OTMAR HILLIGES | RESEARCH STATEMENT

otmar.hilliges@inf.ethz.ch AIT Lab, ETH Zurich, Switzerland ait.ethz.ch

OVERVIEW

I am currently an Associate Professor of computer science at ETH Zurich, where I lead the AIT lab (AIT Lab) and serve as head of the institute for intelligent interactive systems (IIS). My research interests lie at the intersection of computer vision, machine learning and human computer interaction (HCI). My main mission is to endow artificial systems with human level perceptual capabilities and to leverage such systems to develop new ways for humans to interact with complex interactive systems (computers, wearables, robots). Prior to joining ETH, I was a Researcher at Microsoft Research Cambridge (2012-2013). I earned degrees in computer science from Technische Universität München, Germany (Diplom equiv. MSc 2004) and LMU München, Germany (PhD 2009). I have spent two further years as a postdoc at Microsoft Research Cambridge (2010-2012). My work has resulted in more than 100 peer-reviewed papers in the major venues on computer vision, HCI and computer graphics. 20+ patents have been filed in my name on a variety of subjects from surface reconstruction to AR/VR. Amongst other sources of funding, I am a recipient of the prestigious ERC starting grant.

Research agenda: My research spans computer vision, human-computer-interaction and machine learning. I research new methods and algorithms to interpret human motion and other activity from low-level sensor data such as images or body-worn sensors. Areas of interest include but are not limited to the estimation of pose and shape of the human body and hands, estimation of gaze direction and the analysis and generation of complex, non-linear human activity at various levels of abstraction. For example, predicting the motion of a human over several seconds enables artificial agents such as a service robot or self-driving vehicle to reason about their behavior and thus to plan and act in accordance. With such capabilities future artificial systems can become more adaptive and therefore useful for humans. Hence, much of my work is dedicated to the development of new data-driven algorithms and methods for the machine perception and generation of high-level human activity. This includes complex day-to-day tasks ranging from locomotion and other forms of physical motion to more goal oriented tasks such as cooking or other household activities and knowledge work orientated tasks such as reading and writing of text. From a methodological standpoint my work has contributed to the state-of-the art in timeseries analysis and deep generative modelling.

Outline of career to date: My career to-date has spanned academia and industry and has already had significant impact at both levels. During my PhD I studied algorithms for the detection and processing of user input into non-desktop computing systems, including touch enabled devices and mixed and augmented reality settings. As a post-doctoral researcher at Microsoft Research in Cambridge, UK I worked on several impactful research directions, including a depth aware optical-flow based input mechanism for spatial mixed reality applications [11] and was part of a larger effort to develop real-time methods for the 3D scene reconstruction for augmented reality [13, 19]. These publications have received more than 6000 citations and have played a significant role in starting the incubation period of the HoloLens project, now a commercial product from Microsoft. Furthermore, my work has pioneered much research in the direction of estimating human activity and input in mobile and on-the-go scenarios (e.g., [15, 28, 33]). Beyond the academic impact much of this work has had industrial impact as well for example, the ideas described in the UIST '16 paper by Wang et al. [33] now form the core of the recognition engine that ships with Google's Pixel phone. After joining ETH Zurich in 2013 I have built a word-class team working towards the goal of developing and studying ML-powered agents (wearables, robots, computers) that can perceive and reliably interpret sensory input (visual, acoustic, haptic) and reason about user intent, interest, state of knowledge and relevance. My work has been recognized via several best paper awards and honorable mentions at top international venues including ACM CHI, UIST and CSCW, IEEE IROS, ISMAR and 3DV. Moreover, I have recently been granted the prestigious ERC starting grant for my work on computational modelling of human activity and computational design of interactive technologies. I am very well connected and actively collaborates with many colleagues in computer vision, HCI and AI and am a core member of the ETH-MPI center for learning systems (CLS), core faculty of ETH's AI center (ai.ethz.ch), and a fellow of the European Lab for Learning and Intelligent Systems (ELLIS).

SPECIFIC RESEARCH INTERESTS

Human pose estimation and motion modelling: One of the most fundamental components in perceiving and understanding human activity is the estimation of the spatial configuration of the human body (its pose) over time. To this end the AIT lab has contributed many data-driven algorithms for the estimation of full-body (e.g., [29, 27, 26]) and hand-pose configurations from images (e.g., [31, 30]) and other sensors such as inertial-measurement units (IMUs) or even micro-wave radio frequency measurements. A specific focus lies on estimating these quantities in settings that require no or only lightweight instrumentation of the user and thus maybe applicable in real-world settings. For example, recognizing human activity using only the sensors already present on modern smartphones [28, 33], estimating the full-body pose from a sparse set of body-worn IMUs [12]. Similarly, my group recently has been instrumental in demonstrating the feasibility of estimating the full hand pose configuration from monocular images alone via leveraging multi-modal representations

of hands in a variational auto-encoder formulation [31], this area is now a very lively sub-area in computer vision. From a technical perspective a core specialty of my work is the integration of domain-insights into end-to-end trainable deep-learning techniques. For example, embedding of a spatio-temporal graph into an end-to-end trainable CNN to improve the temporal estimation of human poses in videos [29] or embedding of a kinematics model of the human hand to provide an inductive bias for 3D hand pose estimation [30]. Importantly, my work shows experimentally that such hybrid approaches typically outperform both pure learning approaches and traditional iterative optimization-based fitting algorithms. Lastly, I am also interested in the interplay between optimization, model-fitting and parameter estimation and machine-learning in the context of computer vision problems. For example, my team recently proposed a novel gradient-based optimization algorithm for human pose and shape estimation that leverages a neural network to predict a per-parameter update rule which allows for better registration, of details, an order of magnitude faster convergence time and helps in avoiding bad local minima [27]. Similarly, a recent paper has explored the use of neural image synthesis modules in gradient-based iterative 6D object pose estimation enabling instance-level pose estimation from images alone [5].

Eye-gaze estimation and synthesis: To furthermore be able to reason about unobservable high-level states such as tasks and intent it is crucial to remotely analyze the gaze of humans. To this end, the AIT lab has contributed several seminal papers over the last couple of years. Including datasets to study eye-gaze estimation under extreme head-angle and gaze variation [34] and to estimate both gaze and visual stimuli in unison [20]. Furthermore, my group has contributed several methods that combine anatomical models of the oculomotor system with deep neural networks for remote gaze estimation (e.g., [23, 24]). Moreover, I have studied the problem of cross-person gaze estimation in detail, that is to reason about the gaze direction in the presence of unobservable inter-personal differences which pose a major hurdle for highly accurate gaze tracking. For example, in Park et al. [22] a meta-learning formulation is proposed that allows for the personalization of a gaze estimation network with very few samples (as little as three calibration samples) yet achieves very high-accuracy personalized gaze estimates. More recently, a technique has been proposed to overcome these inter-personal differences without the need for any calibration samples. This is achieved by reasoning jointly about the gaze direction and the visual stimulus that is presented on screen [20]. Beyond low-level gaze estimation the group also studies how this information can be used in detecting relevance during decision-making from eye movements for UI adaptation [6] or to reason about user intent and to optimize mixed-reality user interfaces accordingly [7].

Generative modelling of high-level human activity: A cornerstone of human intelligence is the capability to imagine possible future events and thus the ability to plan under uncertainty. To enable artificial agents to successfully co-inhabit a man-made world it is important that such systems have the capability to reason about potential human future actions. Generative modelling of human activity has therefore been a cornerstone of my research agenda, to further the abilities of AI systems in this regard. At a low, bio-mechanical level this includes predicting the pose configuration of the human body over time [9, 3, 14]. At intermediate levels this includes modelling the appearance of faces or the periocular region under fine-grained control of head-pose or gaze direction. This is a challenging problem for traditional computer graphics approaches due to the highly complex structure of the periocular region and the resulting highly non-linear changes in appearance. The AIT lab has contributed several methods for this task based on generative adversarial networks (GANs) [10], and self-supervised disentangling algorithms [35]. Finally, my work has contributed several generative models that capture human activity at a high level of abstraction, that is types of activity that involve entirely unobservable cognitive processes such as modelling of handwritten text [4, 2], flow-charts and diagrams [1] and even of human task interleaving behavior [8].

Applications: AR, VR and Haptics: My research also aims at improving the means with which humans and machines interact. To this end we often leverage advanced machine perception and data-driven user modelling techniques to build advanced intelligent interactive systems. Of particular interest are non-desktop type of interaction paradigms such as sensor-driven mobile applications, Augmented and Virtual Reality. These are interesting because the information that is shown to the user has to be context sensitive to avoid information overload and to increase utility. To this end, my research has been instrumental in advancing the area of computational interaction, in which advanced sensing, adaptive user models and computational design methods are leveraged to adapt the UI at runtime. We have demonstrated the promise of this approach for collaborative user interfaces [21], for UI adaptation based on gaze behavior [6] and in mixed reality [18, 7]. Moreover, we have contributed several algorithms to computationally design and control of novel haptic feedback mechanisms for VR [25, 32, 17, 16].

FUTURE WORK

I am convinced that AI will play an important role in overcoming many of humanity's most urgent issues including urban congestion, rising healthcare costs, an aging society, and climate change. Many of the AI technologies that have a potentially positive impact on these issues rely on advanced computer vision and an understanding of human activity. For example, for self-driving cars to safely navigate crowded cities and be a viable transportation option, it must be able to detect, interpret, and predict human motion. An AI system that supports a medical team in the operating room must be able to interact with humans and understand how they manipulate medical tools. For a care robot to physically interact with its elderly patients, it must be able to predict intent and accurately see the shape and relative position of its patient. And for immersive telepresence systems to replace physical meetings, thereby reducing transportation-induced carbon

emissions, such systems must be able to reconstruct the shape and appearance of humans, even under partial observability, and synthesize realistic human motion. In short, future AI systems will have to be able to perceive human activity at a level that matches our own capabilities.

Therefore, my future work will continue to advance the state-of-the art in perceiving AI systems to an extent that it can be used to solve pressing, real-world problems. The goal of my work is to develop a holistic approach to human-centric machine perception. One that endows artificial agents with human level capabilities of perception, and the ability to reason about human activity in complex settings such that future AI possesses "3D common sense". Finally, I will continue to research new applications of such models and algorithms in computer-vision, graphics, robotics and mixed reality.

REFERENCES

[1] Emre Aksan, Thomas Deselaers, Andrea Tagliasacchi, and **Otmar Hilliges**. Dec. 2020. "CoSE: Compositional Stroke Embeddings". In: *Advances in Neural Information Processing Systems (NeurIPS)*.

[2] Emre Aksan and **Otmar Hilliges**. May 2019. "STCN: Stochastic Temporal Convolutional Networks". In: *International Conference on Learning Representations (ICLR)*.

[3] Emre Aksan, Manuel Kaufmann, and **Otmar Hilliges**. Oct. 2019. "Structured Prediction Helps 3D Human Motion Modelling". In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. doi Link.

[4] Emre Aksan, Fabrizio Pece, and **Otmar Hilliges**. Apr. 2018. "DeepWriting: Making Digital Ink Editable via Deep Generative Modeling". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI '18. doi Link.

[5] Xu Chen, Zijian Dong, Jie Song, Andreas Geiger, and **Otmar Hilliges**. Aug. 2020. "Category Level Object Pose Estimation via Neural Analysis-by-Synthesis". In: *Computer Vision – ECCV 2020*. doi Link.

[6] Anna Feit, Lukas Vordemann, Seonwook Park, Caterina Bérubé, and **Otmar Hilliges**. June 2020. "Detecting Relevance during Decision-Making from Eye Movements for UI Adaptation". In: *Symposium on Eye Tracking Research and Applications*. ETRA '20. doi Link.

[7] Christoph Gebhardt, Brian Hecox, Bas van Opheusden, Daniel Wigdor, James Hillis, **Otmar Hilliges**, and Hrvoje Benko. Oct. 2019. "Learning Cooperative Personalized Policies from Gaze Data". In: *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*. UIST '19. doi Link.

[8] Christoph Gebhardt, Antti Oulasvirta, and **Otmar Hilliges**. Nov. 5, 2020. "Hierarchical Reinforcement Learning Explains Task Interleaving Behavior". In: *Computational Brain & Behavior*. doi Link.

[9] Partha Ghosh, Jie Song, Emre Aksan, and **Otmar Hilliges**. Oct. 2017. "Learning human motion models for long-term predictions". In: *2017 International Conference on 3D Vision (3DV)*. IEEE. doi Link.

[10] Zhe He, Adrian Spurr, Xucong Zhang, and **Otmar Hilliges**. Oct. 2019. "Photo-Realistic Monocular Gaze Redirection Using Generative Adversarial Networks". In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE. doi Link.

[11] **Otmar Hilliges**, David Kim, Shahram Izadi, Malte Weiss, and Andrew Wilson. May 2012. "HoloDesk: Direct 3D Interactions with a Situated See-through Display". In: *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*. CHI '12. doi Link.

[12] Yinghao Huang, Manuel Kaufmann, Emre Aksan, Michael J. Black, **Otmar Hilliges**, and Gerard Pons-Moll. Jan. 2019. "Deep Inertial Poser: Learning to Reconstruct Human Pose from Sparse Inertial Measurements in Real Time". In: *ACM Transactions on Graphics, (Proceedings SIGGRAPH Asia)*. doi Link.

[13] Shahram Izadi, Richard A. Newcombe, David Kim, **Otmar Hilliges**, David Molyneaux, Steve Hodges, Pushmeet Kohli, Jamie Shotton, Andrew J. Davison, and Andrew Fitzgibbon. July 2011. "KinectFusion: Real-time Dynamic 3D Surface Reconstruction and Interaction". In: *ACM SIGGRAPH 2011 Talks*. SIGGRAPH '11. doi Link.

[14] Manuel Kaufmann, Emre Aksan, Jie Song, Fabrizio Pece, Remo Ziegler, and **Otmar Hilliges**. Nov. 2020. "Convolutional Autoencoders for Human Motion Infilling". In: *2020 International Conference on 3D Vision (3DV)*. 3DV '20. IEEE.

[15] David Kim, **Otmar Hilliges**, Shahram Izadi, Alex D. Butler, Jiawen Chen, Iason Oikonomidis, and Patrick Olivier. Oct. 2012. "Digits: Freehand 3D Interactions Anywhere Using a Wrist-worn Gloveless Sensor". In: *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology*. UIST '12. doi Link.

[16] Thomas Langerak, Juan Zarate, David Lindlbauer, Christian Holz, and **Otmar Hilliges**. Oct. 2020. "Omni: Volumetric Sensing and Actuation of Passive Magnetic Tools for Dynamic Haptic Feedback". In: *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. UIST '20. doi Link.

[17] Thomas Langerak, Juan Zarate, Velko Vechev, David Lindlbauer, Daniele Panozzo, and **Otmar Hilliges**. Oct. 2020. "Optimal Control for Electromagnetic Haptic Guidance Systems". In: *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. UIST '20. doi Link.

[18] David Lindlbauer, Anna Maria Feit, and **Otmar Hilliges**. Oct. 2019. "Context-Aware Online Adaptation of Mixed Reality Interfaces". In: *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*. UIST '19. doi Link.

[19] Richard A. Newcombe, Shahram Izadi, **Otmar Hilliges**, David Molyneaux, David Kim, Andrew J. Davison, Pushmeet Kohli, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. Oct. 2011. "KinectFusion: Real-time Dense Surface Mapping and Tracking". In: *Proceedings of the 2011 10th IEEE International Symposium on Mixed and Augmented Reality*. ISMAR '11. doi Link.

[20] Seonwook Park, Emre Aksan, Xucong Zhang, and **Otmar Hilliges**. Aug. 2020. "Towards End-to-End Video-Based Eye-Tracking". In: *Computer Vision – ECCV 2020*. doi Link.

[21] Seonwook Park, Christoph Gebhardt, Roman Rädle, Anna Feit, Hana Vrzakova, Niraj Dayama, Hui-Shyong Yeo, Clemens Klokmose, Aaron Quigley, Antti Oulasvirta, and **Otmar Hilliges**. Apr. 2018. "AdaM: Adapting Multi-User Interfaces for Collaborative Environments in Real-Time". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI '18. doi Link.

[22] Seonwook Park, Shalini De Mello, Pavlo Molchanov, Umar Iqbal, **Otmar Hilliges**, and Jan Kautz. Oct. 2019. "Few-Shot Adaptive Gaze Estimation". In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. doi Link.

[23] Seonwook Park, Adrian Spurr, and **Otmar Hilliges**. Oct. 2018. "Deep Pictorial Gaze Estimation". In: *European Conference on Computer Vision (ECCV)*. ECCV '18. doi Link.

[24] Seonwook Park, Xucong Zhang, Andreas Bulling, and **Otmar Hilliges**. June 2018. "Learning to Find Eye Region Landmarks for Remote Gaze Estimation in Unconstrained Settings". In: *ACM Symposium on Eye Tracking Research and Applications (ETRA)*. ETRA '18. doi Link.

[25] Fabrizio Pece, Juan Jose Zarate, Velko Vechev, Nadine Besse, Olexandr Gudozhnik, Herbert Shea, and **Otmar Hilliges**. Oct. 2017. "MagTics: Flexible and Thin Form Factor Magnetic Actuators for Dynamic and Wearable Haptic Feedback". In: *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*. UIST '17. doi Link.

[26] Jie Song, Bjoern Andres, Michael Black, **Otmar Hilliges**, and Siyu Tang. Oct. 2019. "End-to-End Learning for Graph Decomposition". In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. doi Link.

[27] Jie Song, Xu Chen, and **Otmar Hilliges**. Aug. 2020. "Human Body Model Fitting by Learned Gradient Descent". In: *Computer Vision – ECCV 2020*. doi Link.

[28] Jie Song, Gábor Sörös, Fabrizio Pece, Sean Ryan Fanello, Shahram Izadi, Cem Keskin, and **Otmar Hilliges**. Oct. 2014. "In-air Gestures Around Unmodified Mobile Devices". In: *Proceedings of the 27th annual ACM symposium on User interface software and technology*. UIST '14. doi Link.

[29] Jie Song, Limin Wang, Luc Van Gool, and **Otmar Hilliges**. July 2017. "Thin-Slicing Network: A Deep Structured Model for Pose Estimation in Videos". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. doi Link.

[30] Adrian Spurr, Umar Iqbal, Pavlo Molchanov, **Otmar Hilliges**, and Jan Kautz. Aug. 2020. "Weakly Supervised 3D Hand Pose Estimation via Biomechanical Constraints". In: *Computer Vision – ECCV 2020*. doi Link.

[31] Adrian Spurr, Jie Song, Seonwook Park, and **Otmar Hilliges**. June 2018. "Cross-Modal Deep Variational Hand Pose Estimation". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. doi Link.

[32] Velko Vechev, Juan Zarate, David Lindlbauer, Ronan Hinchet, Herbert Shea, and **Otmar Hilliges**. Mar. 2019. "TacTiles: Dual-mode Low-power Electromagnetic Actuators for Rendering Continuous Contact and Spatial Haptic Patterns in VR". In: *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE. doi Link.

[33] Saiwen Wang, Jie Song, Jaime Lien, Ivan Poupyrev, and **Otmar Hilliges**. Oct. 2016. "Interacting with soli: Exploring fine-grained dynamic gesture recognition in the radio-frequency spectrum". In: *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. doi Link.

[34] Xucong Zhang, Seonwook Park, Thabo Beeler, Derek Bradley, Siyu Tang, and **Otmar Hilliges**. Aug. 2020. "ETH-XGaze: A Large Scale Dataset for Gaze Estimation Under Extreme Head Pose and Gaze Variation". In: *Computer Vision – ECCV 2020*. doi Link.

[35] Yufeng Zheng, Seonwook Park, Xucong Zhang, Shalini De Mello, and **Otmar Hilliges**. Dec. 2020. "Self-Learning Transformations for Improving Gaze and Head Redirection". In: *Neural Information Processing Systems (NeurIPS)*.