Face, Hair, and Body Modeling in the Wild

Research Statement

Ira Kemelmacher-Shlizerman

The future is here. It's just not widely distributed yet.
--William Gibson

Ever since William Gibson's "Neuromancer" in 1984, followed by Neal Stephenson's "Snow Crash" and more recently, Ernest Cline's "Ready Player One" people have been coming up with mind blowing virtual and augmented reality *experiences* that could revolutionize the way we learn new capabilities, communicate across continents, and visualize our world, to name a few.

Today in 2018, such experiences are still mostly limited to games, and the majority of the population is not leveraging those. One reason is that we haven't built the killer computer vision, graphics, AR/VR experience and device yet. Indeed, there is a lot of investment in industry to create the perfect AR device, but not as much research and investment goes into developing applications that can be used day to day by diverse populations.

This is where I find inspiration for my research—to invent experiences that may appeal to wide and diverse populations in their day to day activities. Example areas include: teaching people how to play a musical instrument with the help of computer vision/graphics/learning/AR technology, how to cook a new dish, virtual try-on of different appearances, watching a soccer game as a 3D hologram, or telepresence to enable better and faster visual communication between people. Fig. 1 shows examples of experiences we developed and are working on.

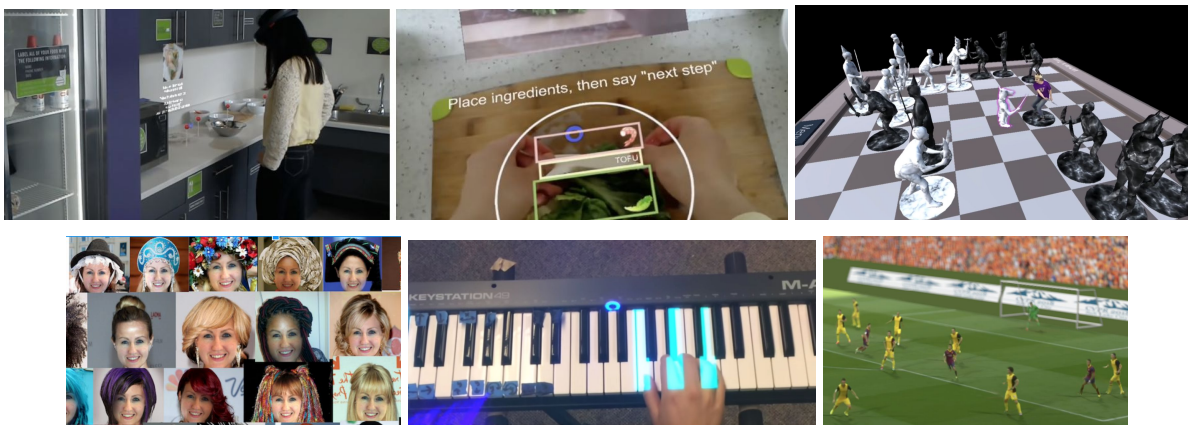

Figure 1. Top: AR cooking app we developed: the kitchen is mapped, and the person in HoloLens is guided throughout recipe steps in their own speed and via visual guides. The guides are overlaid on top of physical objects in the kitchen. Top right: Holochess, as viewed through HoloLens, the little queen figure is me scanned into the game. Bottom: virtual try-on of hairstyles and hats, keyboard app we developed, and 3d hologram of a soccer game. More experiences in: https://courses.cs.washington.edu/courses/cse481v/16sp/home.php and https://homes.cs.washington.edu/~kemelmi/

These and many other applications require major research advances in fundamental areas. One of those is modeling the human face, body, hair, and clothing. Inventing and developing those is a major focus of my research. Below I describe the key research contributions I've made in recent years.

Research Contributions

1. Face Modeling in the Wild

Most of our time is spent among people, and the majority of the photos people take include people. Face modeling is a foundational research problem in computer vision and graphics. Traditionally, most research in this area involves reconstruction from calibrated data via construction of complex setups as light stages, camera rigs, and other types of scanners.

My work in the last decade pioneered face modeling from data in the wild—completely uncalibrated photos as found on the Internet and photos we take with our phones in our day to day activities. Enabling modeling from such collections is important since it unlocks massive new datasets, previously unavailable for 3D analysis. Even more importantly, it opens up the possibility of doing population scale analysis, as only a tiny part of the world's population can afford coming to a lab to be scanned.

I started by developing an algorithm that can reconstruct a **3D shape of a face from a single photo** taken with completely unknown conditions. I observed that shadows on a face reveal a lot about the underlying shape even though the problem is ill-posed, and developed a shape from shading algorithm that works on any photo. Previously shape from shading was only possible in synthetic conditions. [24-28] describe the theory that enabled my approach, it involved modeling of light transport, photometric stereo, and face shape reconstruction. This was the first work that enabled **detailed** face reconstruction from a single photo taken with any light, expressions, and pose (Fig.2). The only other work was based on morphable models and did not allow detailed reconstruction. Current state of the art results typically combine the two approaches: our shading based approach with morphable models.

Figure 2. Example result of my shape from shading algorithm [24-28].

The ability to model a face from any photo automatically, allowed me to **explore big photo collections of people**. I realized that fascinating experiences can be enabled on ordinary photo collections by automatically analyzing thousands of photos. This is how “Face Movies” (https://www.youtube.com/watch?v=fLQtssJDMMc) was born; we showed that by aligning photos to fix the eyes to the same position across the collection, creating a graph where each photo is a node and traversing this graph according to similarity in facial expression, pose, age, and hair, we can visualize 20 years of photos in a beautiful movie that shows how a person changes over time. A key technical contribution was to prove that cross-dissolve (linear interpolation of images over time) creates motion perception--i.e., a step edge will translate under cross-dissolve, enabling smooth yet dynamic transitions across photos.

This experience was tech transferred to Google Inc. and shipped in their product Picasa, appeared on covers of ACM and SIGGRAPH, and published in [22,14]. The underlying algorithmic pieces developed for face movies were useful for a range of applications, e.g., to puppeteer and animate faces by analyzing video frames in [23] with a recent interesting application that appeared in Google’s Art TED talk https://www.youtube.com/watch?v=CjB6DQGaIU0&t=662s as a new way to explore classical paintings (traveling through portraits).

Collection flow [20] was a breakthrough for us and enabled our later series of contributions where detailed and accurate 3D shape [21], 3D flow [15], and texture [11] reconstruction was possible from any video of a person talking (Fig. 3). The motivation was to overcome “brightness constancy”-- a fundamental assumption of any optical flow algorithm which limits the ability to **compute flow under variable illumination**. We proposed that instead of comparing pixels directly, we can use large collections of photos from the class and compare low dimensional illumination sub-spaces. This insight provided a powerful way to modify any flow algorithm to achieve illumination-invariance, by operating on collections rather than pairs of photos. Fig.3 illustrates this idea.

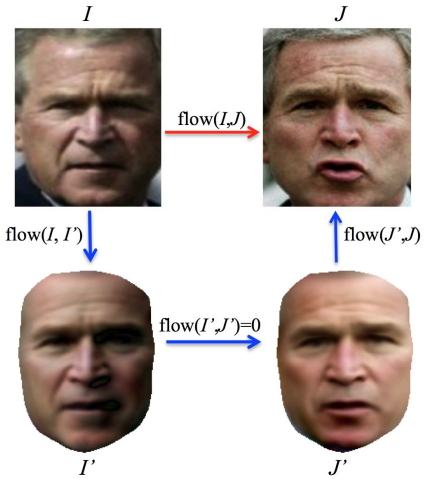

Figure 3. Key idea of collection flow algorithm: instead of comparing pixels directly (flow from I to J), we can use large collections of photos from the class and compare low dimensional illumination sub-spaces (flow from I to J is actually flow composed from I to I’ (which satisfy brightness constancy), I’ to J’, and J’ to J, where flow from I’ to J’ in this case is zero since it’s the same person sub-space).

Our first application of collection flow was to build an “age progression” algorithm for predicting **how children will look when they are older** [16], a technology critical for missing children search. Unlike previous work, our idea was to leverage the Internet for creating low dimensional subspaces of men and women at different ages and learn from those how to synthesize older ages from any given photo of a young person. The results are highly realistic; in our user study, participants were, on average, unable to distinguish our predictions from ground truth photos (Fig.4.) A lot of follow-up work was done in that space and our algorithm is still best performing for young children.

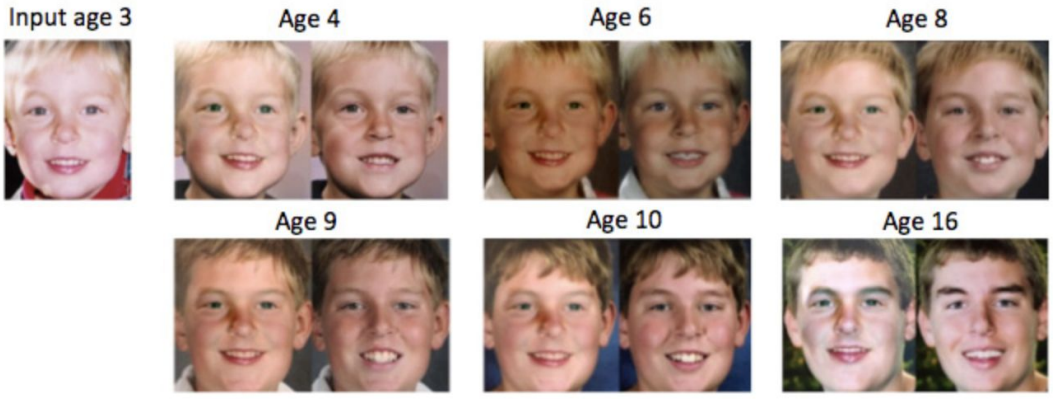

Figure 4. Example result of our age progression algorithm for young children [16]. For each age the photo on the left is synthesized by our method, the photo on the right is ground truth. Note the similarity in each pair. Currently our algorithm is still best performing for young children.

The accuracy of our face reconstruction algorithms [21,15,11] from challenging data found on the Internet, compared to other methods, is additionally derived from several key technical insights: we handled the natural non-rigidity of faces in photos using local view refinement fused into photometric stereo, and leveraged Internet photos together with single-view shading to get highly detailed 3D face reconstructions from monocular YouTube videos. Fig.5 shows example reconstructions of our methods.

Figure 5. Example result of our shading combined with 3D flow algorithm [21,15,20]. Note the highly detailed shape and how it corresponds to the input frames. See videos at: https://www.youtube.com/watch?v=C1iLVAUiC7s and https://www.youtube.com/watch?v=ladqJQLR2bA

Finally, realizing that existing face recognition datasets were too limited, and leveraging the aforementioned technology, I built a **million scale face recognition benchmark called MegaFace** http://megaface.cs.washington.edu/. This is a big effort in data collection, labeling and testing algorithms, and has become the standard benchmark for face recognition for the last five years that is used by thousands of research and industry groups world-wide [10,7]. MegaFace enabled leveling the playing field between academia and industry (previously only companies like Facebook and Google had access to large face datasets with labeled ground truth).

2. Audio - Visual Correlation

A fascinating question that has recently become of high interest to the computer vision and graphics communities is how audio correlates with visual information. I have worked on two aspects of this question: 1) lip sync: can we predict how lips move just from speech signal? [6], 2) can we predict how human body and fingers move while playing an instrument? [4]

In [6] we set an ambitious goal: from only audio of a person speaking, output **photorealistic video of the person saying what’s in the audio**. The key new technical innovation to achieve that goal is training a neural net to predict

lip shape at every point in time from an audio signal, and then combine it with our state of the art face analysis framework (described above), to achieve photorealistic video output. The neural net was trained on 15 hours of video footage of Barack Obama speaking. Our results were striking in their quality, it was hard to distinguish from ground truth video. This created a sensation in the media and has been covered by CNN, BBC, New York Times, and many others. Fig. 6 shows an example synthesized result.

Figure 6. Example synthesis of our algorithm just from audio signal. See video at: https://www.youtube.com/watch?time_continue=2&v=9Yq67CjDqvw

One of the exciting applications of AR could be help people learn new capabilities faster. For example, learn how to play musical instruments. In Fig.1 I show an AR interface that guides the person through piano keys. However, more is required to actually teach complicated pieces. To explore the complexity of music and the corresponding fingers configurations, in [4] we demonstrated a system that learns how a pianist's fingers and body move to create music by training on hours of recitals (Fig.7). The result is given a new musical piece, an avatar plays the musical piece on a virtual piano. While the result is yet perfect, it shows that correlation and predictability are possible. It makes me believe that augmented reality and computer vision has a place in visualizing and teaching people how to play musical instruments.

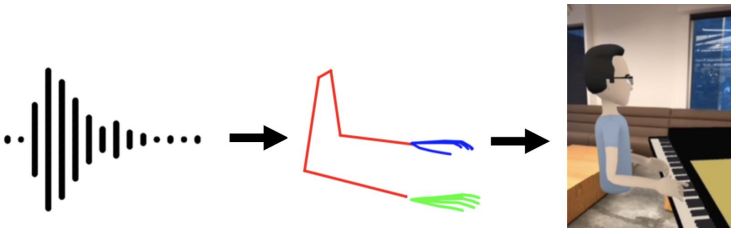

Figure 7. Given music we estimate the configuration of arms and fingers to produce that music and map it onto an avatar in AR.https://arviolin.github.io/AudioBodyDynamics/

3. Body and Hair Modeling

Similarly to face modeling (where we achieved state of the art results), I am working on hair and body modeling from completely uncalibrated data.

We have been working on enabling **automatic head and hair modeling from any video**, e.g., Internet video or selfie video taken with a phone (Fig.8). We built the first system that achieves that goal [2,8]. Previously, state of the art systems assumed significant manual user input or synchronized calibrated cameras. The key contribution of this work, in addition to automatic neural net based segmentation and head analysis algorithms, is a 3D deformation algorithm that modifies a rough 3D head and hair template to fit automatically derived silhouettes in video frames. We demonstrated an end to end system that can take in a selfie video of person's head and output a digital strand based hair model combined with the head and face.

Figure 8. Examples from [2]. Given any video of a person moving naturally (or selfie video) the algorithm reconstructs strand based model of the hair, as well as face, completely automatically. This is the first fully automatic system for hair modeling.

I have also introduced a 2D version of automatic hair modeling from any photo for a **hair and head outfit try-on** application [9] (single author SIGGRAPH paper). The idea was the following: let people imagine how they may look like in different appearances just by uploading their photo and typing a search query, e.g., “curly hair”. The output is various versions of the input person rendered with curly hair. The system works as a search engine, where any phrase can be typed, e.g., “1930” and then the person sees themselves in 1930 style or “India” and it renders the person in indian outfits (Fig.9). A key contribution was to define a concept of similarities in terms of facial shape, pose, lighting, and expression, and use it for image blending. Imagine taking two random photos and trying to blend the face from one with hair from the other; the result likely won’t look good. However, if the people in the two photos have very similar face shapes and very similar poses, then the blend looks much better. The system leveraged millions of photos on the Internet to search for text based queries and match only ones where people are similar. This was the basis for my startup Dreambit acquired by Facebook Inc.

Figure 9. Examples from [9]. The algorithm lets you imagine how you’d look like in different hair styles and head outfits just by entering an image search phrase like “curly hair”. All those photos are synthetic and were created from the same single input via various queries. https://www.youtube.com/watch?v=mILLFK1Rwhk

For body modeling, I have been exploring a number of directions in sports, music and 3D photo visualization, as follows.

In music, we published a first study [4] on how to find **correlation between classical music and body** movements, and given new music to be able to predict how fingers of a pianist would move.

In sports, I am working on **3D hologram visualization of a soccer game**. We published our first work [5] in this space where the input is a single YouTube video of a soccer game and the output is a 3D visualization of the soccer field and players, which can be watched in an AR device (we demonstrate how it works with HoloLens, Fig.10). The key technical contribution was to train a domain specific network to predict meshes of soccer players. We demonstrate state of the art body shape reconstruction results, as well as the first end to end virtual soccer system that was reconstructed from a single youtube video and visualized in AR.

Figure 10. Holographic soccer [5] as viewed from inside the HoloLens. This 3D reconstruction was produced from a single YouTube video of the soccer game. http://grail.cs.washington.edu/projects/soccer/

Finally in [3], from a single photograph, we reconstruct the human figure, and **animate the photo to show in AR**. It creates a magical effect where the human figure in the photo comes to life (Fig.11). This requires segmentation, body estimation, and 3D warping algorithms to robustly work for any given input. A key contribution is a warping algorithm that warps human body template to image silhouettes to include clothing, and various body shapes. We demonstrate diverse examples on sports photos, movie posters, and even Picasso art.

Figure 11. Examples of [3], where a single photo or painting comes to life and the human in the photo/art is animated in AR. (Best visualized using a Hololens.)

4. Novel Experiences

Beyond the algorithms themselves, my goal is to create new experiences and use cases that change how people work and live. Among the experiences we have developed (in addition to what I described above) are how to visualize and explore large photo collections, how to teach people to play simple songs on a keyboard with AR, how to cook with HoloLens as an alternative to looking up recipes on a mobile phone, and how to play chess in AR (Fig.1).

Going forward, I will further work on developing human modeling algorithms while focusing on sports, music, telepresence, and other AR experiences. We've shown very promising results, but just scratched the surface of what's possible in the space. I truly believe that creating technology and building experiences for diverse populations will give us superpowers, and that's what I enjoy doing in concert with advancing computer vision and graphics algorithms.

Education and service:

UW Reality Lab: I co-founded and am co-director of the UW Reality Lab with Steve Seitz and Brian Curless. This center focuses on advancing research and education in VR/AR on all levels--from an undergrad program to graduate research across a variety of fields in Computer Science that contribute to the advancement of VR/AR.

Teaching: I am a strong believer in AR/VR and I am planning our education program (as part of our new UW Reality Lab).

One specific initiative is to create an **undergrad ideas incubator** where prototype experiences and research is done. We are bootstrapping this program right now during the summer with high interest from the undergrad community (120 students applied to our initial program wanting to do AR/VR research).

Identifying which research fits possible future products and startups is also very close to my heart and will be part of this program (my own experience with Google and Facebook worked out well).

I have also **developed and taught a new class** on AR/VR; developing the curriculum around those topic is highly exciting. This was the world's first AR class with HoloLens, where students got access to 40 HoloLenses (before their release) and had 10 weeks to build an application. The class was a big success where the developed applications were even uploaded to the Microsoft store, and included an amazing demo session with high interest from most major AR/VR companies.

Service: I am an active member of both the computer vision and computer graphics community where I was the program coordination chair at CVPR 2016, and have been an area chair for CVPR and on the technical committee of SIGGRAPH for several years.

I've graduated two PhD students: Supasorn Suwajanakorn (joined Google Brain) and Shu Liang (joined Facebook), 2 master students (joined Facebook and a startup), and worked with 10 undergrads.

Papers In Refereed Journals:

[1] Eli Shlizerman, Andrew Denyes, Francis Ge, Ira Kemelmacher-Shlizerman, Video Flights, submitted to SIGGRAPH ASIA, 2018

[2] Shu Liang, Xiufeng Huang, Xianyu Meng, Kunyao Chen, Linda Shapiro, Ira Kemelmacher-Shlizerman, Video to Fully Automatic 3D Head Model, submitted to SIGGRAPH ASIA, 2018

[3] Chung-Yi Weng, Brian Curless, Ira Kemelmacher-Shlizerman, Photo Wake-Up: 3D Character Animation from a Single Photo, submitted to SIGGRAPH ASIA, 2018

[4] Eli Shlizerman, Lucio Dery, Hayden Shochat, Ira Kemelmacher-Shlizerman, “Audio to Body Dynamics”, CVPR, 2018

[5] Konstantinos Rematas, Ira Kemelmacher-Shlizerman, Brian Curless, Steve Seitz, “Soccer on Your Tabletop”, CVPR, 2018

[6] Supasorn Suwajanakorn, Steve Seitz, Ira Kemelmacher-Shlizerman, “Synthesizing Obama: Learning Lip Sync from Audio”, ACM Transactions on Graphics (SIGGRAPH), 2017

[7] Aaron Nech and Ira Kemelmacher-Shlizerman, Level Playing Field for Million-Scale Face Recognition, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017.

[8] Shu Liang, Linda Shapiro, Ira Kemelmacher-Shlizerman, Head Reconstruction from Internet Photos, European Conference on Computer Vision (ECCV) 2016

[9] I. Kemelmacher-Shlizerman, Transfiguring Portraits, ACM Transactions on Graphics (SIGGRAPH) 2016

[10] I. Kemelmacher-Shlizerman, E. Brossard, S. Seitz, D. Miller, The MegaFace Benchmark: Million Faces for Large Scale Recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016

[11] S. Suwajanakorn, S. Seitz, I. Kemelmacher-Shlizerman, What makes Tom Hanks look like Tom Hanks, International Conference on Computer Vision (ICCV), Chile, 2015

[12] K. Tuite and I. Kemelmacher-Shlizerman, The Meme Quiz: A Facial Expression Game Combining Human Agency and Machine Involvement, Foundations of Digital Games (FDG), 2015

[13] S. Liang, I. Kemelmacher-Shlizerman, L. Shapiro, 3D Face Hallucination from a Single Depth Frame, International Conference on 3D Vision (3DV), Tokyo, Dec 2014

[14] I. Kemelmacher-Shlizerman, E. Shechtman, R. Garg, S. Seitz, Moving Portraits, Comm. of the ACM, Research Highlights, Sep 2014

[15] S. Suwajanakorn, I. Kemelmacher-Shlizerman, S. Seitz, Total Moving Face Reconstruction, European Conference on Computer Vision (ECCV), Zurich, Sep 2014.

[16] I. Kemelmacher-Shlizerman, S. Suwajanakorn, S. Seitz, Illumination-aware Age Progression, Proc. of Computer Vision Pattern Recognition (CVPR), June 2014.

[17] I. Kemelmacher-Shlizerman, Internet-based morphable model. International Conf. Computer Vision (ICCV), Sydney, Dec 2013

[18] I. Kemelmacher-Shlizerman, R. Basri, B. Nadler, 3D Face Reconstruction from Single Two-tone and Color images, Shape Perception in Human and Computer Vision, Springer, 2013

[19] M. Arie-Nachimson, S. Kovalsky, I.Kemelmacher-Shlizerman, A.Singer, and R. Basri, Global Motion Estimation from Point Matches, International Conf. on 3D Vision (3DV), 2012.

[20] I. Kemelmacher-Shlizerman and S.M. Seitz, Collection Flow, Proc. of Computer Vision Pattern Recognition (CVPR), 2012

[21] I. Kemelmacher-Shlizerman, S.M. Seitz, Face Reconstruction in the Wild, International Conference on Computer Vision (ICCV), 2011

[22] I. Kemelmacher-Shlizerman, E. Shechtman, R. Garg, S.M.Seitz, Exploring Photobios, ACM Transactions on Graphics (SIGGRAPH) 2011.

[23] I. Kemelmacher-Shlizerman, A. Sankar, E. Shechtman, S.M. Seitz, Being John Malkovich, European Conference on Computer Vision (ECCV), 2010

[24] I. Kemelmacher-Shlizerman, R. Basri, 3D Face Reconstruction from a single image using a single reference face shape, IEEE Trans. on Pattern Analysis and Machine Int. (PAMI), 2010

[25] I. Kemelmacher-Shlizerman, R. Basri, B. Nadler, 3D Shape reconstruction of Mooney Faces, IEEE Conf. on Computer Vision and Pattern Recog. (CVPR) 2008

[26] D. Mahajan, I. Kemelmacher-Shlizerman, R. Ramamoorthi, P.N. Belhumeur, A Theory of Locally Low dimensional Light Transport, ACM Trans. on Graphics (SIGGRAPH), 2007

[27] R. Basri, D.W.Jacobs, I.Kemelmacher, Photometric Stereo with General Unknown lighting, International Journal of Computer Vision (IJCV), 2007

[28] I. Kemelmacher-Shlizerman, R. Basri, Molding Face Shapes by Example, European Conf. Computer Vision (ECCV), 2006

[29] I. Kemelmacher-Shlizerman, R. Basri, Indexing with Unknown Illumination and Pose, IEEE Conf. on Computer Vision and Pattern Recog. (CVPR) 2005.