Report from Dagstuhl Seminar 19102

3D Morphable Models

Edited by

Bernhard Egger[1], William Smith[2], Christian Theobalt[3], and Thomas Vetter[4]

1 **MIT – Cambridge, US,** egger@mit.edu
2 **University of York, GB,** william.smith@york.ac.uk
3 **MPI für Informatik – Saarbrücken, DE,** theobalt@mpi-inf.mpg.de
4 **Universität Basel, CH,** thomas.vetter@unibas.ch

Abstract

3D Morphable Models is a statistical object model separating shape from appearance variation. Typically, they are used as a statistical prior in computer graphics and vision. This report summarizes the Dagstuhl seminar on 3D Morphable Models, March 3-8, 2019. It was a first specific meeting of a broader group of people working with 3D Morphable Models of faces and bodies. This meeting of 26 researchers was held 20 years after the seminal work was published at Siggraph. We summarize the discussions, presentations and results of this workshop.

Seminar March 3–8, 2019 – http://www.dagstuhl.de/19102
2012 ACM Subject Classification Computing methodologies → Computer graphics, Computing methodologies → Computer vision
Keywords and phrases 3D Computer Vision, Analysis-by-Synthesis, Computer Graphics, Generative Models, Statistical Modelling
Digital Object Identifier 10.4230/DagRep.9.3.16

1 Executive Summary

Bernhard Egger (MIT – Cambridge, US)
William Smith (University of York, GB)
Christian Theobalt (MPI für Informatik – Saarbrücken, DE)
Thomas Vetter (Universität Basel, CH)

A total of 45 people was invited to this seminar in the first round of invitations. The seminar was fully booked after the first round and 26 researchers from academia and industry participated in the seminar. 21 researchers presented their work in around 15-30 minutes presentations, an abstract of each presentation is included in this report. Besides those presentations participants where presenting their shared data and software in a specific slot. We collected this information in a list of shared resources which we made publicly available[1]. This overview and exchange was one of the aims we had initially in mind when organizing the workshop. In the beginning of the workshop we collected ideas for discussions in our flexible sessions, those ideas are also contained in this report. We then structured the seminar fixing

[1] https://github.com/3d-morphable-models/curated-list-of-awesome-3D-Morphable-Model-software-and-data

the topics of discussion for the flexible sessions. The summaries of those discussions are also contained in this report. One slot was reserved for a joint group discussion on upcoming ethical concerns on the methods we are developing. This interesting and well organized discussion was an initiative from the participants and not foreseen by the organizers. Another bigger discussion was around the topic of how to compare different approaches and how to establish a benchmark. We did not completely converge on a final solution but we identified currently available benchmarks and we discussed how a gold-standard benchmark would look like. Another aim of the workshop was to initiate an edited book or a survey paper with broad support. Arising from the workshop a group of 13 junior and senior researchers started to work on a joint survey and perspective paper on 20 years of Morphable Face Models. Discussions and presentations were followed by vivid discussions on current challenges and future research directions. To future nurture the ideas of the seminar we started a google group for discussions, sharing news and exchanging students [2]. The group would like to meet again at Dagstuhl in 2022. The program was more dense than expected and we would like to have more time for discussions in groups after a set of talks. We would like to highlight 5 main discussion points:

- To what degree of detail we need to model in 3D and physically adequate, what can we learn from semi-supervised or unsupervised 2D data?
- Is the model depending on the application or is there a golden standard model that is able to fit all applications?
- The current revolution of deep learning in computer vision enables a lot of novel strategies and speeds up the models, however, other challenges in modeling, synthesis and inverse rendering remain and new deep learning specific challenges are introduced.
- What are the ethical implications of the models and systems we are building?
- How will the field develop in the next 20 years? Which challenges should we focus on?

We started the seminar with a short introduction of everybody. The homework was to introduce themself with at most one slide and prepare one important question, challenge or goal you would like to discuss during the seminar.

- **Thabo Beeler**: Non-Linear Morphable Models. How to get off Model in a meaningful way?
- **Florian Bernard**: Deeper integration of models of human knowledge and algorithms into learning systems. What are potential perspectives? How to best approach this?
- **Michael J. Black**: What's next? Increasing realism? Deep representations? Something else?
- **Volker Blanz:** Expressive model also reproduces non-face structures! How to discriminate between face and non-face? Future: better regularization, rely on trained regressors, recognize glasses ?
- **Bernhard Egger:** What to model? What to learn?
- **Victoria Fernandez Abrevaya:** How far are we from closing the gap between high-quality and low-quality capture devices, and can we use 3DMM for this?
- **Patrik Huber:** What is missing to reliably reconstruct realistic 3D faces from mostly uncontrolled 2D footage?
- **Ron Kimmel:** Geometry is the art of finding the "right" parametrization. Deep Learning is a technology that exploits convenient parametric spaces (CNN) for classification. Any hope for unification? Is translating geometry into algebra the answer?

[2] https://groups.google.com/forum/#!forum/3d-morphable-models

- **Tatsuro Koizumi:** How to evaluate and assure the robustness of neural network-based reconstruction? How to improve the stability of self-supervised training?
- **Adam Edward Kortylewski:** Can we resolve the limitations of Deep Learning with Generative Object Models?
- **Yeara Kozlov:** Can physically based face modeling be replaced by machine learning?
- **Andreas Morel-Forster:** Fast posterior estimation – A contradiction?
- **Nick Pears:** How to build deeper, wider models?
- **Gerard Pons-Moll:** Is the Euclidean 3D space the right space to model humans, clothing and hair?
- **Emanuele Rodolà:** Can we make inverse spectral geometry useful in practice?
- **Sami Romdhani:** How to combine Deep Learning and 3D Equations to generate images?
- **Javier Romero:** How can Deep Nets learn from unstructured, uncalibrated views?
- **Shunsuke Saito:** Is there an unified representation to represent digital human without explicitly having prior for each component?
- **William Smith:** Self-supervision: holy grail or just re-discovering gradient descent-based analysis-by-synthesis? How do we make sure the gradients of our losses are really useful (Appearance loss: meaningless when far from good solution, Landmark loss: ambiguous (and not self-supervised), Rasterization: not differentiable)?
- **Ayush Tewari:** How can we build high quality 3D morphable models from 2D data?
- **Christian Theobalt:** Can we build a 4D Real World Reconstruction Loop? Ethical, Privacy, Security Questions of Parametric/Morphable Model Building and Reconstruction Algorithms
- **Thomas Vetter:** Did we learn much about this optimization problem (inverse rendering)?
- **Stefanie Wuhrer:** How to effectively learn parametric human models from captured data using minimal supervision?
- **Michael Zollhöfer:** What is the best representation for deep learning-based 3D reconstruction and image synthesis?
- **Silvia Zuffi:** How to model skin dynamics from video?

After the individual introductions, we discussed those ideas in discussion groups to identify points to discuss during the seminar. The following list is the unfiltered result of our brainstorming on open questions and challenges.

- Where to spend the next 20 years? Perfection: finer detail? Move it: Movement, new representation, new goals, new data? Break it: hair, clothing, new representation, new goals, new data?
- Why aren't we focusing on fixing the obvious errors?
- Optimization: Why aren't we doing more to understand our objective function and adopt the algorithms?
- How to predict distributions instead of point estimates?
- How much detail to model vs. overfitting?
- How to evaluate Photorealism?
- Should vision people be more aware of graphics standard for photorealism?
- Is it important to understand?
- Do we need correspondences to build 3D models and predictions?
- How to learn 3D from 2D?
- How to adapt models over time (without calibration)?
- How to deal with multi-view and video in CNNs?

- Which courses/skills are required?
- Use for society?
- What to leave for industry?
- What is the role of academia within industry (collaboration vs. isolation)?
- Representations (beyond triangle meshes) to deal with category discontinuities, e.g. smooth surface vs. hair
- Evaluation of shape and appearance reconstruction
- Connections between deep learning and parametric models
- Role of axiomic models in learning
- Comparability: Benchmark and metrics
- Future prediction of motion
- Self-supervision
- Differentiable inverse rendering

2 Table of Contents

3 Overview of Talks

3.1 Shape Synthesis with Local-Global Tensors

Thabo Beeler (Disney Research – Zürich, CH)

Joint work of Thabo Beeler, Mengjiao Wang, Derek Bradley, Stefanos Zafeiriou

Global 3DMMs are extremely popular due to their simplicity and robustness. This robustness, however, comes at the price of flexibility as 3DMMs can only represent data that is 'within model'. For something that exhibits as much variation as the human face, this effectively means that only coarse scale features and coarse scale deformation may be captured by a global 3DMM. We explore the idea to couple such a global 3DMM with local 3DMMs in order to enrich the expressive power of the statistical model whilst not sacrificing too much of the robustness. We demonstrate our proposed coupling of global/local tensor models on the task to synthesize expressions for a person that are both expressive and preserve the identity of the subject, starting from just a neutral scan of the subject.

3.2 Combinatorial Non-Rigid Shape-to-Image Matching

Florian Bernard (MPI für Informatik – Saarbrücken, DE)

Joint work of Florian Bernard, Frank R. Schmidt, Johan Thunberg, Daniel Cremers
Main reference Florian Bernard, Frank R. Schmidt, Johan Thunberg, Daniel Cremers: "A Combinatorial Solution to Non-Rigid 3D Shape-to-Image Matching", CoRR, Vol. abs/1611.05241, 2016.
URL https://arxiv.org/abs/1611.05241

We propose a combinatorial solution for the problem of non-rigidly matching a 3D shape to 3D image data. To this end, we model the shape as a triangular mesh and allow each triangle of this mesh to be rigidly transformed to achieve a suitable matching to the image. By penalizing the distance and the relative rotation between neighboring triangles, our matching compromises between image and shape information. We resolve two major challenges: Firstly, we address the resulting large and NP-hard combinatorial problem with a suitable graph-theoretic approach. Secondly, we propose an efficient discretization of the unbounded 6-dimensional Lie group SE(3). In contrast to existing local (gradient descent) optimization methods, we obtain solutions that do not require a good initialization and that are within a bound of the optimal solution.

3.3 Expressive human body models for communication and interaction

Michael J. Black (MPI für Intelligente Systeme – Tübingen, DE)

Bodies in computer vision have often been an afterthought. Human pose is often represented by 10-12 body joints in 2D or 3D. This is inspired by Johannson's moving light display experiments, which showed that some human actions can be recognized from the motion of the major joints of the body. The joints of the body, however, do not capture everything

all that we need to understand human behavior. In our work we have focused on 3D body shape, represented as a triangulated mesh. Shape gives us more information about a person related to their health, age, fitness, and clothing size. But shape is also useful because our body surface is critical to our physical interactions with the world. We cannot interpenetrate objects and they cannot interpenetrate us. Consequently we developed the SMPL body model, which is widely used in research and industry. It is simple, efficient, posable, and compatible with most graphics packages. It is also differentiable and easy to integrate into optimization or deep learning methods. While popular, SMPL has drawbacks for representing human actions and interactions. Specifically, the face does not move and the hands are rigid. To facilitate the analysis of human actions, interactions and emotions, we have developed a new 3D model of human body pose, hand pose, and facial expression that we estimate from a single monocular image. To achieve this, we use thousands of 3D scans to train a unified, 3D model of the human body, SMPL-X, that extends SMPL with fully articulated hands and an expressive face. We estimate the parameters of SMPL-X directly from images. Specifically, we estimate 2D image features bottom-up and then optimize the SMPL-X model parameters to fit the the features top-down. This is a step towards automatic expressive human capture from monocular RGB data.

3.4 Morphable Texture Coordinates

Volker Blanz (Universität Siegen, DE)

In 3D Morphable Models, the assignment of texture and surface structures to vertices is usually permanent. The talk presents a method that slides textures and displacement maps along the surface. It proposes a linear model of texture coordinates and is illustrated on the example of eyeball rotation.

3.5 Modeling, Reconstruction, and Animation of 3D Faces

Timo Bolkart (MPI für Intelligente Systeme – Tübingen, DE)

Joint work of Timo Bolkart, Anurag Ranjan, Soubhik Sanyal, Michael J. Black, Haiwen Feng, Daniel Cudeiro, Cassidy Laidlaw

Learned 3D representations of human faces are useful for computer vision problems such as 3D face reconstruction from images, as well as graphics applications such as character generation and animation.

Traditional models learn a linear or multilinear latent representation of a face. Due to this linearity, they cannot capture extreme deformations and non-linear expressions. Our convolutional mesh autoencoder (CoMA) [1] applies spectral graph convolutions to the mesh surface and introduces mesh sampling operations to enable a hierarchical mesh representation that captures non-linear shape and expression variations in multiple scales. Compared to traditional methods, CoMA requires 75% fewer parameters and reaches a 50% lower reconstruction error.

Second, we present RingNet [3] to reconstruct 3D faces from single images without any 2D-to-3D supervision. Our key observation is that an individual's face shape is constant across images, regardless of expression, pose, lighting, etc. RingNet leverages multiple images of a person and automatically detected 2D face features. It uses a novel loss that encourages the face shape to be similar when the identity is the same and different for different people. We achieve invariance to expression by representing the face using the statistical FLAME model [2]. Once trained, our method takes a single image and outputs the parameters of FLAME, which can be readily animated.

Audio-driven facial animation from audio has been widely explored, but achieving realistic, human-like performance is still unsolved. This is due to the lack of available 3D datasets, models, and standard evaluation metrics. We introduce a novel 3D speech dataset (12 subjects, 40 sentences each) and train a model that animates 3D faces from speech. Our learned Voice Operated Character Animation model (VOCA) [4] takes any speech signal as input (from any language) and then animates a wide range of adult faces, not seen during training. This makes VOCA suitable for tasks like in-game video, virtual reality avatars, or any scenario when the speaker, speech, or language is not known in advance.

References

1 A. Ranjan, T. Bolkart, S. Sanyal, M. J. Black, *Generating 3D faces using Convolutional Mesh Autoencoders*, ECCV 2018.
2 T. Li, T. Bolkart, M. J. Black, H. Li, J. Romero, *Learning a model of facial shape and expression from 4D scans*, Siggraph Asia 2017.
3 S. Sanyal, T. Bolkart, H. Feng, M. J. Black, *Learning to Regress 3D Face Shape and Expression from an Image without 3D Supervision*, CVPR 2019.
4 D. Cudeiro, T. Bolkart, C. Laidlaw, A. Ranjan, M. J. Black, *Capture, Learning, and Synthesis of 3D Speaking Styles*, CVPR 2019.

3.6 Attributes, Illumination and Occlusion

Bernhard Egger (MIT – Cambridge, US)

Main reference Bernhard Egger: "Semantic Morphable Models" PhD Thesis, University of Basel, 2017.
URL https://doi.org/10.5451/unibas-006722192

In my presentation, I was talking about research challenges that were otherwise not covered in the seminar. For occlusions, we proposed a joint segmentation and model adaptation framework [1]. To initialize this hard optimization task we rely on a RANSAC based robust illumination estimation. An illumination prior from real-world images is estimated and arises as a nice side product. We also built a first joint shape, albedo and attribute model using Copula Component Analysis and use it for both Analysis and Synthesis [2, 3]. I proposed that all in our community should focus on obvious problems like occlusions.

References

1 Bernhard Egger, Sandro Schönborn, Andreas Schneider, Adam Kortylewski, Andreas Morel-Forster, Clemens Blumer, Thomas Vetter: *Occlusion-Aware 3D Morphable Models and an Illumination Prior for Face Image Analysis*. International Journal of Computer Vision 126(12): 1269-1287 (2018)
2 Bernhard Egger, Dinu Kaufmann, Sandro Schönborn, Volker Roth, Thomas Vetter: *Copula Eigenfaces – Semiparametric Principal Component Analysis for Facial Appearance*

Modeling 11th International Conference on Computer Graphics Theory and Applications (GRAPP), February 27-29, 2016

3 Bernhard Egger, Dinu Kaufmann, Sandro Schönborn, Volker Roth, Thomas Vetter: *Copula Eigenfaces with Attributes: Semiparametric Principal Component Analysis for a Combined Color, Shape and Attribute Model* In International Joint Conference on Computer Vision, Imaging and Computer Graphics (pp. 95-112). Springer, Cham (2016, February). Communications in Computer and Information Science book series (CCIS, volume 693), 2017

3.7 Interaction between invariant structures for shape analysis

Ron Kimmel (Technion – Haifa, IL)

A classical approach for surface classification is to find a compact algebraic representation for each surface that would be similar for objects within the same class and preserve dissimilarities between classes. Self functional maps were suggested by Halimi and the lecturer as a surface representation that satisfies these properties, translating the geometric problem of surface classification into an algebraic form of classifying matrices. The proposed map transforms a given surface into a universal isometry invariant form defined by a unique matrix. The suggested representation is realized by applying the functional maps framework to map the surface into itself. The idea is to use two different metric spaces of the same surface for which the functional map serves as a signature. As an example we suggested the regular and the scale invariant surface laplacian operators to construct two families of eigenfunctions. The result is a matrix that encodes the interaction between the eigenfunctions resulted from two different Riemannian manifolds of the same surface. Using this representation, geometric shape similarity is converted into algebraic distances between matrices. I will also comment on some of our efforts to migrate geometry into the arena of deep learning, in a sense learning to understand.

3.8 Can we resolve the limitations of deep learning with generative object models?

Adam Edward Kortylewski (Johns Hopkins Univ. – Baltimore, US)

This talk describes major limitations of current deep learning approaches to facial image analysis such as the lack of generalization from biased training data and the sensitivity to partial occlusion. I will discuss the relevant work of our group leveraging synthetically generated face images for overcoming those limitations [1, 2, 3]. Towards the end of the talk, I will hypothesize that integrating generative object models – such as 3DMMs – into deep neural networks would provide a means for overcoming those limitations.

References

1 Kortylewski, A., Egger, B., Schneider, A., Gerig, T., Morel-Forster, A., Vetter, T. (2018). Empirically analyzing the effect of dataset biases on deep face recognition systems. In Pro-

ceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (pp. 2093-2102).
2 Kortylewski, A., Schneider, A., Gerig, T., Egger, B., Morel-Forster, A., Vetter, T. (2018). Training deep face recognition systems with synthetic data. arXiv preprint arXiv:1802.05891.
3 Kortylewski, A., Egger, B., Schneider, A., Gerig, T., Morel-Forster, A., Vetter, T. (2019). Analyzing and Reducing the Damage of Dataset Bias to Face Recognition with Synthetic Data. Conference on Computer Vision and Pattern Recognition (CVPR) Workshops

3.9 Data Driven Inversion of Faces

Yeara Kozlov (ETH Zürich, CH)

Facial animation is one of the most challenging problems in computer graphics, and it is often solved using linear heuristics like blend-shape rigging. More expressive approaches like physical simulation have emerged, but these methods are very difficult to tune, especially when simulating a real actor's face. We propose to use a simple finite element volume for face animation and present an instrument free, non-intrusive method for recovering the required simulation parameters. Our method involves reconstructing a very small number of head poses of the actor in 3D, where the head poses span different configurations of force directions due to gravity. Our algorithm can then automatically recover both the gravity-free rest shape of the face as well as the spatially-varying physical material stiffness such that a simulation will match the captured targets as closely as possible. We present preliminary results and discuss the challenges in using our method on faces.

3.10 Inside and Outside the Scanner Room: On How to Capture and Model People from Data.

Gerard Pons-Moll (MPI für Informatik – Saarbrücken, DE)

The research community has made significant progress in modeling people's faces, hands and bodies from data, and currently several models are publicly available. The standard approach is to capture data coming from 3D/4D scanners and learn models from it. Such an approach provides a very useful first step, but it does not scale to the real world. If we want to learn rich models that include clothing, interactions of people, and interactions with the environment geometry, we require new approaches that learn from ubiquitous data such as plain RGB-images and video. In this talk, I will describe some of our works on capturing and learning models of human pose, shape, soft-tissue, and clothing from 3D scans as well as from plain video. I will conclude the talk outlining the next challenges in building digital humans and perceiving them from sensory data.

3.11 Isospectralization, or How to Hear Shape, Style, and Correspondence

Emanuele Rodolà (Sapienza University of Rome, IT)

Main reference Luca Cosmo, Mikhail Panine, Arianna Rampini, Maks Ovsjanikov, Michael Bronstein, Emanuele Rodolà: "Isospectralization, or how to hear shape, style, and correspondence". to appear in Proc. CVPR 2019, Long Beach, CA, USA, 2019.

The question whether one can recover the shape of a geometric object from its Laplacian spectrum ('hear the shape of the drum') is a classical problem in spectral geometry with a broad range of implications and applications. While theoretically the answer to this question is negative (there exist examples of iso-spectral but non-isometric manifolds) little is known about the practical possibility of using the spectrum for shape reconstruction and optimization. In this talk, I will introduce a numerical procedure called isospectralization [1], consisting of deforming one shape to make its Laplacian spectrum match that of another. By implementing isospectralization using modern differentiable programming techniques, we showed that the *practical* problem of recovering shapes from the Laplacian spectrum is solvable. I will finally exemplify the applications of this procedure in some of the classical and notoriously hard problems in geometry processing, computer vision, and graphics such as shape reconstruction, style transfer, and non-isometric shape matching.

References

1 Luca Cosmo, Mikhail Panine, Arianna Rampini, Maks Ovsjanikov, Michael Bronstein, Emanuele Rodolà. *Isospectralization, or how to hear shape, style, and correspondence*. Proc. CVPR, 2019

3.12 Deep 3D Morphable Models

Sami Romdhani (IDEMIA, FR)

Recently, Generative Adversarial Networks (GANs) have addressed a lot of attention. Indeed, this is because they are capable to generate synthetic face images at an unprecedented level of realism and quality. One of the main limitation of the GANs, though, is their inability to let the user control the type of face image generated. For instance, even though a face with some pose or some illumination can be generated, there is no control over these parameters. Hence, it is not possible to generate a face image of a random individual at different poses or different illumination conditions. This is, however, something that the 3D Morphable Model does very well, by leveraging the 3D equations grounded in physics. Hence, there is a need to build a generator that can synthesize highly realistic images as GAN can, while giving control over semantic parameters such as pose or expression, as 3D MM can.

3.13 Deep Learning from Unstructured, Uncalibrated Views

Javier Romero (Amazon Research – Barcelona, ES)

It is said that one of the most important professors in our field once described the three main problems in computer vision to be correspondences, correspondences, and correspondences. Deep networks have attacked successfully the problem of extracting correspondences between two images in a number of problems (optical flow, stereo matching, etc). However, there is still little work on deep networks producing coherent output (keypoint estimation, segmentation) representations when presented with unstructured, non-calibrated multiview RGB data. This work probably requires deep networks to either consume some notion of correspondences or producing it internally in a way that its estimations are preserved across them. In a world in which it is common to have multiple images or videos from a particular object of interest, it is important to let neural networks exploit effectively this input. I would like to present this challenging, unsolved question to the audience of the workshop, with the focus on extracting key points and dense registrations of people from multiple images.

3.14 Top-Down Human Digitization In the Wild

Shunsuke Saito (USC – Los Angeles, US)

3D morphable models have been a popular choice for compact facial shape and appearance representation for two decades. However, extending such representation to non-parametric structures such as hair and clothed human bodies poses a significant challenge due to their immense variation in shape and topology. To this end, we introduce an effective and unified data representation based on deep learning that can represent the entire human body, including the face, hair, body, and clothing. I will present several possible representations for human digitization and show several highlights of our recent progress on high-fidelity geometry/texture using deep convolutional neural networks. I will also discuss the pros and cons when inferring both parametric and non-parametric data when modeling humans.

3.15 Three Ambiguities

William Smith (University of York, GB)

Joint work of William Smith, Anil Bas, Ye Yu, Chao Zhang, Behrend Heeren, Martin Rumpf

The problem of providing a physical explanation of an image, i.e. inverse rendering geometry, reflectance and illumination from a single image, is an ill-posed problem. In this talk, I will consider three specific ambiguities that arise. First, when using a morphable model to solve the shape-from-correspondence problem (e.g. fitting a model to landmarks) there is a nonlinear subspace of 3D shapes that all project to the given 2D positions. In particular, this is significant when camera calibration is unknown and hence distance from the camera to

object is unconstrained. Second, the general task of single image inverse rendering is highly ambiguous. For example, the shaded versus painted hypothesis and ambiguity between low-frequency lighting and texture effects. I described InverseRenderNet, a self-supervised deep neural network that learns this task by exploiting a prior on natural illumination and multiview supervision to ensure photometric invariants are consistently estimated across lighting. Third, I considered the problem of dealing with rigid body motion superposed on top of nonlinear shape deformation. Building a statistical model of the intrinsic shape variation, invariant to how the shapes are aligned requires RBM-invariant modeling. I proposed a hybrid statistical/physical model that uses the discrete shell energy as a local distance measure and time-discrete principal geodesic analysis to build the statistical model.

References

1 Anil Bas and William AP Smith. What does 2d geometric information really tell us about 3d face shape? *arXiv preprint arXiv:1708.06703*, 2017.
2 Behrend Heeren, Chao Zhang, Martin Rumpf, and William Smith. Principal geodesic analysis in the space of discrete shells. *Computer Graphics Forum (Proceedings of SGP)*, 37(5):173–184, 2018.
3 William AP Smith. The perspective face shape ambiguity. In *Perspectives in Shape Analysis*, pages 299–319. Springer, 2016.
4 Ye Yu and William AP Smith. Inverserendernet: Learning single image inverse rendering. In *Proc. CVPR*, 2019.

3.16 Building 3D Morphable Face Models from Videos

Ayush Tewari (MPI für Informatik – Saarbrücken, DE)

Joint work of Ayush Tewari, Florian Bernard, Pablo Garrido, Gaurav Bharaj, Mohamed Elgharib, Hans-Peter Seidel, Patrick Pérez, Michael Zollhöfer, Christian Theobalt
Main reference Ayush Tewari, Florian Bernard, Pablo Garrido, Gaurav Bharaj, Mohamed Elgharib, Hans-Peter Seidel, Patrick Pérez, Michael Zollhöfer, Christian Theobalt: "FML: Face Model Learning from Videos". The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA 2019.

Reconstructing the 3D geometry and appearance of faces from monocular images is challenging, as it requires inverting the image formation process. Parametric face models (3DMMs), built using limited 3D scan data are used to constrain this ill-posed problem. However, these models lack details and can only represent a very limited subset of identities. I will talk about building large-scale face models just from videos, which allows for reconstruction in-the-wild. The presented method learns to reconstruct all faces in a large video dataset while building a low dimensional 3D face model at the same time. Models built from videos can generalize better across identities, compared to classical morphable models, due to the abundance of video data on the internet. I will talk about how this idea can be used for building higher quality and detailed 3D morphable models of faces.

3.17 To Optimize, To Learn, Or to Integrate

Christian Theobalt (MPI für Informatik – Saarbrücken, DE)

URL http://gvv.mpi-inf.mpg.de/

Reconstructing models of the real world in motion from sparse or even single camera images is a focus of research in my group Graphics, Vision and Video at the Max-Planck-Institute for Informatics in Saarbruecken. In particular, reconstructing the space-time coherent geometry, deformation, material and illumination of the scene is of interest. Using our work on monocular face performance capture, I discuss several classes of algorithms developed for models of the world in motion from a single color camera. In particular, I visited model-based analysis-by-synthesis approaches, learning-based regression or classification approaches, as well as new algorithms we developed that deeply integrate model-based and deep learning-based algorithms in an end-to-end trainable manner. I also discuss the pros and cons of the different classes of methods and opened up the question of how the deeply integrated approaches could be able to drive a real-world reconstruction loop.

References

1 P. Garrido, M. Zollhöfer, C. Wu, D. Bradley, P. Perez, T. Beeler, and C. Theobalt Corrective 3D reconstruction of lips from monocular video ACM Transactions on Graphics (Proc. of SIGGRAPH Asia 2016)
2 J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt and M. Nießner Face2Face: Real-time Face Capture and Reenactment of RGB Videos Proc. Computer Vision and Pattern Recognition (Oral) (CVPR 2016)
3 A. Tewari, M. Zollhöfer, H. Kim, P. Garrido , F. Bernard, P. Perez and C. Theobalt MoFA: Model-based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction International Conference on Computer Vision (ICCV), 2017 (Oral)
4 A. Tewari, M. Zollhöfer, P. Garrido , F. Bernard, H. Kim, P. Perez and C. Theobalt Self-supervised Multi-level Face Model Learning for Monocular Reconstruction at over 250 Hz, Proc. CVPR 2018 (Oral)
5 H. Kim, M. Zollhöfer, A. Tewari, J. Thies, C. Richardt and C. Theobalt InverseFaceNet: Deep Single-Shot Inverse Face Rendering From A Single Image Computer Vision and Pattern Recognition (CVPR), 2018
6 H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Nießner, P. Perez, C. Richardt, M. Zollhöfer and C. Theobalt, Deep Video Portraits, ACM Transactions on Graphics (Proc. SIGGRAPH 2018)
7 A. Tewari, F. Bernard, P. Garrido, G. Bharaj, M. Elgharib, H-P. Seidel, P. Perez, M. Zollhöfer and C. Theobalt, FML: Face Model Learning from Videos, CVPR 2019 (Oral)

3.18 Probabilistic Morphable Models

Thomas Vetter (Universität Basel, CH)

Main reference Thomas Gerig, Andreas Morel-Forster, Clemens Blumer, Bernhard Egger, Marcel Lüthi, Sandro Schönborn, Thomas Vetter: "Morphable Face Models – An Open Framework", in Proc. of the 13th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2018, Xi'an, China, May 15-19, 2018, pp. 75–82, IEEE Computer Society, 2018.
URL https://doi.org/10.1109/FG.2018.00021

Probabilistic Morphable Models extend the classical Morphable Model approach in two terms. First the shape and texture variability of the models is formalized as Gaussian Process and second, the model to target registration utilizes data-driven Markov Chain Monte Carlo optimization. The step from PCA based representations to Gaussian Processes unifies several different deformations models, such as spline, free-form or data based to a single formal description. This is of conceptual importance since it connects the rich field of Gaussian Processes to the Morphable Model approach. On the practical side, it is now sufficient to implement only a single software framework for a whole class of different deformation models. The second novelty, the stochastic optimization framework aims for two main improvements. The model fitting problem is inherently difficult since its non-convexity and the high dimensional parameter space. The model fitting starts in general by some initial parameter guess. But the local optima problem makes it necessary to consider a certain uncertainty of these initialization steps to avoid that a bad initial guess hinders to overall optimization. Another shortcoming of previous methods is that the optimization results only in a single "optimal" value but does not inform about the quality of the result or similar results. We propose to compute the full posterior parameter distribution for a given target image. This leads to a full Bayesian Approach for model to image registration. We propose to compute an approximation of the full posterior based on a stochastic optimization framework using Metropolis-Hastings Filtering. This approach does not only inform about the certainty of the solution it also enables an easy approach to integrate uncertain guesses for the initialization of the optimization procedure. Overall our Probabilistic Morphable Model technique is a fully probabilistic approach enabling Bayesian inference on images.

References

1 Lüthi, M., Gerig, T., Jud, C., & Vetter, T. *Gaussian process morphable models*. IEEE transactions on pattern analysis and machine intelligence, 40(8), 1860–187, (2018).
2 Schönborn, S., Egger, B., Morel-Forster, A., & Vetter, T. *Markov chain monte carlo for automated face image analysis*. International Journal of Computer Vision, 123(2), 160–18, (2017).
3 Gerig, T., Morel-Forster, A., Blumer, C., Egger, B., Luthi, M., Schönborn, S., & Vetter, T. *Morphable face models-an open framework*. In 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018) (pp. 75–82). IEEE, (2018, May).

3.19 Building 3D Morphable Models with Minimal Supervision

Stefanie Wuhrer (INRIA – Grenoble, FR)

Joint work of Timo Bolkart, Victoria Fernandez Abrevaya, Adnane Boukhayma, Edmond Boyer, Stefanie Wuhrer

3D Morphable Models (3DMMs) are commonly used in many virtual and augmented reality applications. Recently, a number of static and dynamic databases of 3D face scans have been published, whose acquisition was facilitated by increasingly affordable 3D face scanners. In spite of this progress, building high-quality 3DMMs that can benefit from these datasets remains challenging in practice as the raw 3D face scans need to be registered. We are interested in models that decouple different factors of variation (e.g. identity, expression or age), and in this case, the data additionally needs to be labeled. As inaccuracies in the registrations and labeling directly deteriorate the quality of the resulting 3DMM, these steps are often completed with manual interaction in practice. The goal of our work is to build 3DMMs with minimal supervision. To achieve this, we have developed groupwise methods that take advantage of the full training database. I will present our works that allow to improve registration accuracy by taking advantage of the minimum description length principle [1]. We will further discuss how autoencoders [2] and generative adversarial networks [3] can be used to efficiently train from datasets that combine existing 3D face databases where only sparse label information is available. For the second part of this presentation, I will give an outlook on upcoming challenges. A first challenging problem is to learn models of dynamic 3D face deformations. In this scenario, minimal supervision is critical. A second open problem we will discuss is how to model correlations between different dynamically deforming body parts, such as face and tongue movements.

References

1 Timo Bolkart, Stefanie Wuhrer. A Groupwise Multilinear Correspondence Optimization for 3D Faces. ICCV 2015.
2 Victoria Fernández Abrevaya, Stefanie Wuhrer, Edmond Boyer. Multilinear Autoencoder for 3D Face Model Learning. WACV 2018.
3 Victoria Fernandez Abrevaya, Adnane Boukhayma, Stefanie Wuhrer, Edmond Boyer. A Generative 3D Facial Model by Adversarial Training. arXiv:1902.03619, 2019.

3.20 Learning 2D and 3D Deep Generative Models

Michael Zollhöfer (Stanford University, US)

Joint work of Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Nießner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, Christian Theobalt, Vincent Sitzmann, Felix Heide, Gordon Wetzstein

Generative 2D rendering-to-video translation networks that take renderings of parametric model instances as input enable to bridge the domain gap between synthetic computer graphics and real imagery. With the ability to freely control the underlying parametric face model, we are able to demonstrate a large variety of video rewrite applications. For instance, we can reenact the full head using interactive user-controlled editing and realize high-fidelity visual dubbing. While this approach of bridging the domain gap in 2D screen

space enables several existing applications, it has limited generalization capabilities and does not easily scale to large head rotations. We address this lack of 3D understanding of such generative neural networks by introducing a persistent 3D feature embedding. At its core, our approach is based on a Cartesian 3D grid of embedded features that learn to make use of the underlying 3D scene structure. Our approach thus combines insights from 3D geometric computer vision with recent advances in learning image-to-image mappings based on adversarial loss functions.

References

1 Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Nießner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, Christian Theobalt. *Deep video portraits*. ACM Trans. Graph. 37(4): 163:1–163:14 (2018)

2 Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, Michael Zollhöfer. *DeepVoxels: Learning Persistent 3D Feature Embeddings*. CoRR abs/1812.01024 (2018)

3.21 Modeling Animal Shape

Silvia Zuffi (IMATI – Milano, IT)

I will present our recent work on modeling animal shape, the SMAL model. SMAL model is a 3D articulated model that can represent animals including lions, tigers, horses, cows, hippos, dogs. We learn the model from a small set of 3D scans of toy figurines in arbitrary poses that we align to a common template using a novel approach. From the aligned toys, brought into a reference pose, we learn a linear shape space over a large variety of animal species.

4 Working groups

4.1 Discussion: The Ehics and Regulation of Photo-realistic Human Generation

Michael J. Black (MPI für Intelligente Systeme – Tübingen, DE)

Advances in 3DMMs and related technologies have created the ability to synthesize images of people that are indistinguishable from real images, to manipulate photos of people to change identity or expression, and to edit videos to change what people are saying. Photo-realistic face generation and manipulation have the ability to change our perception of history and our perception of each other. By doing so, it has the power to change behavior and the future. There are both positive and negative applications of this technology and the question is if and how it should be regulated. In this session, we explored several case studies at different levels from regulating the research, the technology, the output of the technology, particular uses, and users themselves. As an outcome of this session, there are plans to

write a position paper and to possibly organize a Dagstuhl seminar focused on the ethical questions that would bring together people with different levels of expertise from science, business, government, history, psychology, sociology, law enforcement, and ethics.

4.2 Discussion: 3D Morphable Models – 10 Years Perspective

Patrik Huber (University of Surrey, GB)

In this session, the topic was to discuss where our field might be in 10 years time. The discussion was started with a comment that recent work done in Unity by Vicon, puppeteering a Chinese woman, involved a lot of manual work, but the results are much better than what our computer vision algorithms can currently produce. In 10 years time, we can perhaps achieve such quality with computer vision algorithms and with much fewer manual intervention. Another way would be to look at the question from the future and ask ourselves what the additional value is that our tech can provide. Simple answers would be that there are many uses in health care, for example, entertainment specifically for elderly or lonely people, or to detect when a person falls – and then we could work towards reliable algorithms for these tasks.

It was pointed out that in vision, we would like to understand images. We work hard to generate priors for image analysis or generation, here with the 3D morphable models. We can model the world in the way we understand it from physics (which would be the more traditional way with morphable models and a rendering pipeline), or, being done more and more recently, model it with some sort of function that we don't fully understand anymore (which would correspond to many of today's deep-learning based approaches). Traditionally, we have been able to understand each parameter of our models and have an intuition about them, and they were not just abstract latent variables of a neural network. One big question, therefore, is how we will create priors in the future: Will we have machines creating priors automatically, or will we still teach the machines? One general point of the audience was that we are likely to continue the "black-box-way", but develop the tools to understand those more abstract models and latent spaces much better. There will likely also be a process to record and diagnose failures of those systems, and then improve them accordingly – much like in today's complex airplane systems, where it is also nigh impossible to test for all potentially occurring events. It was further pointed out how good of a representation linear models and PCA specifically are. One needs only a small number of data points to represent quite complex things, and there will always be cases where only few data is available, so those simple models are unlikely to disappear.

It was also briefly discussed how in the last few years, the community got a lot better at optimization, for example, 3D human pose estimation. A lot of information can be estimated already with a few sparse points, for example, facial landmarks or body joints. One important and continuing research direction is selecting descriptive features of an object, where currently neural networks do a very good job at learning the feature selection. In that case, the prior is contained in the training data (often with biases we may and may not be aware of), which ties back to the earlier discussion about learning priors and whether we will still be using "manual" priors in a few years time. In essence, where a few decades ago students were tweaking optimization parameters, in a very similar way, we are today tweaking optimization parameters of neural networks.

A final point discussed was that most current research only tackles one specific task, like face or full-body reconstruction, in isolation. There is research starting to emerge that combines those individual tasks, and it has been brought up whether to put together those individual methods is only an engineering task, or whether it is a fundamental research question that also requires new methods and models. It is also still an unanswered question whether we can use the same technology, for example, full-body models, to model the whole world. We are currently learning all these specialized models, which coincides with the mental models that exist in our brains, but at some points, we have to put those separate bits together. In our brains we have hierarchical concepts that we seamlessly relate and connect with each other – for example a chair can be seen as a global object, with many different varieties, and if we inspect a chair much closer, we might go down to the level of the materials that the chair consists of, which is a different generic concept. This is something that we will likely have to address more deeply in the computer vision community. Currently, most of those models that we are using are also "forward-only", meaning they are trained once, and then when deployed, are not able to adapt or learn new things.

The discussion didn't come to too much of a conclusion, but the audience would probably agree that learning priors from data will continue to be a hot topic in the next 10 years, and models and algorithms for specific tasks will be put together to yield a more holistic reconstruction of human bodies including detailed reconstruction of their faces, hands, and clothing, up to a level that is currently only achievable by computer graphics with laboursome manual work.

4.3 Discussion: Representation Group 1

Adam Edward Kortylewski (Johns Hopkins Univ. – Baltimore, US)

Main reference Adam Kortylewski, Bernhard Egger, Andreas Schneider, Thomas Gerig, Andreas Morel-Forster, Thomas Vetter: "Empirically Analyzing the Effect of Dataset Biases on Deep Face Recognition Systems", in Proc. of the 2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2018, Salt Lake City, UT, USA, June 18-22, 2018, pp. 2093–2102, IEEE Computer Society, 2018.
URL http://dx.doi.org/10.1109/CVPRW.2018.00283

We discussed the properties of good representations and found that this is highly dependent on the down-stream task. An open research question is whether universal representations can be learned that suit multiple-down stream tasks (see Multi-task learning literature). Finally, it is important to be aware of a trade-off in the properties of representations w.r.t the interpretability of the representation. An interpretable representation is likely not the most efficient possible. This trade-off should be taken into account in the discussion on interpretable representations.

4.4 Discussion: Representation Group 2

Yeara Kozlov (ETH Zürich, CH)

The representations used in computer graphics and vision influence the available toolset and the problems the community chooses to address. A good representation would have the following properties: semantic, compact, complete, has specificity, is differentiable, unique and allow for high-quality rendering. A representation that might be useful to solve technical problems might not be able to give insight into the problem of vision and vice-versa. It was suggested that the community should revisit historical representations such that we can deal with challenging problems that are not currently addressed.

4.5 Discussion: Levels of Detail for Modeling

Javier Romero (Amazon Research – Barcelona, ES)

We talked about three topics related to levels of detail. One discussion was focused on geometry statistics at different levels of detail. Classic 3D Morphable models have done a good job in fitting and sampling low-frequency details but typically fail at modeling high frequencies. Adversarial nets, one of the current solutions to increase details in reconstructions, exploits exactly the differences between real crisp data and smooth models by encouraging the generator to make them indistinguishable. There are also classic solutions to increase detail in models, like multilinear wavelet models (https://arxiv.org/pdf/1401.2818.pdf). These models have shown good fitting power but are not suitable for sampling. A second discussed topic was the dependency between the level of detail and the task to be solved. Hollywood movies require a great level of detail to achieve photorealism. However, synthetic data generated for training deep networks probably have very different requirements. It has been shown that one of the important aspects of synthetic data for this problem is the blending between foreground and background (https://arxiv.org/pdf/1710.10710.pdf), but it's unclear what is important for more fine-grained tasks like a detailed 3D reconstruction of faces and bodies. Intuitively, it seems like reconstruction would benefit from material estimation in the synthetic assets, but maybe not to the level of subsurface scattering or facial microstructures. Finally, a third topic we discussed is the perceived level of detail by humans. The field focuses on capturing and reproducing high frequencies in the face with high accuracy, although it is unclear how much of that detail can be kept in memory by us humans. An individual would still be recognizable even when rendered with the wrinkles from a different person, as soon as the wrinkles are consistent with his age and general facial features. This maybe suggests adversarial functions that try to push generators for consistency rather than precise high-frequency matching.

4.6 Discussion: Academia and Industry

Shunsuke Saito (USC – Los Angeles, US)

In this session, we discussed the current issues and future of relationship between industry and academia. Throughout the discussion, we all agreed that we should actively seek the collaboration on the premise that society should support different interests. The main discussion point is three-fold: 1) openness vs protectionism 2) human resource 3) education. First, while academia prefers openness to facilitate knowledge creation as community, industry tends to keep knowledge within a company to keep superiority in the market. Therefore it is essential to incentivize industry to release their knowledge and data (e.g., making a challenge/competition on open questions in a conference). Another short-term solution is to encourage targeted collaboration, where a company provides a specific university or group with their proprietary data to solve open questions together. Such collaboration can benefit both academia and industry by bringing another aspect to the problems. Secondly regarding human resources, it has been increasingly difficult for universities to hire not only competitive students but also senior researchers due to large gap in terms of monetary rewards. However, given the fact that internet bubble in early 90s created the exact same situation between industry and academia, we conclude that we should learn how to handle the situation from the history rather than finding out a solution from scratch. Lastly, industry has been more influential on academia by occupying committee members of a conference or creating new demand in the market, which creates pressure on universities to change the curriculum. To avoid conflict of interest and encourage diversity in research, the entire research community may need to take responsibility on educating junior researchers by providing guideline on this matter including paper review. In conclusion, we view siblings as an ideal form of relationship between industry and academia, where they can play in harmony but do not always seek the same interest.

4.7 Discussion: Inverse Rendering

Ayush Tewari (MPI für Informatik – Saarbrücken, DE)

The goal of inverse rendering is to estimate parameters of a forward rendering model. Estimating these rendering parameters from a single image is difficult, because of the ambiguities between different parameters. Priors (for e.g. 3D Morphable Models) of different objects help in resolving many of these ambiguities. Certain applications like image editing might not require accurate estimation of different parameters. In that case, we can make strong assumptions about the world. However, the accurate inverse can be required for other applications. Differentiable rendering is the key to solve inverse rendering problems. However, rendering techniques are usually non-differentiable. Occlusions and gradients around the boundary of the 3D surface lead to non-differentiability of the rendering function. Rasterization techniques are fast but cannot deal with transparency and complex light effects. Ray tracing can deal with these phenomena but is typically slower. However, with faster hardware and better software, differentiable raytracing could be widely used in the near future.

Participants

- Thabo Beeler
Disney Research – Zürich, CH
- Florian Bernard
MPI für Informatik – Saarbrücken, DE
- Michael J. Black
MPI für Intelligente Systeme – Tübingen, DE
- Volker Blanz
Universität Siegen, DE
- Timo Bolkart
MPI für Intelligente Systeme – Tübingen, DE
- Bernhard Egger
MIT – Cambridge, US
- Victoria Fernandez Abrevaya
INRIA – Grenoble, FR
- Patrik Huber
University of Surrey, GB
- Ron Kimmel
Technion – Haifa, IL
- Tatsuro Koizumi
University of York, GB
- Adam Edward Kortylewski
Johns Hopkins Univ. – Baltimore, US
- Yeara Kozlov
ETH Zürich, CH
- Andreas Morel-Forster
Universität Basel, CH
- Nick Pears
University of York, GB
- Gerard Pons-Moll
MPI für Informatik – Saarbrücken, DE
- Emanuele Rodolà
Sapienza University of Rome, IT
- Sami Romdhani
IDEMIA, FR
- Javier Romero
Amazon Research – Barcelona, ES
- Shunsuke Saito
USC – Los Angeles, US
- William Smith
University of York, GB
- Ayush Tewari
MPI für Informatik – Saarbrücken, DE
- Christian Theobalt
MPI für Informatik – Saarbrücken, DE
- Thomas Vetter
Universität Basel, CH
- Stefanie Wuhrer
INRIA – Grenoble, FR
- Michael Zollhöfer
Stanford University, US
- Silvia Zuffi
IMATI – Milano, IT