object detection as a 2D detection task, and we use only 2D training data, we use CAD models to inform our detection model about the 3D structure of objects. This enables reasoning about occlusion [96] and transfer learning of geometric information across different object instances [80].

In [53] we extended our work into a system that is capable of extracting detailed CAD models from unconstrained images. We combine several estimation steps that infer viewpoint, object identity and position into a coherent system that results in very fine grained and detailed hypotheses about the objects present in the scene. Following careful design, in each stage the method constantly improves the performance and achieves state-of-the-art performance in simultaneous 2D bounding box and viewpoint estimation on the challenging Pascal3D+ dataset.

More information: https://ps.is.tuebingen.mpg.de/project/objects-in-3d

Object Detection

Jürgen Gall, Abhilash Srikantha, N. Razavi, L. Van Gool

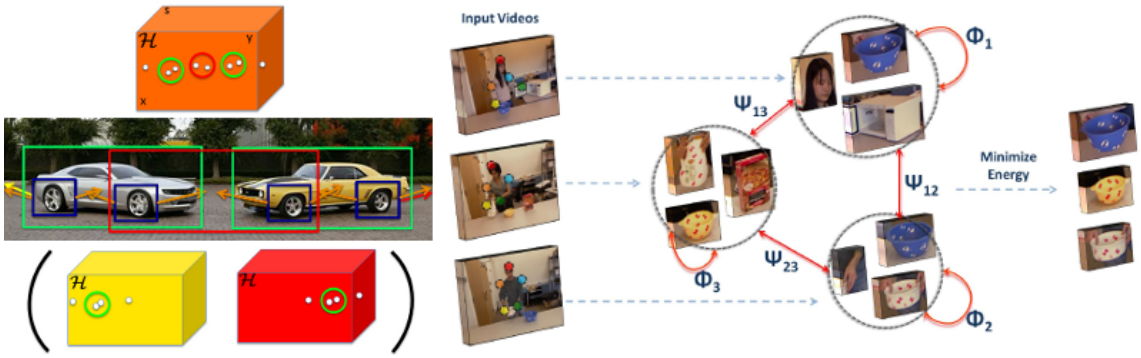


Figure 1.23: Left: During detection, Hough forests cast weighted votes to a Hough space (orange) where objects are detected by localizing modes. Improved performance can be realized by using latent Hough spaces thereby relaxing the patch independece criterion. Right: Instances of common objects in videos are discovered by defining a model that encodes similarity of their appearance and functionality.

Object detection for real world applications is still a challenging problem. While increased data can partly solve the problem, the ability of detectors to process large data sets in reasonable time becomes another important issue besides accuracy.

A family of methods that can handle large amount of training data efficiently, and that are inherently suited for multi-class problems, are based on random forests, which are ensembles of randomized decision trees that can be applied to regression or classification tasks. Since object detection involves both classifying patches belonging to an object and using them to regress the location and scale of the object, random forests for object detection need to be trained to satisfy both objectives [128].

While object detection based on Hough forests allows parts observed in different training instances to support a single object hypothesis, it also produces false positives by accumulating votes that are consistent in location but inconsistent in other properties like pose, color, shape or type. To address this problem, Hough forests can be augmented with latent variables in order to enforce consistency among votes [113]. To this end, only votes that agree on the assignment of the latent variable are allowed to support a single hypothesis.

In order to avoid an expensive manual labeling process, or to learn object classes autonomously without human intervention, we propose a framework for object discovery in activity-labeled videos [71]. Since small objects like pens are difficult to discover only based on appearance, we introduce similarity based on object functionality, which can be estimated from relative human-object motion during the activity. We show that functionality is an important cue for discovering objects from activities in RGB(D) video datasets.

More information: https://ps.is.tuebingen.mpg.de/project/object-detection

Computer Vision Performance Evaluation

Andreas Geiger, Jonas Wulff, Moritz Menze, Daniel Butler, Michael Black

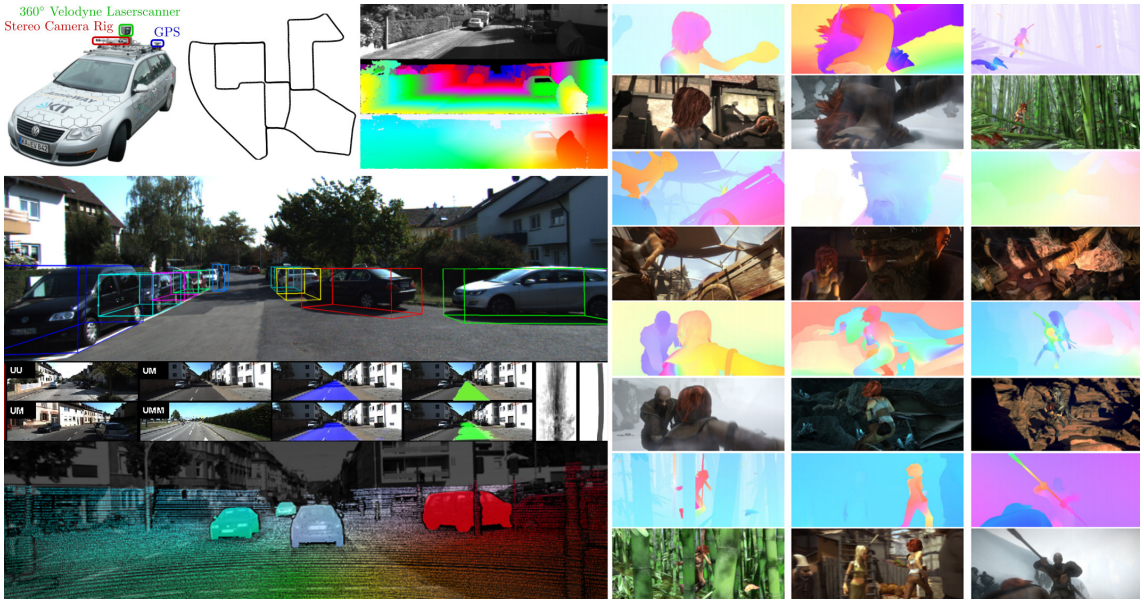


Figure 1.24: The KITTI Vision Benchmark Suite (left) and the MPI Sintel Benchmark (right) provide ground truth data and evaluation servers for benchmarking vision algorithms. So far, more than 400 methods have been evaluated on our benchmarks.

While ground truth datasets spur innovation, many current datasets for evaluating stereo, optical flow, scene flow and other tasks are restricted in terms of size, complexity, and diversity, making it difficult to train and test on realistic data. For example, we co-authored the Middlebury flow dataset [38], which arguably set a standard for the field but was limited in terms of complexity.

In [26, 57], we took advantage of an autonomous driving platform to develop challenging real-world benchmarks for stereo, optical flow, scene flow, visual odometry/SLAM, 3D object detection, 3D tracking and road/lane detection. Accurate ground truth is provided by a Velodyne laser scanner and a GPS localization system. Our datasets are captured by driving around a mid-size city of Karlsruhe, in rural areas and on highways with up to 15 cars and 30 pedestrians visible per image. For each of our benchmarks, we also provide a set of evaluation metrics and a server for evaluating results on the test set. Our experiments showed that moving outside the laboratory to the real world was critical. We continue to develop new ground truth to push the field further.

In [106, 107], we proposed a novel optical flow, stereo and scene flow data set derived from the open source 3D animated short film Sintel. We extracted 35 sequences displaying different environments, characters/objects, and actions and showed that the image and motion statistics of Sintel are similar to natural movies. Using the 3D source data, we created an optical flow data set exhibits important features not present in previous datasets: long sequences, large motions, non-rigidly moving objects, specular reflections, motion blur, defocus blur, and atmospheric effects. We released the ground truth optical flow for 23 training sequences and withheld the remaining 12 sequences for evaluation purposes. When released in 2012, the best methods had an average endpoint error of around 10 pixels. The dataset has focused the community on core problems and only 3.5 years later, there are over 70 methods evaluated on the benchmark with the best methods are approaching 5 pixels in error.

More information: https://ps.is.tuebingen.mpg.de/project/datasets

Human Pose, Shape and Action

Michael Black, Peter Gehler, Javier Romero, Federica Bogo, Silvia Zuffi, Hueihan Jhuang, Matthew Loper, Jürgen Gall, Cordelia Schmid

Figure 1.25: We propose novel challenging datasets for human pose estimation, 3D mesh registration and action recognition: a) MPII Human Pose, including around 25000 images of over 40000 people with annotated 2D body joints; b) FAUST, collecting 300 real human body scans with automatically computed ground-truth correspondences; c) J-HMDB, a dataset for action recognition with annotated human joints, segmentation, and optical flow.

Human pose estimation, 3D mesh registration and action recognition techniques have made significant progress during the last years. However, most existing datasets to evaluate them are inadequate for capturing the challenges of real-world scenarios. We introduce novel datasets and benchmarks, all publicly available for research purposes.

In [131], we describe the datasets currently available for pose estimation and the performance of state-of-the-art methods on them. In [81], we introduce a novel benchmark for pose estimation, "MPII Human Pose", that makes a significant advance with respect to previous work in terms of diversity and difficulty. It includes around 25000 images containing over 40000 people performing more than 400 different activities. We provide a rich set of labels including body joint positions, occlusion labels, and activity labels. Given these rich annotations we perform a detailed analysis of the leading human pose estimation approaches, gaining insights for the successes and failures of these methods.

FAUST [77] is the first dataset for 3D mesh registration providing both real data (300 human body scans of different people in a wide range of poses) and automatically computed ground-truth correspondences between them. We define a benchmark on FAUST, and find that current shape registration methods have trouble with this real-world data.

With the "Joints for the HMDB" dataset (J-HMDB) we focus on action recognition [87]. We annotate complex videos using a 2D "puppet" body model to obtain "ground truth" joint locations as well as optical flow and segmentation. We evaluate current methods using this dataset by systematically replacing the input to various algorithms with ground truth. This enables us to discover what is important – e.g., should we improve flow algorithms, or enable pose estimation? We find that high-level pose features greatly outperform low/mid level features; in particular, pose over time is critical. Our analysis and the J-HMDB dataset should facilitate a deeper understanding of action recognition algorithms.

More information: https://ps.is.tuebingen.mpg.de/project/evaluating-humans

Neural Prosthetics and Decoding

Michael Black, J. Donoghue, J. Simeral, S.-P. Kim, L. Hochberg, M. Homer, C. Vargas-Irwin

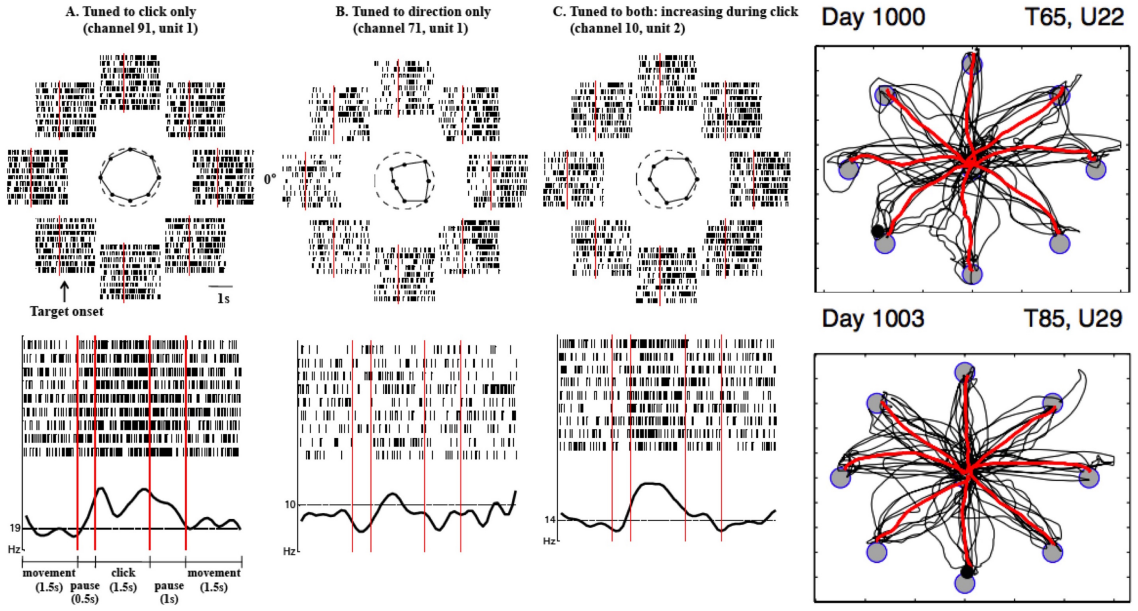


Figure 1.26: Left top (A-C): Firing of three different neurons in the brain of a paralyzed human. B and C show directional tuning. Left bottom (A-C): Firing rate of the same neurons as a function of imagined "clicking". A and C modulate with clicking. Right: Decoded trajectories in a center-out task from a population of motor cortical neurons in a human at around 1000 days after implantation.

We use motion capture together with electrode arrays, implanted in the motor cortex of monkeys, to learn how motor cortical activity relates to movement and to create new algorithms to decode this activity. Translating these models to paralyzed humans allows us to restore or improve lost function in people with central nervous system injury by directly coupling brains with computers, allowing people to control a computer cursor with their thoughts.

We developed a point-and-click intracortical Brain Computer Interface (iBCI) that enables humans with tetraplegia to volitionally move a 2D computer cursor in any desired direction on a computer screen, hold it still, and click on an area of interest [37]. This direct brain-computer interface extracts both discrete (click) and continuous (cursor velocity) signals from a single small population of neurons in human motor cortex. Enabling this is a multi-state probabilistic decoding algorithm that simultaneously decodes neural spiking activity and outputs either a click signal or the velocity of the cursor. The algorithm combines a linear classifier, which determines whether the user is intending to click or move the cursor, with a Kalman filter that translates the neural population activity into cursor velocity. We present a paradigm for training the multi-state decoding algorithm using neural activity observed during imagined actions. We quantified point-and-click performance using various human-computer interaction measurements for pointing devices. We found that participants could control the cursor motion and click on specified targets, suggesting that signals from a small ensemble of motor cortical neurons (40) can be used for natural point-and-click 2D cursor control of a personal computer. Furthermore in [36] we showed that such devices could be used to decode intended cursor movement over 1000 days after implantation.

Our ongoing work focuses on developing new non-linear decoding algorithms [14, 19] and on analyzing the motor control of grasping in non-human primates [10].

More information: https://ps.is.tuebingen.mpg.de/project/neural-prosthetics

Markerless Animal Motion Capture

Michael Black, Oren Freifeld, K. Shenoy, P. Nuyujukian, J. Foster

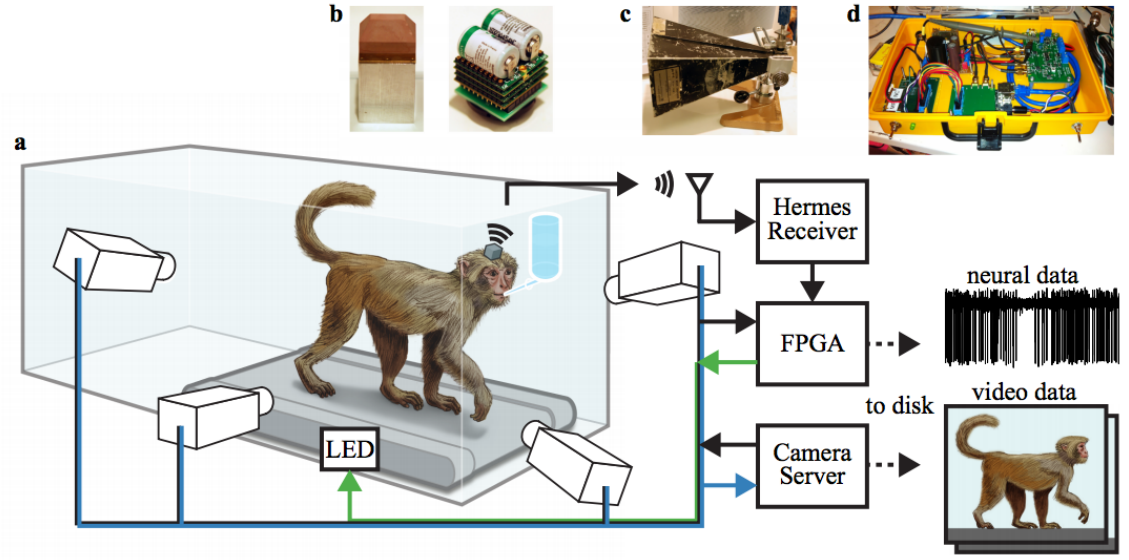


Figure 1.27: Markerless behavior capture. (a) Unconstrained behavior is recorded synchronously with video cameras while broadband neural activity is recorded and transmitted wirelessly. Images of the Stanford HermesE (b) transmitter and recording electronics, (c) receiving antenna, and (d) receiver and FPGA.

Experiments in motor neurophysiology often involve animals performing repeated actions. Stereotyped and practiced actions facilitate data analysis by allowing the experimenter to average neural firing activity across multiple trials. Does such activity, however, reflect what happens during natural, unconstrained, and spontaneous movement? Do models of neural activity developed in such settings translate to intracortical brain-machine interfaces (iBMIs) where humans need to control devices in a dynamically changing context?

To answer these questions we must record natural behavior and the corresponding neural activity [16, 116, 124]. For the latter, our collaborators at Stanford have developed an implantable device that enables the wireless recording of spiking activity from many neurons at once (see Figure). For the former, traditional marker-based motion capture techniques are impractical for full-body animal tracking. Consequently we exploit our work on human tracking to enable markerless articulated animal tracking.

Our system transmits neural activity from an intracortical array using a head-mounted device and records behavior using multi-camera markerless motion capture. We demonstrate this with the first recordings from motor cortex of rhesus monkeys walking quadrupedally on a treadmill. We find that multi-unit threshold-crossings encode the phase of walking and that the average firing rate covaries with the speed of individual steps.

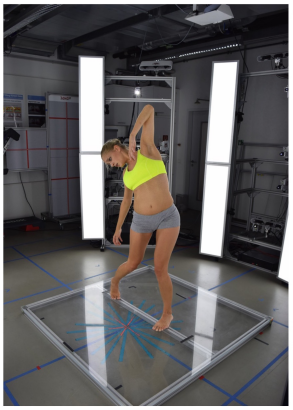
Freely-moving animal models may allow neuroscientists to examine a wider range of behaviors and can provide a flexible experimental paradigm for examining the neural mechanisms that underlie movement generation across behaviors and environments. For iBMIs, freely-moving animal models have the potential to aid prosthetic design by enabling the study of how neural encoding changes with posture, environment, and other real-world context changes. Understanding behavior in more naturalistic settings is essential for overall progress of basic motor neuroscience and for the successful translation of BMIs to people with paralysis.

More information: https://ps.is.tuebingen.mpg.de/project/markerless-animal-mocap

Artist-in-Residence Program (AIR)

Emma-Jayne Holderness

The artist in residence program (AIR) in Perceiving Systems matches artists with researchers who often use similar media, though in different contexts. Bringing artists into the scientific environment creates surprising synergies by juxtaposing different viewpoints and methodologies, enabling researchers to see their work in a different light, sparking creative thinking, and ultimately leading to innovative ideas.



Lilla and Bill Outcault, Sept. 2015. Lilla and Bill were the first artists to use 3D body scanners as an artistic tool. They produce marionettes or avatars that are anonymous yet universal figures that portray the concept of human frailty. Their project with Perceiving Systems involved six dancers with individual choreographies that express feelings of emotion through the body including anxiety, fear, panic and insanity.

http://locurto-outcault.com

Helga Griffiths, Feb. 2016. Helga has won multiple prizes and grants and her work, which has been widely exhibited internationally, often builds on collaborations with scientists. Most of Helga's work involves the integration of various sensory stimuli, producing "multi-sense" installations that play with the boundaries of conventional perception.

Helga's current project, called "negative space," involves data capture of identical twin dancer choreographies in our 4D scanner. Helga is interested in the space created between the dancers and what this space means about the relationship between them.

http://www.helgagriffiths.de

Peter Evers. Peter Evers' work explores the relationship between human identity and technology. His exhibition "Irresistible drift" contains video and motion capture sequences from Perceiving Systems and was exhibited at Belfast Exposed. The work shows a human subject instrumented for performance capture against a green-screen backdrop. He sees the subject as reduced to a puppet by the technology their every movement recorded for research. Research on body shape in Perceiving Systems can be seen as the most advanced approach for "capturing" the image and identity of people. This brings up issues of what is private and what is public? How will our avatars behave and represent us in the digital world? Can they be appropriated by others?

http://www.belfastexposed.org/exhibition/Irresistible_Drift

More information: https://ps.is.tuebingen.mpg.de/project/jobs/art

1.3 Equipment

4D Scanner. Human bodies and clothing deform in complex ways and exhibit interesting dynamics. To study and model 3D shapes in motion we commissioned a unique 4D scanner that captures the full 3D human body shape at 60 frames per second. Built for us by 3dMD (Atlanta, GA), the system uses 22 pairs of stereo cameras, 22 color cameras, and custom speckle projectors. The speckle patterns allow accurate stereo reconstruction of 3D shape. This speckle pattern alternates at 120fps with large white-light LED panels that provide a smooth nearly uniform illumination. Each frame in the "4D movie" is a 3D mesh with approximately 150,000 points.

This facility enables us to study body shape in motion, understand how bodies deform, and to capture the dynamics of soft tissue motions at high spatial and temporal resolution. With the world's first true 4D body scanner we have captured bodies in motion in ways that have never been seen before. Our custom protocols and software are providing new insights into body shape and motion for graphics, medicine, psychology, and computer vision.

With this scanner we have captured hundreds of thousands of 3D scans. Processing this data to make it useful involves registering a 3D template mesh to all the scans in a process we call 4Cap, for 4D motion capture. Our current research focuses on extending these techniques to accurately capture the 4D motions of clothing on the body.

3D Scanner. We also capture static 3D scans of bodies, faces, hand and other objects. In our Body Lab, we house a full body 3D stereo capture system, custom built by 3dMD. At the time of installation, it was both the largest and highest resolution system of its kind in the world. The aluminium frame has 22 modular, medical standard scanning units. Each of those contains a pair of stereo cameras for computing shape and geometry, utilizing either one or two speckle projector units and a single 5 megapixel colour camera for capturing texture. A series of flash units illuminate the subject during capture. Five Windows based PCs control the camera pods and communicate over gigabit ethernet and internally via Firewire 800. Capture takes a fraction of a second, and the system can fire several times per minute.

The system is normally configured to capture a 3D scan of a full human body, in a full range of poses. Since the capture time is so fast (2ms), fast motions such as jumping or skipping can be accurately frozen in time. To date we've taken over 29,000 scans with this system, and have ongoing body shape trials.

Motion Capture. The department also uses traditional motion capture systems. We share the use of a 12 camera T10-series VICON tracking system with the Autonomous Motion department. We also have an Optitrack system that is synchronized with our 4D capture system, enabling us to precisely evaluate how the systems compare in terms of motion and shape recovery. We also collaborate with researchers in Biological Cybernetics, where they have extensive motion capture facilities including IMU-based Xsens suits, a Vicon system, and various other sensing and display technologies.

Video Capture. Our video capture facility supports the development of new markerless motion and shape capture systems. The facility contains a 4DViews system with 12 synchronized 4 megapixel cameras that capture video at full resolution at nearly 60fps, or at lower resolutions at up to 200fps. A series of Linux PCs drive the cameras over gigabit ethernet. The cameras themselves are mounted on a rigid frame or on high quality Manfrotto^{TM} tripods and magic arms, to allow complete freedom of configuration within the lab itself. Four large fluorescent lamps provide static, consistent lighting to the performance space, which is surrounded by a green-screen curtain. 4DViews provides an extensive software suite for viewing the raw camera footage and managing the whole system, and a local, high performance NAS appliance serves 40TB of online storage.

1.4 Awards & Honors

2016

Cordelia Schmid, Humboldt Research Award.

2015

The **Middlebury Dataset** was awarded the 2015 IEEE Mark Everingham Prize for service and contributions to the computer vision community. Michael Black was part of the team behind the optical flow benchmark.

Jonas Wulff: Outstanding Reviewer Award, Int. Conf. on Computer Vision (ICCV).

Ali Osman Ulusoy, Andreas Geiger, and Michael J. Black: Best Paper Award, International Conference on 3D Vision (3DV), 2015, for the paper "Towards Probabilistic Volumetric Reconstruction using Ray Potentials."

Andreas Geiger and Chaohui Wang: 2015 Best Paper Award, German Conference on Pattern Recognition (GCPR), for the paper “Joint 3D Object and Layout Inference from a single RGB-D Image.”

Federica Bogo: Third prize at the Science2Start business plan competition.

Michael J. Black: elected foreign member of the *Royal Swedish Academy of Sciences,* class for engineering sciences.

Andreas Geiger: KIT Doctoral Award for his PhD thesis on “Probabilistic Models for 3D Urban Scene Understanding from Movable Platforms.”

2014

Andreas Geiger: Ernst-Schoemperlen-Prize for his PhD thesis, awarded by the KIT Center for Mobility Systems

2013

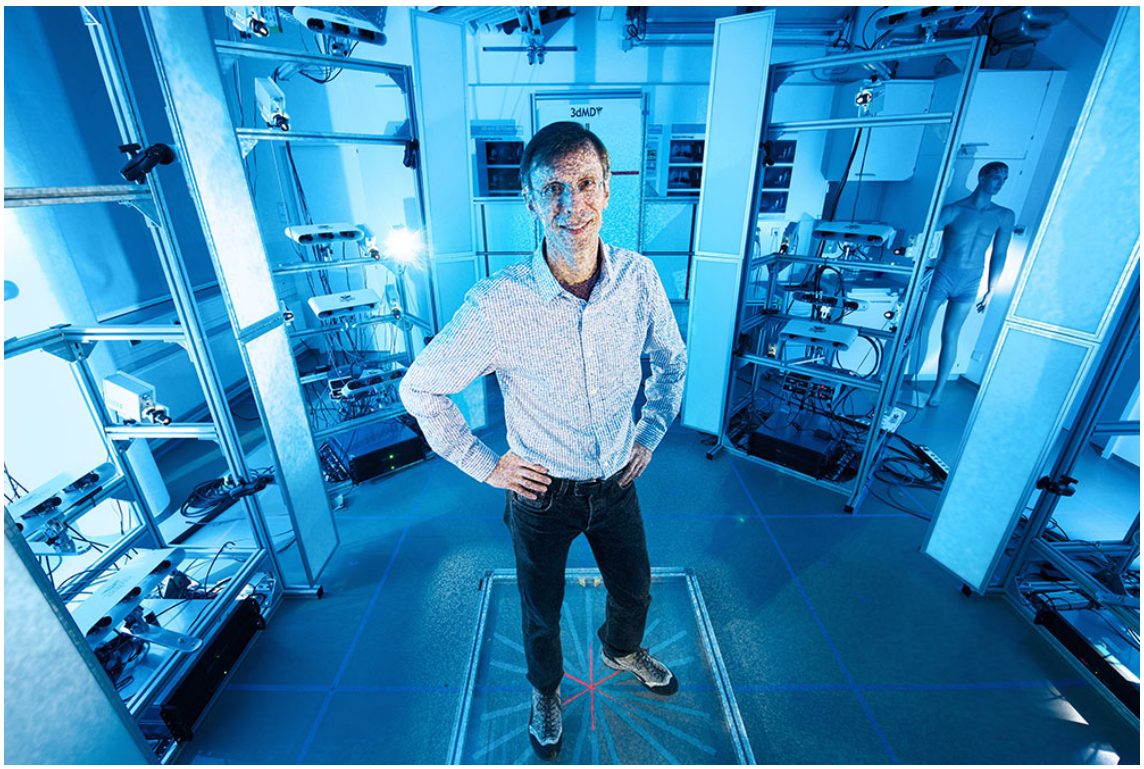
Michael J. Black: *Helmholtz Prize* for work that has stood the test of time; for the paper: Black, M. J., and Anandan, P., “A framework for the robust estimation of optical flow,” IEEE International Conference on Computer Vision, ICCV, pages 231-236, Berlin, Germany. May 1993.

Gerard Pons-Moll: Best Science Paper Award at the British Machine Vision Conference (BMVC) 2013 for the paper “Metric Regression Forests for Human Pose Estimation,” with J. Taylor, J. Shotton, A. Hertzmann, and A. Fitzgibbon.

Andreas Geiger: Best Paper Runner-Up Award at CVPR 2013 for the paper “Lost! Leveraging the Crowd for Probabilistic Visual Self-Localization” with M. Brubaker and R. Urtasun.

Juergen Gall: Emmy-Noether-Program Stipend from the German Science Foundation.

1.5 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.

Dr. Black's research spans computer vision, computer graphics, and computational neuroscience. In vision and graphics he is most known for his work on optical flow, robust statistical methods, human motion capture and analysis, 3D body shape modeling, and probabilistic models of the visual world. In computational neuroscience his work focuses on probabilistic models of the neural code and applications of neural decoding in human neural prosthetics.

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 MPI-Sintel dataset.

He is a co-founder and member of the board of directors of Body Labs Inc., which is commercializing his team's research on 3D human body shape.

Dr. Michael J. Black

Appointments

01/2011 – present	Director at the Max Planck Institute for Intelligent Systems
05/2012 – present	Honorary Professor, Department for Computer Science, University of Tübingen
04/2014 – present	Visiting Professor, Dept. of Inf. Tech. and Electrical Eng., ETH Zurich
01/2011 – present	Adjunct Professor, Dept. of Computer Science, Brown University
02/2013 – 06/2015	Managing Director of the MPI for Intelligent Systems, Stuttgart and Tübingen
05/2011 – 06/2013	Visiting Professor, Electrical Engineering, Stanford University

Awards & Honors (Selected)

2015	Elected foreign member of the Royal Swedish Academy of Sciences
2015	Best Paper Award, International Conference on 3D Vision (3DV)
2013	Helmholtz Prize for work that has stood the test of time
2010	Koenderink Prize for Fundamental Contributions in Computer Vision
2005	Marr Prize, Honorable Mention, Int. Conf. on Computer Vision, ICCV
1999	Marr Prize, Honorable Mention, Int. Conf. on Computer Vision, ICCV
1991	IEEE Computer Society, Outstanding Paper Award, CVPR

Selected Organization and Community Service (2011-2015)

2015	Co-organizer of the workshop "Scenes from Video II", Colchagua Valley, Chile
2015	Co-organizer of the Tutorial "How to build a digital human body" at ICCV
2015	Co-Director, Max Planck ETH Center for Learning Systems (CLS)
2015	Co-organizer of the Computational Vision Summer School, Bernstein Center
2014	Co-organizer of the Computer Vision Workshop, with ETH Zurich
2013	Co-organizer of the workshop "Scenes from Video", Barossa Valley, Australia
2012	Area Chair of the European Conference on Computer Vision (ECCV)
2012	Co-organizer of the Computational Vision Summer School, Bernstein Center

Memberships (2011–2015)

Royal Swedish Academy of Science, since 2015
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
Bernstein Center for Computational Neuroscience (BCCN), Tübingen, member since 2011
Canadian Institute for Advanced Research (CiFAR), Associate 2006 – 2014
Institute for Electrical and Electronics Engineers (IEEE): Senior Member since 2008
Society for Neuroscience (SfN): 2001-2014
Brown Institute for Brain Science (BIBS), member since 2000

Startup Activity and Board Memberships (2011 – 2015)

Body Labs Inc., New York, NY, Co-founder, Member of the Board, 2013 – present
Willow Garage, Palo Alto, CA, Advisory Board, 2008 – 2013
Videosurf Inc., San Mateo, CA, Scientific Advisory Board, 2006 – 2011. Sold to Microsoft

Selected Keynote, Conference, Workshop, and Public Talks (2011-2015)

Keynote, *12th IEEE Int. Conf. on Advanced Video and Signal-based Surveillance, AVSS*, Karlsruhe, 2015
Plenary, *Int. Conf. on Robotics and Automation (ICRA)*, Karlsruhe,May 2013
Keynote, *Swedish Society for Automated Image Analysis, SSBA*, Stockholm, 2012
Keynote, *Vision, Modeling and Visualization Workshop (VMV)*, Berlin, 2011
Keynote, *International Workshop on Human Activity Understanding from 3D Data*, 2011
Public talk, *Science Notes*, WAHRnehmung in Tübingen, 2015
FMX Conference on Animation, Effects, Games and Transmedia, Stuttgart, 2015

1.6 Publications

1.6.1 Books

2014

[1] S. Nowozin, P. V. Gehler, J. Jancsary, C. H. Lampert. Advanced Structured Prediction. Neural Information Processing Series. MIT Press, 432 pages, 2014.

2012

[2] A. Fossati, J. Gall, H. Grabner, X. Ren, K. Konolige. Consumer Depth Cameras for Computer Vision - Research Topics and Applications. Advances in Computer Vision and Pattern Recognition. Springer, 2012.

1.6.2 Proceedings

2015

[3] J. Gall, P. Gehler, B. Leibe, editors. Proceedings of the 37th German Conference on Pattern Recognition. Springer, 2015.

1.6.3 Journal Articles

2016

[4] 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.

[5] 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)*, 2016.

[6] T. von Marcard, G. Pons-Moll, B. Rosenhahn. Human Pose Estimation from Video and IMUs. *Transactions on Pattern Analysis and Machine Intelligence PAMI*, 2016.

2015

[7] S. Hauberg, A. Feragen, R. Enficiaud, M. Black. Scalable Robust Principal Component Analysis using Grassmann Averages. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 2015 (cited on page 10).

[8] B. Pepik, M. Stark, P. Gehler, B. Schiele. Multi-view and 3D Deformable Part Models. *Pattern Analysis and Machine Intelligence* **37** (11): 14, 2015 (cited on page 32).

[9] 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, 2015 (cited on pages 20–22).

[10] C. E. Vargas-Irwin, L. Franquemont, M. J. Black, J. P. Donoghue. Linking Objects to Actions: Encoding of Target Object and Grasping Strategy in Primate Ventral Premotor Cortex. *Journal of Neuroscience* **35** (30): 10888–10897, 2015 (cited on page 36).

[11] M. A. Brubaker, A. Geiger, R. Urtasun. Map-Based Probabilistic Visual Self-Localization. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 2015.

[12] G. Pons-Moll, J. Romero, N. Mahmood, M. J. Black. Dyna: A Model of Dynamic Human Shape in Motion. *ACM Transactions on Graphics, (Proc. SIGGRAPH)* **34** (4): 120:1–120:14, 2015 (cited on pages 21, 22).

[13] G. Pons-Moll, J. Taylor, J. Shotton, A. Hertzmann, A. Fitzgibbon. Metric Regression Forests for Correspondence Estimation. *International Journal of Computer Vision*: 1–13, 2015.

[14] C. E. Vargas-Irwin, D. M. Brandman, J. B. Zimmermann, J. P. Donoghue, M. J. Black. Spike train SIMilarity Space (SSIMS): A framework for single neuron and ensemble data analysis. *Neural Computation* **27** (1): 1–31, 2015 (cited on page 36).

2014

[15] 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, 2014 (cited on page 16).

[16] J. Foster, P. Nuyujukian, O. Freifeld, H. Gao, R. Walker, S. Ryu, T. Meng, B. Murmann, M. Black, K. Shenoy. A freely-moving monkey treadmill model. *J. of Neural Engineering* **11** (4): 046020, 2014 (cited on page 37).

[17] I. Piryankova, J. Stefanucci, J. Romero, S. de la Rosa, M. Black, B. Mohler. Can I recognize my body's weight? The influence of shape and texture on the perception of self. *ACM Transactions on Applied Perception for the Symposium on Applied Perception* **11** (3): 13:1–13:18, 2014 (cited on page 25).

[18] A. Tsoli, N. Mahmood, M. J. Black. Breathing Life into Shape: Capturing, Modeling and Animating 3D Human Breathing. *ACM Transactions on Graphics, (Proc. SIGGRAPH)* **33** (4): 52:1–52:11, 2014 (cited on page 21).

[19] M. L. Homer, J. A. Perge, M. J. Black, M. T. Harrison, S. S. Cash, L. R. Hochberg. Adaptive Offset Correction for Intracortical Brain Computer Interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **22** (2): 239–248, 2014 (cited on page 36).

[20] A. Geiger, M. Lauer, C. Wojek, C. Stiller, R. Urtasun. 3D Traffic Scene Understanding from Movable Platforms. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* **36** (5): 1012–1025, 2014 (cited on page 31).

[21] D. Sun, S. Roth, M. J. Black. A Quantitative Analysis of Current Practices in Optical Flow Estimation and the Principles behind Them. *International Journal of Computer Vision (IJCV)* **106** (2): 115–137, 2014 (cited on page 26).

2013

[22] P. Hennig, M. Kiefel. Quasi-Newton Methods: A New Direction. *Journal of Machine Learning Research* **14** (1): 843–865, 2013.

[23] J. Romero, T. Feix, C. Ek, H. Kjellstrom, D. Kragic. Extracting Postural Synergies for Robotic Grasping. *Robotics, IEEE Transactions on* **29** (6): 1342–1352, 2013.

[24] A. Lehmann, P. Gehler, L. VanGool. Branch&Rank for Efficient Object Detection. *International Journal of Computer Vision*, 2013.

[25] C. Wang, N. Komodakis, N. Paragios. Markov Random Field Modeling, Inference & Learning in Computer Vision & Image Understanding: A Survey. *Computer Vision and Image Understanding (CVIU)* **117** (11): 1610–1627, 2013 (cited on page 28).

[26] A. Geiger, P. Lenz, C. Stiller, R. Urtasun. Vision meets Robotics: The KITTI Dataset. *International Journal of Robotics Research* **32** (11): 1231–1237, 2013 (cited on page 34).

[27] J. Romero, H. Kjellström, C. H. Ek, D. Kragic. Non-parametric hand pose estimation with object context. *Image and Vision Computing* **31** (8): 555–564, 2013.

[28] A. Sekunova, M. Black, L. Parkinson, J. J. S. Barton. Viewpoint and pose in body-form adaptation. *Perception* **42** (2): 176–186, 2013.

[29] Y. Liu, J. Gall, C. Stoll, Q. Dai, H.-P. Seidel, C. Theobalt. Markerless Motion Capture of Multiple Characters Using Multi-view Image Segmentation. *Transactions on Pattern Analysis and Machine Intelligence* **35** (11): 2720–2735, 2013 (cited on page 17).

[30] S. Hauberg, F. Lauze, K. S. Pedersen. Unscented Kalman Filtering on Riemannian Manifolds. *Journal of Mathematical Imaging and Vision* **46** (1): 103–120, 2013.

[31] G. Fanelli, M. Dantone, J. Gall, A. Fossati, L. van Gool. Random Forests for Real Time 3D Face Analysis. *International Journal of Computer Vision* **101** (3): 437–458, 2013.

2012

[32] G. Stanley, J. Jin, Y. Wang, G. Desbordes, Q. Wang, M. Black, J.-M. Alonso. Visual Orientation and Directional Selectivity Through Thalamic Synchrony. *Journal of Neuroscience* **32** (26): 9073–9088, 2012 (cited on page 11).

[33] P. Guan, L. Reiss, D. Hirshberg, A. Weiss, M. J. Black. DRAPE: DRessing Any PErson. *ACM Trans. on Graphics (Proc. SIGGRAPH)* **31** (4): 35:1–35:10, 2012.

[34] A. Yao, J. Gall, L. van Gool. Coupled Action Recognition and Pose Estimation from Multiple Views. *International Journal of Computer Vision* **100** (1): 16–37, 2012 (cited on page 18).

2011

[35] L. Sigal, M. Isard, H. Haussecker, M. J. Black. Loose-limbed People: Estimating 3D Human Pose and Motion Using Non-parametric Belief Propagation. *International Journal of Computer Vision* **98** (1): 15–48, 2011 (cited on pages 17, 23).

[36] J. D. Simeral, S.-P. Kim, M. J. Black, J. P. Donoghue, L. R. Hochberg. Neural control of cursor trajectory and click by a human with tetraplegia 1000 days after implant of an intracortical microelectrode array. *J. of Neural Engineering* **8** (2): 025027, 2011 (cited on page 36).

[37] S.-P. Kim, J. D. Simeral, L. R. Hochberg, J. P. Donoghue, G. M. Friehs, M. J. Black. Point-and-Click Cursor Control With an Intracortical Neural Interface System by Humans With Tetraplegia. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **19** (2): 193–203, 2011 (cited on page 36).

[38] S. Baker, D. Scharstein, J. P. Lewis, S. Roth, M. J. Black, R. Szeliski. A Database and Evaluation Methodology for Optical Flow. *International Journal of Computer Vision* **92** (1): 1–31, 2011 (cited on page 34).

1.6.4 Conference Papers

2016

[39] 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 (cited on page 25).

2015

[40] P. Lenz, A. Geiger, R. Urtasun. FollowMe: Efficient Online Min-Cost Flow Tracking with Bounded Memory and Computation. In *International Conference on Computer Vision (ICCV)*, 2015.

[41] D. Tzionas, J. Gall. 3D Object Reconstruction from Hand-Object Interactions. In *International Conference on Computer Vision (ICCV)*, 2015 (cited on page 19).

[42] F. Bogo, M. J. Black, M. Loper, J. Romero. Detailed Full-Body Reconstructions of Moving People from Monocular RGB-D Sequences. In *International Conference on Computer Vision (ICCV)*, pages 2300–2308, 2015 (cited on pages 16, 22, 24).

[43] N. Kong, M. J. Black. Intrinsic Depth: Improving Depth Transfer with Intrinsic Images. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3514–3522, 2015 (cited on page 29).

[44] M. Kiefel, V. Jampani, P. V. Gehler. Permutohedral Lattice CNNs. In *ICLR Workshop Track*, 2015.

[45] C. Zhou, F. Güney, Y. Wang, A. Geiger. Exploiting Object Similarity in 3D Reconstruction. In *International Conference on Computer Vision (ICCV)*, 2015.

[46] A. C. Wellerdiek, M. Breidt, M. N. Geuss, S. Streuber, U. Kloos, M. J. Black, B. J. Mohler. Perception of Strength and Power of Realistic Male Characters. In *Proc. ACM SIGGRAPH Symposium on Applied Perception, SAP'15*, pages 7–14, 2015 (cited on page 25).

[47] A. O. Ulusoy, A. Geiger, M. J. Black. Towards Probabilistic Volumetric Reconstruction using Ray Potentials. In *3D Vision (3DV), 2015 3rd International Conference on*, pages 10–18, 2015.

[48] J. Romero, M. Loper, M. J. Black. FlowCap: 2D Human Pose from Optical Flow. In *Pattern Recognition, Proc. 37th German Conference on Pattern Recognition (GCPR)*. Vol. LNCS 9358, pages 412–423, 2015 (cited on page 15).

[49] M. Menze, C. Heipke, A. Geiger. Discrete Optimization for Optical Flow. In *German Conference on Pattern Recognition (GCPR)*. Vol. 9358, pages 16–28, 2015 (cited on page 26).

[50] A. Geiger, C. Wang. Joint 3D Object and Layout Inference from a single RGB-D Image. In *German Conference on Pattern Recognition (GCPR)*. Vol. 9358. Lecture Notes in Computer Science, pages 183–195, 2015 (cited on pages 28, 31).

[51] L. Sevilla-Lara, J. Wulff, K. Sunkavalli, E. Shechtman. Smooth Loops from Unconstrained Video. In *Computer Graphics Forum (Proceedings of EGSR)*, 2015.

[52] M. Menze, C. Heipke, A. Geiger. Joint 3D Estimation of Vehicles and Scene Flow. In *Proc. of the ISPRS Workshop on Image Sequence Analysis (ISA)*, 2015 (cited on page 28).

[53] B. Pepik, M. Stark, P. Gehler, T. Ritschel, B. Schiele. 3D Object Class Detection in the Wild. In *Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2015 (cited on page 32).

[54] S. Zuffi, M. J. Black. The Stitched Puppet: A Graphical Model of 3D Human Shape and Pose. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2015)*, pages 3537–3546, 2015 (cited on page 23).

[55] J. Wulff, M. J. Black. Efficient Sparse-to-Dense Optical Flow Estimation using a Learned Basis and Layers. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2015)*, pages 120–130, 2015 (cited on pages 26, 27).

[56] I. Akhter, M. J. Black. Pose-Conditioned Joint Angle Limits for 3D Human Pose Reconstruction. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2015)*, pages 1446–1455, 2015 (cited on page 18).

[57] M. Menze, A. Geiger. Object Scene Flow for Autonomous Vehicles. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2015*, pages 3061–3070, 2015 (cited on pages 28, 34).

[58] F. Güney, A. Geiger. Displets: Resolving Stereo Ambiguities using Object Knowledge. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2015*, 2015 (cited on page 28).

[59] V. Jampani*, S. M. A. Eslami*, D. Tarlow, P. Kohli, J. Winn. Consensus Message Passing for Layered Graphical Models. In *Eighteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*. Vol. 38, pages 425–433, 2015.

[60] V. Jampani*, R. Gadde*, P. V. Gehler. Efficient Facade Segmentation using Auto-Context. In *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on*, pages 1038–1045, 2015 (cited on page 31).

[61] V. Jampani, S. Nowozin, M. Loper, P. V. Gehler. The Informed Sampler: A Discriminative Approach to Bayesian Inference in Generative Computer Vision Models. In *Special Issue on Generative Models in Computer Vision and Medical Imaging*. Vol. 136, pages 32–44, 2015 (cited on page 30).

2014

[62] M. Kiefel, C. Schuler, P. Hennig. Probabilistic Progress Bars. In *Conference on Pattern Recognition (GCPR)*. Vol. 8753. Lecture Notes in Computer Science, pages 331–341, 2014.

[63] J. Wulff, M. J. Black. Modeling Blurred Video with Layers. In *Computer Vision – ECCV 2014*. Vol. 8694. Lecture Notes in Computer Science, pages 236–252, 2014 (cited on page 27).

[64] L. Sevilla-Lara, D. Sun, E. G. Learned-Miller, M. J. Black. Optical Flow Estimation with Channel Constancy. In *Computer Vision – ECCV 2014*. Vol. 8689. Lecture Notes in Computer Science, pages 423–438, 2014 (cited on page 26).

[65] Z. Hong, C. Wang, X. Mei, D. Prokhorov, D. Tao. Tracking using Multilevel Quantizations. In *Computer Vision – ECCV 2014*. Vol. 8694. Lecture Notes in Computer Science, pages 155–171, 2014.

[66] M. M. Loper, M. J. Black. OpenDR: An Approximate Differentiable Renderer. In *Computer Vision – ECCV 2014*. Vol. 8695. Lecture Notes in Computer Science, pages 154–169, 2014 (cited on page 30).

[67] N. Kong, P. V. Gehler, M. J. Black. Intrinsic Video. In *Computer Vision – ECCV 2014*. Vol. 8690. Lecture Notes in Computer Science, pages 360–375, 2014 (cited on page 29).

[68] J. Bohg, J. Romero, A. Herzog, S. Schaal. Robot Arm Pose Estimation through Pixel-Wise Part Classification. In *IEEE International Conference on Robotics and Automation (ICRA) 2014*, pages 3143–3150, 2014.

[69] D. Tzionas, A. Srikantha, P. Aponte, J. Gall. Capturing Hand Motion with an RGB-D Sensor, Fusing a Generative Model with Salient Points. In *German Conference on Pattern Recognition (GCPR)*. Lecture Notes in Computer Science, pages 1–13, 2014.

[70] M. Kiefel, P. Gehler. Human Pose Estimation with Fields of Parts. In *Computer Vision – ECCV 2014*. Vol. LNCS 8693. Lecture Notes in Computer Science, pages 331–346, 2014 (cited on page 14).

[71] A. Srikantha, J. Gall. Discovering Object Classes from Activities. In *European Conference on Computer Vision*. Vol. 8694. Lecture Notes in Computer Science, pages 415–430, 2014 (cited on page 33).

[72] F. Bogo, J. Romero, E. Peserico, M. J. Black. Automated Detection of New or Evolving Melanocytic Lesions Using a 3D Body Model. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Vol. 8673. Lecture Notes in Computer Science, pages 593–600, 2014.

[73] M. Schoenbein, A. Geiger. Omnidirectional 3D Reconstruction in Augmented Manhattan Worlds. In *International Conference on Intelligent Robots and Systems*, pages 716–723, 2014.

[74] A. Srikantha, J. Gall. Hough-based Object Detection with Grouped Features. In *International Conference on Image Processing*, pages 1653–1657, 2014.

[75] J. Pacheco, S. Zuffi, M. J. Black, E. Sudderth. Preserving Modes and Messages via Diverse Particle Selection. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*. Vol. 32. 1, pages 1152–1160, 2014 (cited on pages 10, 14, 23).

[76] G. Pons-Moll, D. J. Fleet, B. Rosenhahn. Posebits for Monocular Human Pose Estimation. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2345–2352, 2014.

[77] F. Bogo, J. Romero, M. Loper, M. J. Black. FAUST: Dataset and evaluation for 3D mesh registration. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 3794–3801, 2014 (cited on pages 20, 35).

[78] S. Hauberg, A. Feragen, M. J. Black. Grassmann Averages for Scalable Robust PCA. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 3810–3817, 2014.

[79] O. Freifeld, S. Hauberg, M. J. Black. Model Transport: Towards Scalable Transfer Learning on Manifolds. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1378–1385, 2014 (cited on page 9).

[80] B. Pepik, M. Stark, P. Gehler, B. Schiele. Multi-View Priors for Learning Detectors from Sparse Viewpoint Data. In *International Conference on Learning Representations*, 2014 (cited on page 32).

[81] M. Andriluka, L. Pishchulin, P. Gehler, B. Schiele. Human Pose Estimation: New Benchmark and State of the Art Analysis. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 3686–3693, 2014 (cited on page 35).

[82] M. Schoenbein, T. Strauss, A. Geiger. Calibrating and Centering Quasi-Central Catadioptric Cameras. In *IEEE International Conference on Robotics and Automation*, pages 4443–4450, 2014.

[83] M. Roser, M. Dunbabin, A. Geiger. Simultaneous Underwater Visibility Assessment, Enhancement and Improved Stereo. In *IEEE International Conference on Robotics and Automation*, pages 3840–3847, 2014.

[84] A. M. Lehrmann, P. V. Gehler, S. Nowozin. Efficient Non-linear Markov Models for Human Motion. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1314–1321, 2014 (cited on page 18).

[85] P. Hennig, S. Hauberg. Probabilistic Solutions to Differential Equations and their Application to Riemannian Statistics. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics*. Vol. 33. JMLR: Workshop and Conference Proceedings, pages 347–355, 2014.

[86] A. Tsoli, M. Loper, M. J. Black. Model-based Anthropometry: Predicting Measurements from 3D Human Scans in Multiple Poses. In *Proceedings Winter Conference on Applications of Computer Vision*, pages 83–90, 2014.

2013

[87] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, M. J. Black. Towards understanding action recognition. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3192–3199, 2013 (cited on pages 15, 35).

[88] Y. Zeng, C. Wang, X. Gu, D. Samaras, N. Paragios. A Generic Deformation Model for Dense Non-Rigid Surface Registration: a Higher-Order MRF-based Approach. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3360 3367, 2013.

[89] A. M. Lehrmann, P. Gehler, S. Nowozin. A Non-parametric Bayesian Network Prior of Human Pose. In *Proceedings IEEE Conf. on Computer Vision (ICCV)*, pages 1281–1288, 2013 (cited on page 18).

[90] L. Pishchulin, M. Andriluka, P. Gehler, B. Schiele. Strong Appearance and Expressive Spatial Models for Human Pose Estimation. In *International Conference on Computer Vision (ICCV)*, pages 3487–3494, 2013 (cited on page 14).

[91] S. Zuffi, J. Romero, C. Schmid, M. J. Black. Estimating Human Pose with Flowing Puppets. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3312–3319, 2013 (cited on page 15).

[92] H. Zhang, A. Geiger, R. Urtasun. Understanding High-Level Semantics by Modeling Traffic Patterns. In *International Conference on Computer Vision*, pages 3056–3063, 2013 (cited on page 31).

[93] M. Homer, M. Harrison, M. J. Black, J. Perge, S. Cash, G. Friehs, L. Hochberg. Mixing Decoded Cursor Velocity and Position from an Offline Kalman Filter Improves Cursor Control in People with Tetraplegia. In *6th International IEEE EMBS Conference on Neural Engineering*, pages 715–718, 2013.

[94] D. Tzionas, J. Gall. A Comparison of Directional Distances for Hand Pose Estimation. In *German Conference on Pattern Recognition (GCPR)*. Vol. 8142. Lecture Notes in Computer Science, pages 131–141, 2013.

[95] M. A. Brubaker, A. Geiger, R. Urtasun. Lost! Leveraging the Crowd for Probabilistic Visual Self-Localization. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2013)*, pages 3057–3064, 2013.

[96] B. Pepik, M. Stark, P. Gehler, B. Schiele. Occlusion Patterns for Object Class Detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2013 (cited on page 32).

[97] L. Pishchulin, M. Andriluka, P. Gehler, B. Schiele. Poselet conditioned pictorial structures. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 588–595, 2013 (cited on page 14).

[98] D. Sun, J. Wulff, E. Sudderth, H. Pfister, M. Black. A fully-connected layered model of foreground and background flow. In *IEEE Conf. on Computer Vision and Pattern Recognition, (CVPR 2013)*, pages 2451–2458, 2013 (cited on page 27).

[99] M. Dantone, J. Gall, C. Leistner, L. van Gool. Human Pose Estimation using Body Parts Dependent Joint Regressors. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 3041–3048, 2013.

2012

[100] P. Hennig, M. Kiefel. Quasi-Newton Methods: A New Direction. In *Proceedings of the 29th International Conference on Machine Learning*. ICML '12, pages 25–32, 2012.

[101] S. Ghosh, E. Sudderth, M. Loper, M. Black. From Deformations to Parts: Motion-based Segmentation of 3D Objects. In *Advances in Neural Information Processing Systems 25 (NIPS)*, pages 2006–2014, 2012 (cited on page 23).

[102] D. Hirshberg, M. Loper, E. Rachlin, M. Black. Coregistration: Simultaneous alignment and modeling of articulated 3D shape. In *European Conf. on Computer Vision (ECCV)*. LNCS 7577, Part IV, pages 242–255, 2012 (cited on pages 20, 22).

[103] S. Hauberg, O. Freifeld, M. J. Black. A Geometric Take on Metric Learning. In *Advances in Neural Information Processing Systems (NIPS) 25*, pages 2033–2041, 2012 (cited on page 10).

[104] S. Hauberg, K. S. Pedersen. Spatial Measures between Human Poses for Classification and Understanding. In *Articulated Motion and Deformable Objects*. Vol. 7378. LNCS, pages 26–36, 2012 (cited on page 18).

[105] M. Ristin, J. Gall, L. van Gool. Local Context Priors for Object Proposal Generation. In *Asian Conference on Computer Vision (ACCV)*. Vol. 7724. LNCS, pages 57–70, 2012.

[106] 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, 2012 (cited on page 34).

[107] J. Wulff, D. J. Butler, G. B. Stanley, M. J. Black. Lessons and insights from creating a synthetic optical flow benchmark. In *ECCV Workshop on Unsolved Problems in Optical Flow and Stereo Estimation*. Part II, LNCS 7584, pages 168–177, 2012 (cited on page 34).

[108] O. Freifeld, M. J. Black. Lie Bodies: A Manifold Representation of 3D Human Shape. In *European Conf. on Computer Vision (ECCV)*. Part I, LNCS 7572, pages 1–14, 2012 (cited on page 9).

[109] S. Pellegrini, J. Gall, L. Sigal, L. van Gool. Destination Flow for Crowd Simulation. In *Workshop on Analysis and Retrieval of Tracked Events and Motion in Imagery Streams*. Vol. 7585. LNCS, pages 162–171, 2012.

[110] N. Razavi, N. Alvar, J. Gall, L. van Gool. Sparsity Potentials for Detecting Objects with the Hough Transform. In *British Machine Vision Conference (BMVC)*, pages 11.1–11.10, 2012.

[111] A. L'opez-M'endez, J. Gall, J. Casas, L. van Gool. Metric Learning from Poses for Temporal Clustering of Human Motion. In *British Machine Vision Conference (BMVC)*, pages 49.1–49.12, 2012.

[112] L. Ballan, A. Taneja, J. Gall, L. van Gool, M. Pollefeys. Motion Capture of Hands in Action using Discriminative Salient Points. In *European Conference on Computer Vision (ECCV)*. Vol. 7577. LNCS, pages 640–653, 2012 (cited on page 19).

[113] N. Razavi, J. Gall, P. Kohli, L. van Gool. Latent Hough Transform for Object Detection. In *European Conference on Computer Vision (ECCV)*. Vol. 7574. LNCS, pages 312–325, 2012 (cited on page 33).

[114] B. Pepik, P. Gehler, M. Stark, B. Schiele. 3D2PM – 3D Deformable Part Models. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Lecture Notes in Computer Science, pages 356–370, 2012 (cited on page 32).

[115] C. Dann, P. Gehler, S. Roth, S. Nowozin. Pottics – The Potts Topic Model for Semantic Image Segmentation. In *Proceedings of 34th DAGM Symposium*. Lecture Notes in Computer Science, pages 397–407, 2012 (cited on page 31).

[116] J. D. Foster, P. Nuyujukian, O. Freifeld, S. Ryu, M. J. Black, K. V. Shenoy. A framework for relating neural activity to freely moving behavior. In *34th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'12)*, pages 2736–2739, 2012 (cited on pages 11, 37).

[117] A. Yao, J. Gall, C. Leistner, L. van Gool. Interactive Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3242–3249, 2012.

[118] M. Dantone, J. Gall, G. Fanelli, L. van Gool. Real-time Facial Feature Detection using Conditional Regression Forests. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2578–2585, 2012.

[119] B. Pepik, M. Stark, P. Gehler, B. Schiele. Teaching 3D Geometry to Deformable Part Models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3362–3369, 2012 (cited on page 32).

[120] D. Sun, E. Sudderth, M. J. Black. Layered segmentation and optical flow estimation over time. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1768–1775, 2012 (cited on page 27).

[121] S. Zuffi, O. Freifeld, M. J. Black. From pictorial structures to deformable structures. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 3546–3553, 2012 (cited on pages 14, 15, 23).

2011

[122] P. Gehler, C. Rother, M. Kiefel, L. Zhang, B. Schölkopf. Recovering Intrinsic Images with a Global Sparsity Prior on Reflectance. In *Advances in Neural Information Processing Systems 24*, pages 765–773, 2011 (cited on page 29).

[123] D. A. Hirshberg, M. Loper, E. Rachlin, A. Tsoli, A. Weiss, B. Corner, M. J. Black. Evaluating the Automated Alignment of 3D Human Body Scans. In *2nd International Conference on 3D Body Scanning Technologies*, pages 76–86, 2011 (cited on page 20).

[124] J. Foster, O. Freifeld, P. Nuyujukian, S. Ryu, M. J. Black, K. Shenoy. Combining wireless neural recording and video capture for the analysis of natural gait. In *Proc. 5th Int. IEEE EMBS Conf. on Neural Engineering*, pages 613–616, 2011 (cited on page 37).

[125] A. Weiss, D. Hirshberg, M. Black. Home 3D body scans from noisy image and range data. In *Int. Conf. on Computer Vision (ICCV)*, pages 1951–1958, 2011 (cited on page 24).

[126] A. Tsoli, M. J. Black. Shape and pose-invariant correspondences using probabilistic geodesic surface embedding. In *33rd Annual Symposium of the German Association for Pattern Recognition (DAGM)*. Vol. 6835. Lecture Notes in Computer Science, pages 256–265, 2011.

1.6.5 Book Chapters

2014

[127] J. Gall. Simulated Annealing. In. *Encyclopedia of Computer Vision*. Ed. by K. Ikeuchi. Springer Verlag, pages 737–741, 2014.

2013

[128] J. Gall, V. Lempitsky. Class-Specific Hough Forests for Object Detection. In *Decision Forests for Computer Vision and Medical Image Analysis*. Ed. by A. Criminisi, J. Shotton. Springer. Chap. 11, pages 143–157, 2013 (cited on page 33).

2012

[129] A. Weiss, D. Hirshberg, M. J. Black. Home 3D body scans from noisy image and range data. In *Consumer Depth Cameras for Computer Vision: Research Topics and Applications*. Ed. by A. Fossati, J. Gall, H. Grabner, X. Ren, K. Konolige. Springer-Verlag. Chap. 6, pages 99–118, 2012 (cited on page 24).

[130] J. Gall, N. Razavi, L. van Gool. An Introduction to Random Forests for Multi-class Object Detection. In *Outdoor and Large-Scale Real-World Scene Analysis*. Ed. by F. Dellaert, J.-M. Frahm, M. Pollefeys, B. Rosenhahn, L. Leal-Taix'e. Vol. 7474. LNCS. Springer, pages 243–263, 2012.

2011

[131] M. Andriluka, L. Sigal, M. J. Black. Benchmark datasets for pose estimation and tracking. In *Visual Analysis of Humans: Looking at People*. Ed. by Moesland, Hilton, Kr"uger, Sigal. Springer-Verlag, London, pages 253–274, 2011 (cited on page 35).

[132] S. Roth, M. J. Black. Steerable random fields for image restoration and inpainting. In *Markov Random Fields for Vision and Image Processing*. Ed. by A. Blake, P. Kohli, C. Rother. MIT Press, pages 377–387, 2011 (cited on page 28).

1.6.6 Theses

PhD Theses

[133] F. Bogo. From Scans to Models: Registration of 3D Human Shapes Exploiting Texture Information. PhD thesis. University of Padova, 2015.

[134] S. Zuffi. Shape Models of the Human Body for Distributed Inference. PhD thesis. Brown University, 2015.

[135] A. Tsoli. Modeling the Human Body in 3D: Data Registration and Human Shape Representation. PhD thesis. Brown University, Department of Computer Science, 2014.

[136] O. Freifeld. Statistics on Manifolds with Applications to Modeling Shape Deformations. PhD thesis. Brown University, 2013.

[137] D. Sun. From Pixels to Layers: Joint Motion Estimation and Segmentation. PhD thesis. Brown University, Department of Computer Science, 2012.

[138] P. Guan. Virtual Human Bodies with Clothing and Hair: From Images to Animation. PhD thesis. Brown University, Department of Computer Science, 2012.