Video to Fully Automatic 3D Hair Model

SHU LIANG, University of Washington
XIUFENG HUANG, Owlii
XIANYU MENG, Owlii
KUNYAO CHEN, Owlii
LINDA G. SHAPIRO, University of Washington
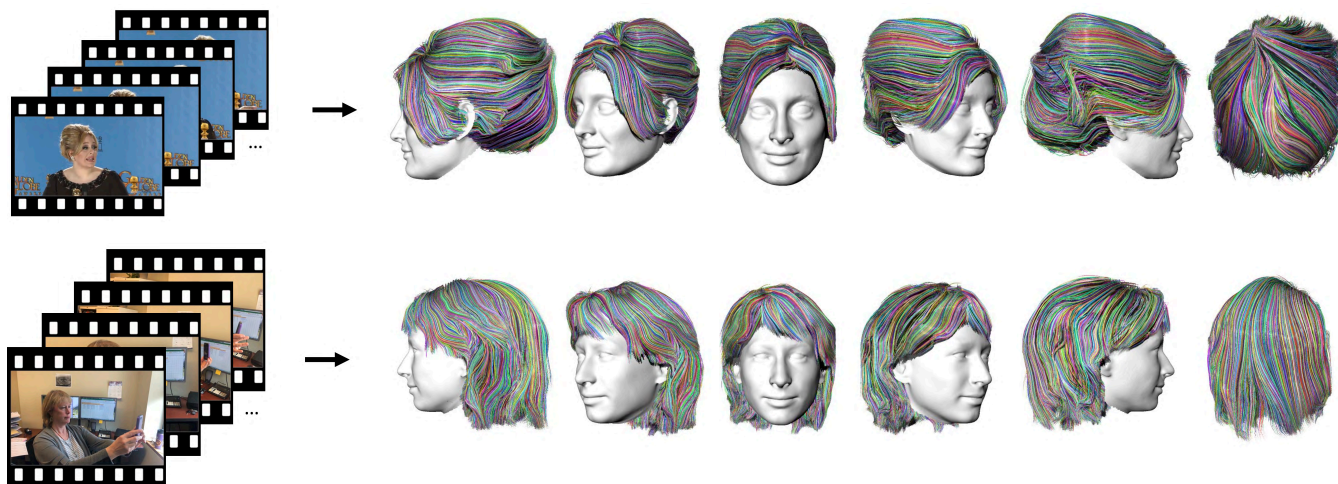IRA KEMELMACHER-SHLIZERMAN, University of Washington

Fig. 1. Two example video inputs (left) and automatic reconstructions produced by our method (right). Input to the algorithm is any video that captures a person's head, e.g., in the figure we show a celebrity video and a selfie video captured by a mobile phone (figure shows observer viewpoint).

Imagine taking a selfie video with your mobile phone and getting as output a 3D model of your head (face and 3D hair strands) that can be later used in VR, AR, and any other domain. State of the art hair reconstruction methods allow either a single photo (thus compromising 3D quality) or multiple views, but they require manual user interaction (manual hair segmentation and capture of fixed camera views that span full 360°). In this paper, we describe a system that can completely automatically create a reconstruction from any video (even a selfie video), and we don't require specific views, since taking your −90°, 90°, and full back views is not feasible in a selfie capture.

In the core of our system, in addition to the automatization components, hair strands are estimated and deformed in 3D (rather than 2D as in state of the art) thus enabling superior results. We provide qualitative, quantitative,

and Mechanical Turk human studies that support the proposed system, and show results on a diverse variety of videos (8 different celebrity videos, 9 selfie mobile videos, spanning age, gender, hair length, type, and styling).

CCS Concepts: • **Computing methodologies → Shape modeling**;

Additional Key Words and Phrases: Hair Reconstruction, Selfie video, 3D hair, in the wild

1 INTRODUCTION

> Her glossy hair is so well resolved that Hiro can see individual strands refracting the light into tiny rainbows.
>
> *Snow Crash, Neal Stephenson*

Any future virtual and augmented reality application that includes people must have a robust and fast system to capture a person's head (3D hair strands and face). The simplest capture scenario is taking a single selfie photo with a cell phone [Hu et al. 2017]. 3D reconstruction from a single selfie [Chai et al. 2016], however, by definition, will not produce high fidelity results due to the ill-posedness of the problem—a single view does not show sides of
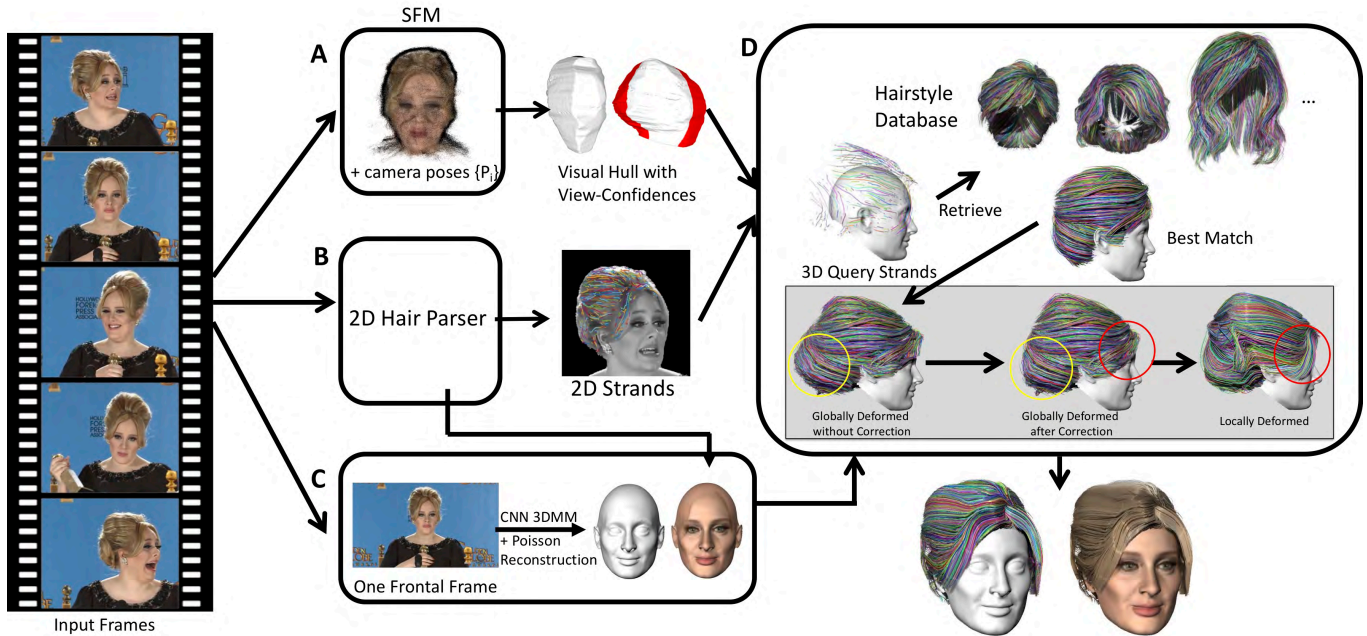
Fig. 2. Overview of our method. The input to the algorithm is a video: (A) structure from motion is applied to the video to get camera poses, depth maps and a visual hull shape with view-confidence values, (B) hair segmentation and gradient direction networks are trained to apply on each frame and recover 2D strands, (C) the segmentations are used to recover the texture of the face area, and a 3D face morphable model is used to estimate face and bald head shapes. The core of the algorithm is (D) where the depth maps and 2D strands are used to create 3D strands, which are used to query a hair database; the strands of the best match are refined globally and locally to fit the input photos.

the person. Using multiple frames, however, will create an accurate reconstruction. Indeed, a state-of-the-art method for hair modeling [Zhang et al. 2017] needs four views, but it requires spanning the full 360° (front, back, and sides), as well as user interaction. This paper proposes to use a video as input and introduces solutions to three obstacles that prevent state-of-the-art work [Chai et al. 2016; Hu et al. 2017; Zhang et al. 2017] from being applicable in simple automatic self capture:

(1) Fixed views: [Zhang et al. 2017] requires four views (front, back, and two side views), those are hard to acquire accurately with self capture. In this paper, we do not constrain the views; instead we use any available video frames in which the subjects talk or capture themselves. For the input videos which do not have a full view range from −90 to 90 degrees, we will correct the incomplete views with a hair shape from a database. Camera poses are estimated automatically with structure from motion. Results with various view ranges are demonstrated (as low as 90 degrees range). [Chai et al. 2016; Hu et al. 2017] assume a single frontal image.

(2) Hair segmentation: [Zhang et al. 2017] relies on the user to label hair and face pixels. In this paper, we use automatic hair segmentation and don't require any user input. Using general video frames, rather than four fixed views, introduces motion blur, varying lighting, and resolution issues, all of which our system overcomes.

(3) Accuracy: our method compares and deforms hair strands in 3D rather than 2D, and the availability of the back view is not required as in [Zhang et al. 2017]. It achieves higher accuracy results as demonstrated with qualitative, quantitative, and human studies. Intersection of union rate of the hair region compared to ground truth photos is on average 80% for our method (compared to 60% by [Zhang et al. 2017]). Amazon Turk raters prefer our results 72.4% over the four-view method [Zhang et al. 2017] and 90.8% over the single-view method of [Hu et al. 2017].

In addition to [Zhang et al. 2017] (and a previously multi-view based method by [Vanakittistien et al. 2016] that also required user interaction), there is a large body of work for modeling hair from photos. Earlier works assumed laboratory calibrated photos, e.g., [Hu et al. 2014a]. More recently [Chai et al. 2016, 2013; Hu et al. 2015, 2017] showed how to reconstruct hair from a single photo. Interesting hair-related applications inspire further research, e.g., depth-based portraits [Chai et al. 2016], effective avatars for games [Hu et al. 2017], photo-based hair morphing [Weng et al. 2013], and hair-try-outs [Kemelmacher-Shlizerman 2016]. Enabling reconstruction "in the wild" is an open problem where Internet photos have been explored [Liang et al. 2016], as well as structure from motion [Ichim et al. 2015] on a mobile video input. Both methods output a rough structure of the hair and head, without hair strands. This paper proposes a system that can take in an in-the-wild video and automatically output a full head model with a 3D hair-strand model.

2 RELATED WORK

In this section, we describe work that focuses on face, head, and hair modeling.

Face modeling (no hair, or head) has progressed tremendously in the last decade. Beginning with high-detailed head geometry with a stereo capturing system [Alexander et al. 2013; Beeler et al. 2010; Debevec 2012], then RGB-D-based methods like dynamic fusion [Newcombe et al. 2015] and non-rigid reconstruction methods [Thies et al. 2015; Zollhöfer et al. 2014] allowed capture to be real-time and much easier with off-the-shelf devices. [Blanz and Vetter 1999] proposed a 3D morphable face model to represent any person's face shape using a linear combination of face bases, [Richardson et al. 2016, 2017; Tran et al. 2017] proposed CNN-based systems and [Kemelmacher-Shlizerman and Basri 2011; Kemelmacher-Shlizerman and Seitz 2011; Suwajanakorn et al. 2014, 2015] showed how to estimate highly detailed shapes from Internet photos and videos.

Head modeling: Towards creating a full head model from a single photo, [Chai et al. 2015] modeled the face with hair as a 2.5D portrait, using head shape priors and shading information. [Maninchedda et al. 2016] allowed multiple views and used volumetric shape priors to reconstruct the geometry of a human head starting from structure-from-motion dense stereo matching. [Cao et al. 2016] showed that a full 3D head with a rough but morphable hair model can be reconstructed from a set of captured images with hair depth estimated from each image and then fused together. Finally, [Liang et al. 2016] explored how to reconstruct a full head model of a person from Internet photos. These methods focused on rough head and hair modeling without reconstruction of hair strands.

Hair strand modeling is key to digital human capture. Capturing the nuances of person's hair shape is critical, since the hair style is a unique characteristic of a person and inaccuracies can change their appearance dramatically.

Multi-view camera rigs and a controlled capturing environment were able to acquire hair shape with high fidelity [Hu et al. 2014a; Luo et al. 2013; Paris et al. 2004, 2008; Ward et al. 2007], and more recently with RGB-D cameras [Hu et al. 2014b]. With a single RGB image, [Chai et al. 2013, 2012] showed that strands can be recovered from per pixel gradients, and with a use of a database of synthetic hairstyles, [Hu et al. 2015] created a natural-looking hair model. A key requirement was to have a user draw directional strokes on the image to initialize strand construction. A fully automatic approach was proposed recently by [Chai et al. 2016] and [Hu et al. 2017] with CNN-based methods for hair segmentation, direction classification and a larger database for retrieving the best match for a single-view input.

Since single-view modeling is ill-posed by definition, [Vanakittistien et al. 2016] used a hand-held cell phone camera to take photos from 8 views of the head to recover the hair strands, and [Zhang et al. 2017] proposed a method to reconstruct the hair from four-view images starting from a rough shape retrieved from a database and synthesized hair textures to provide hair-growing directions to create detailed strands. However, both methods need human interactions for hair segmentation and pose alignment. This paper solves those constraints and also demonstrates higher accuracy results compared to those methods.
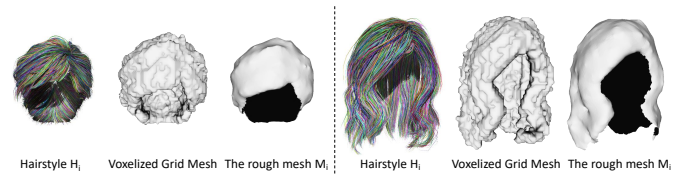

Fig. 3. Example hair styles from the dataset. For each hairstyle H_i, we create its corresponding rough mesh M_i as described in the text.

3 OVERVIEW

Figure 2 provides an overview of our method and the key components. The input to the algorithm is a video sequence of a person talking and moving naturally, as in a TV interview. There are four algorithmic components (correspond to the labeling of boxes in Figure 2):

(A) video frames are used to create a structure-from-motion model, estimate camera poses as well as per-frame depth, and compute a rough visual hull of the person with view-confidences, (B) two models are trained: one for hair segmentation, and another for hair direction; given those models, 2D hair strands are estimated and hair segmentation results are transferred to the visual hull to define the hair region, (C) the masks from the previous stage are used to separate the hair from the face and run the morphable model (3DMM) to estimate the face shape and later create the texture of the full head.

(D) is a key contribution in which first depth maps and 2D hair strands are combined to create 3D strands, and then 3D strands are used to query a database of hair styles. The match is deformed according to the visual hull, then corrected based on the region of confidence of the visual hull. Finally, it is deformed on the local strand level to fit the input strands. Texture is estimated from input frames to create the final hair model. The full head shape is a combination of the face model and the hair strands. In the next sections, we describe each of these components.

4 INPUT FRAMES TO ROUGH HEAD SHAPE

This section describes part (A) in Figure 2. We begin by preprocessing each frame i using semantic segmentation to roughly separate the person from the background [Zheng et al. 2015] resulting in masks S_i. Our goal is to estimate camera pose per frame and to create a rough initial structure from all the frames. Since the background is masked out, having the head moving while the camera is fixed is roughly equivalent to the head being fixed while the camera is moving; thus we use structure from motion [Wu 2011] to estimate camera pose P_i per frame and per-frame depth D_i using [Goesele et al. 2007].

Given S_i and P_i per frame, we estimate an initial visual hull of the head using shape-from-silhouette [Laurentini 1994]. The method takes a list of pairs P_i and S_i as input and carves a 3D voxel space to obtain a rough shape of the head. Meanwhile, each segmented video frame is processed using the IntraFace software [Xiong and De la Torre 2013], which provides 49 inner facial landmarks per frame. The 2D facial landmarks are transferred to 3D using D_i and averaged.

Hairstyle Before Global Deformation   Rough Mesh before Deformation   Rough Mesh after Deformation   Hairstyle After Global Deformation

a) a comparison of before and after global deformation

Reference Frame   Before Local Deformation   After Local Deformation

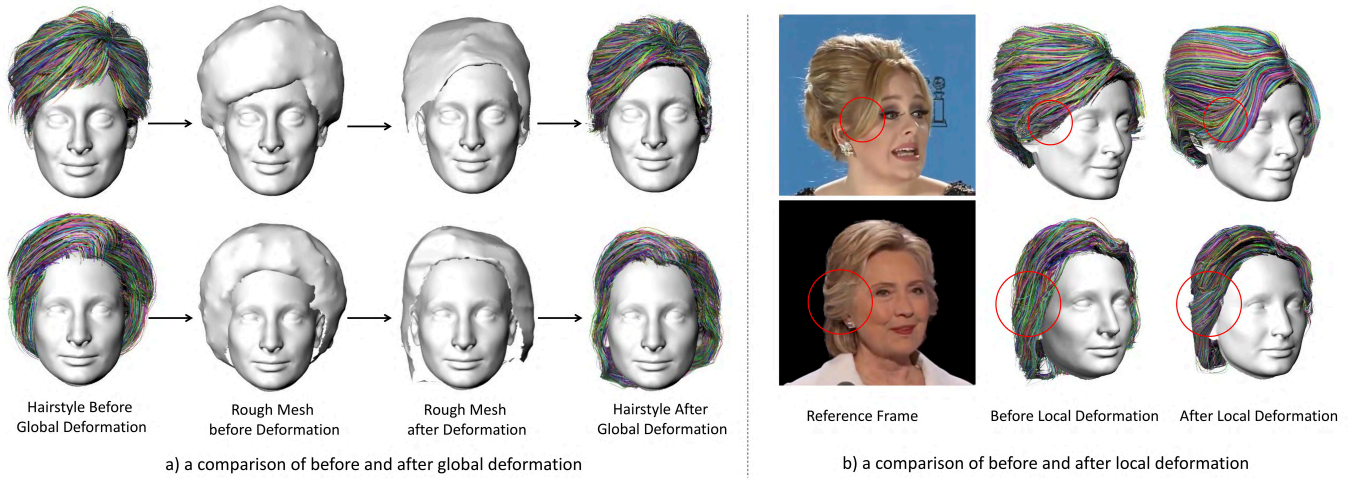b) a comparison of before and after local deformation

Fig. 4. In (a), we show a comparison of before and after global deformation. The retrieved hairstyle is deformed under the control of its rough mesh to fit the visual hull shape. In (b), we show a comparison of before and after local deformation. A video frame is shown as a reference that after local deformation, we are able to recover more personalized hair details.

The hair segmentation classifier trained in step(B) (Section 5) is run on all of our video frames. Each pixel is assigned a probability of being in the hair region. We drop the video frames with large motion blurs by calculating the surface area S_i of the detected hair region on each frame. Assuming the head size is relatively fixed across frames, a valid frame should have a hair region size of at least $0.33\overline{S_i}$. The corresponding probabilities of the valid frames are transferred to the visual-hull shape. A vertex with a mean probability larger than 0.5 is considered hair. Thus, we extract the hair part X_h out of the visual-hull as shown in Figure 6(a), and the remaining is the skin part.

The resultant visual-hull shape is relatively rough due to the non-rigid nature of the subject's head and might be stretched due to the incomplete views. Ideally, assuming the camera distance is always larger than the size of the head, we will get a complete visual hull if our video covers a full azimuth range of -90 to 90 degree. However, for in-the-wild videos, we usually cannot guarantee full coverage. We rigidly align the rough visual hull to a generic head model using 3D facial landmarks, and each camera pose P_i is also transformed to a corresponding P'_i based on this alignment. We connect each P'_i to the center of the generic head (the origin point in our case) and calculate the azimuth angle γ_i of each camera. The vertices on the visual hull with an azimuth angle in $[min(\gamma_i) - \pi/2, max(\gamma_i) - \pi/2] \cup [min(\gamma_i) + \pi/2, max(\gamma_i) + \pi/2]$ as illustrated in Figure 6(a) are denoted *high-confidence* vertices.

5 IMAGES TO 2D STRANDS

This section describes part (B) in Figure 2. Inspired by the strand direction estimation method of [Chai et al. 2016], we trained our own hair segmentation and hair directional classifiers to label and predict the hair direction in hair pixels of each video frame. We manually label hair directional labels on our data and train a classifier using the manual labels directly as groundtruth as in [Chai et al. 2016].

More details on our hair segmentation method can be found in the appendix. Results of the classifier are shown on examples in Figure 5 (1st and 3rd row).

To estimate 2D strands, we select one video frame every $\pi/8$ degrees according to its camera azimuth angle γ_i spanning the camera view range $[min(A_i), max(A_i)]$. Similar to previous hair orientation estimation methods [Chai et al. 2012; Jakob et al. 2009], we filter each image with a bank of oriented filters that are uniformly sampled in $[0, \pi)$. We choose the orientation θ_p with the maximum response for each pixel to get the 2D non-directional orientation map for each image. We further trace the hair strands on each non-directional orientation map following the method in [Chai et al. 2012] as shown in Figure 5 (2nd and 4th rows). The hair-direction labels are used to resolve the ambiguity of each traced 2D hair strand. If half of the points in a single strand have opposite directions to their directional labels, we flip the direction of the strand.

6 FACE MODEL

This section corresponds to part (C) in Figure 2. Each segmented video frame from the previous stage is processed using the IntraFace software [Xiong and De la Torre 2013], which provides head pose, and 49 inner facial landmarks. From all the frames, the frame that is closest to frontal face is picked first (where yaw and pitch are approximately 0), and fed to a morphable-model-based face model estimator [Tran et al. 2017]. This method generates a linear combination of the Basel 3D face dataset [Blanz and Vetter 1999] with both identity shape, expression weights and texture. Here, we only use the identity weights to generate a neutral face shape. In the future, it will be easy to add facial expressions to our method. The result of the estimation is a masked face model.

We complete the head shape using a generic 3D head model from the Facewarehouse dataset [Cao et al. 2013]. We choose to use the Basel dataset instead of using the Facewarehouse dataset to fit
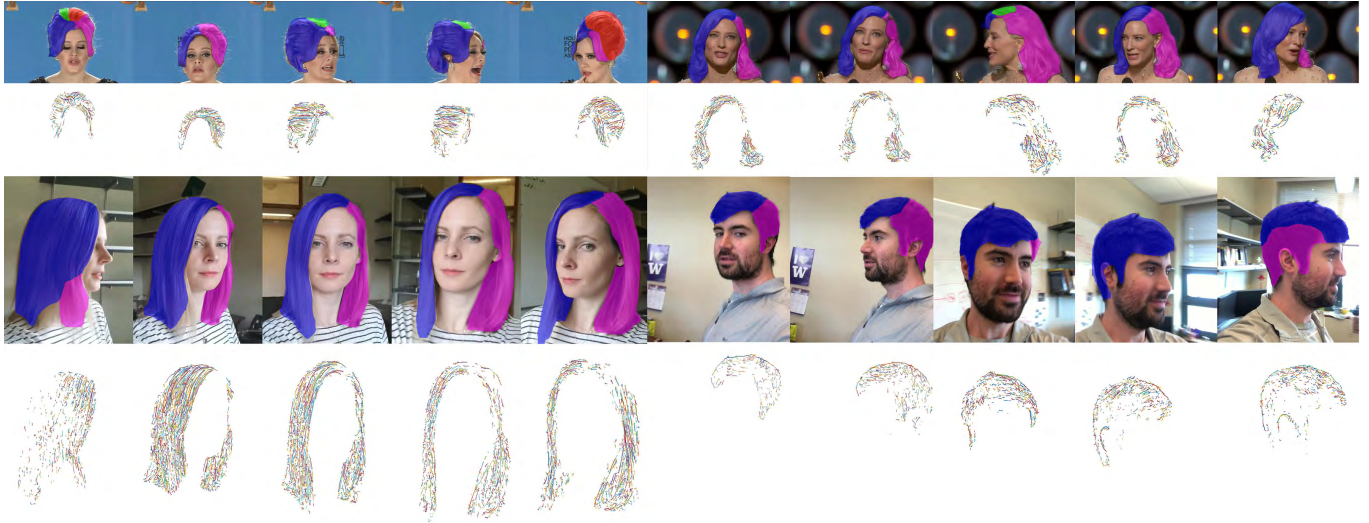
Fig. 5. Examples of Figure 2(B). Hair segmentation, directional labels and 2D hair strands of example video frames. For the color of the directional subregions, red stands for $[0, 0.5\pi)$, pink stands for $[0.5\pi, \pi)$, blue stands for $[\pi, 1.5\pi)$ and green stands for $[1.5\pi, 2\pi)$.

directly, because the Facewarehouse dataset contains only about $11k$ vertices for the whole head, while the Basel dataset contains about $53k$ for just the face region in which more facial shape details are provided. We pre-define 66 3D facial landmarks (49 landmarks on the inner face and 17 landmarks on the face contour and the ears) on both the 3D face dataset used by [Tran et al. 2017] and the generic head shape. Since all the face shapes in the 3D face dataset are in dense correspondence, we transfer these 66 landmarks to all the output 3D faces. We then deform the generic shape towards the 3D face shape using the landmarks, following [Liang et al. 2014]. We fuse the deformed generic head shape and the face shape using Poisson surface reconstruction [Kazhdan and Hoppe 2013] and get a complete head shape.

For the texture of the head, we project the full head to the selected frontal image and extract per-vertex colors from the non-hair region of the frontal image. We complete the side-view textures by projecting the head to all the frames. For the remaining invisible region, we assign an average skin color.

7 3D HAIR STRAND ESTIMATION

This section corresponds to part (D) in Figure 2. By utilizing a video, we deform the hairstyles in 3D instead of 2D because we are able to take advantage of the shape information of all the frames and its content continuity to estimate the per-frame pose.

2D to initial 3D strands: Each video frame i has an estimation of 2D strands; those are projected to depths D_i to estimate 3D strands. Large peaks of the 3D hair strands (distance to the neighboring vertex larger than 0.002 with a reference head width of 0.2) are removed. A merging procedure is performed to decrease future retrieval time (and reduce duplicate strands) as follows: for each pair of strands, if their directions are the same, the pairwise point-to-point distances of the vertices in these two 3D strands are checked, and the overlapping line segments are combined. If the directions

are not the same, no merging occurs. This process iterates until around 100 3D strands are obtained.

3D Strands to Query a Hair Database: The recovered 3D strands in the previous stage are sparse and incomplete; thus we use them to query a database of hair models and adjust the retrieved matches with global and local deformations to create a full hair model. The sparseness is a result of resolution of the video frames, motion blur, quality, and coverage in views (in all of our input videos, the back of the head is not visible). While being sparse, the strands do capture the person's specific hairstyle. We describe the algorithm below.

We use the hair dataset created by [Chai et al. 2016], which contains 35,000 different hairstyles, each hairstyle model consisting of more than 10,000 hair strands. For each database hair model, we create a voxel grid around each hair strand vertex and combine all voxel grids into a voxelized mesh. The shape is further smoothed using Laplacian mesh smoothing [Sorkine et al. 2004]. In order to remove the inner layer of the shape, a ray is shot from each vertex with the direction equivalent to the one from the center of the head to the current vertex. If the ray intersects any other part of the rough shape, the vertex is removed, because it is in the inner surface; otherwise it is kept. The resulting shape has 5,000 to 7,000 vertices. The final cleaned shape M (shown in Figure 3) will be used for retrieval and deformation.

For each 3D hair strand in our query hairstyle Q, the closest 3D hair strand from a hair style H is determined using the following distance:

$$Distance(Q, H) = \sum_{s_i \in Q} \sum_{p(s_i)} \min_{s_j \in H, \mathbf{n}_{p(s_i)} \cdot \mathbf{n}_{p(s_j)} > 0} |p(s_i) - p(s_j)|, \quad (1)$$

where s_i is a hair strand of Q and $p(s_i)$ is a vertex in strand s_i of H, $\mathbf{n}_{p(s_i)}$ is the tangent vector direction at $p(s_i)$.

This point-to-line distance comparison is very time-consuming. We performed experiments to accelerate the retrieval speed by pruning using a rough mesh X_h of the head obtained from step (A) (Section 4) and step (B) (Section 5) as follows:

(1) **Hairstyle boundary:** Only the hairstyles with x-range in the range of $(0.8(maxX\{X_h\} - minX\{X_h\}),$
$1.2(maxX\{X_h\} - minX\{X_h\}))$ and y-range in the range of $(0.8(maxY\{X_h\} - minY\{X_h\}), 1.2(maxY\{X_h\} - minY\{X_h\}))$ are considered.

(2) **Area of the hairstyle:** The surface area of the rough mesh X_h and of each hairstyle mesh M_i are computed. Only the hairstyles with surface area in the range of $(0.8S_{X_h}, 1.5S_{X_h})$ are considered.

Next, the retrieved matches are deformed in global and local fashion in 3D instead of 2D, taking advantage of the multi-view information in the video. Figure 4 illustrates the deformation process.

View Correction and Global Deformation After the top 20 best matching hairstyles are found, each retrieved hairstyle M_i is deformed towards the rough shape X_h (created by Step (A)(B) and shown in Figure 6(a)) using deformable registration [Allen et al. 2003] producing deformed hairstyle mesh M_i'. Furthermore, step (A) defines regions of *low-confidence* of X_h (see Section 4), so further correction is needed on the corresponding regions of M_i'. The azimuth angle of each vertex is calculated on M_i', and the vertices that are outside the confident region are considered invalid as shown in Figure 6(c) marked as red. Naturally, we think of using the original shape of M_i to correct the invalid region. The correction is based on the idea of Laplacian Mesh Editing [Sorkine et al. 2004]. We denote the valid view range as $[R_1, R_2] \cup [R_3, R_4]$ for simplification. We assign a confidence value c_i to each vertex v_i' on M_i' and minimize the following energy function.

$$E(v_i') = \sum_{i=1}^{n}(1 - c_i)||\mathcal{L}(v_i') - \mathcal{L}(v_i)||^2 + \lambda \sum_{c_i=1}||v_i' - x_i||^2 \quad (2)$$

where $\mathcal{L}$ is a Laplacian operator, v_i is the vertex position before deformation, x_i is the closest point of v_i' on X_h after direct deformable registration, λ is 10^{-5}. The confidence value c_i is 1 for the valid region and is defined for the invalid region as follows:

$$c_i = exp(-\frac{||\gamma(v_i') - R_j||^2}{2\sigma^2}) \begin{cases} j = 1, \gamma(v_i') \in (-\pi, R_1) \\ j = 2, \gamma(v_i') \in (R_2, 0) \\ j = 3, \gamma(v_i') \in [0, R_3) \\ j = 4, \gamma(v_i') \in (R_4, \pi] \end{cases}$$

where $\sigma = \pi/18$. As shown in Figure 6(c), the stretched red region of M_i' is corrected to have a natural look in (d). After the correction, a transformation matrix T is obtained for each vertex on M_i.

We further deform the hair strands in H_i as shown in Figure 4(a). Each vertex v_i in M_i works as an anchor point for the hairstyle deformation. For each point p in H_i, its deformation will be decided by a set of neighboring anchor points as

$$T_p = \frac{\alpha I + \sum_{v_i \in N(p)} w_i T_i}{\alpha + \sum_{v_i \in N(p)} w_i}, \quad (3)$$

where $N(p)$ is the set of neighboring anchor points chosen to be the top 10 closest vertices of M_i. I is an identity matrix, and w_i is defined as a Gaussian function

$$w_i = exp(-\frac{||p - v_i||^2}{2\sigma^2}) \quad (4)$$

In our experiments, we set α to 0.01 and σ to 0.015, while the width of our reference head is 0.2. We deform the top 20 best matching hairstyles, and use the same distance function as proposed in Equation 1 to find the final best match. We show a comparison in Figure 4 (a) before and after global deformation.

Local Deformation To add locally personalized hair details, we follow a method similar to [Fu et al. 2007] by converting the deformed hair strands into a 3D orientation field V. The orientation of the extracted hair strands from the video frames are also added to the 3D orientation field to bend the hair strands in the surface layer of the hairstyle. For each 3D query strand, we set an influence volume with a radius of 2 around it and diffuse it to fill in its surrounding voxels. The best matching hairstyle and the query 3D hair strands all contribute to the 3D orientation field by

$$E(\mathbf{v_i}) = \sum_{i \in V} ||\Delta(\mathbf{v_i})||^2 + w_1 \sum_{i \in C} ||\mathbf{v_i} - \mathbf{c_i}||^2 + w_2 \sum_{i \in Q} ||\mathbf{v_i} - \mathbf{q_i}||^2, \quad (5)$$

where Δ is the discrete Laplacian operator, C is the boundary constraints with the known directions $\mathbf{c_i}$ at certain voxel grids that contain 3D strands from the best-matching hairstyle, and Q is the boundary constraints from query hair strands.

In our experiment, we set the 3D orientation grid size to be $80 \times 80 \times 80$, w_1 to be 1 and w_2 to be 0.1. We show a comparison of results with and without the local deformation in Figure 4(b). Notice the personalized hair strands in the red circles. We avoid the artifacts of hair going inside the head by growing new hair strands out of V from the scalp region of the complete head shape of Section 6 with pre-defined hair root points.

Hair Texture The color of each hair strand vertex is averaged from all the frames, and the unseen regions are assigned by an average color of the visible regions.

8 EXPERIMENTS

In this section we describe the parameters used and the data collection process; we show results as well as comparisons to state-of-the-art methods.

8.1 Data Collection and Processing of Video Sequences

We collected 8 video clips of celebrities by searching for key words like "Hillary speech", "Adele award" on YouTube with an HD filter. The typical resolution of our videos is 720p (1280×720) with video duration around 40 seconds sampled at 10 fps (380 frames for Adele, 340 and 240 frames for Cate Blanchett, 250 and 350 frames for Hillary Clinton, 500 frames for Justin Trudeau, 390 frames for Theresa May, 310 frames for Angela Merkel). The camera view point is relatively fixed across all the frames, while the subject is making a speech with his/her head turning. We processed our frames at 10fps. We ran the face detection method of [Xiong and De la Torre 2013] on all the frames to determine a bounding box around the head (box height varies from 200 to 600). Our online video sequences typically
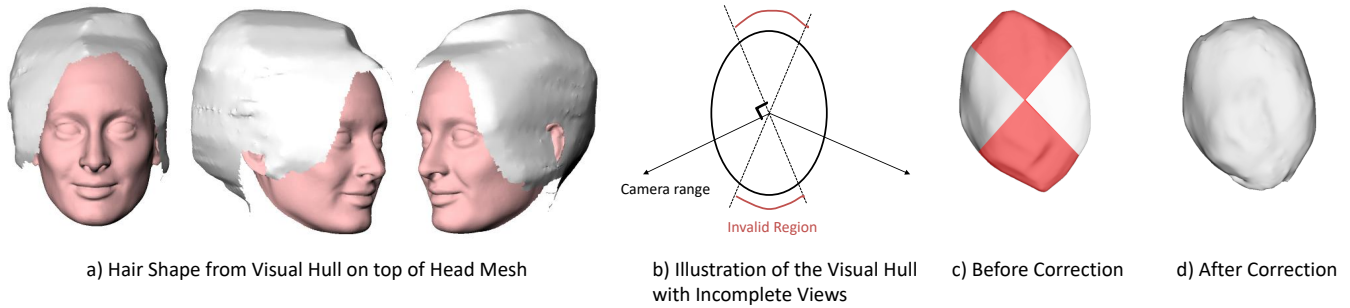
a) Hair Shape from Visual Hull on top of Head Mesh   b) Illustration of the Visual Hull with Incomplete Views   c) Before Correction   d) After Correction

Fig. 6. In (a), we show the hair part shape X_h extracted from the visual hull (Figure 2(A)) plus the hair labels in Figure 2(B) on top of the head mesh from Figure 2(C). (b) shows an illustration of the top-down view of the visual hull with camera range and invalid regions. In (c)(d), we show a candidate hairstyle mesh M_i before and after correction.

cover the frontal, left and right view of the person. The minimum view range we have is for Angela Merkel: only −15 to 75 degrees. There are no back views of any person's head in the videos.

For mobile selfie videos, 9 subjects were asked to take a selfie video of themselves from left to right and switch hands in the front using their own smart phones (video resolution varies from 720p to 1080p). The subjects were not required to stay rigid and could take the video at their ease. The videos were taken in arbitrary environments and the lightings were not controlled. Note that the quality of mobile selfie videos are usually worse than the online videos due to large motion blurs caused by hand moving, auto focus, and auto exposure from phone cameras, although a higher frame resolution is accessible. The selfie video is approximately 15 seconds sampled at 20fps (369 frames, 277 frames, 229 frames, 300 frames, 267 frames, 376 frames, 383 frames, 235 frames and 256 frames for each individual from top to bottom of Figure 7).

Later, the semantic segmentation method of [Zheng et al. 2015] was run on video frames to remove the background and foreground occlusions such as microphones. We ran VisualSFM [Wu 2011] on the pre-processed frames. In [Wu 2011], the non-rigid face expression change might cause large distortions in the reconstructed views; thus we set radial distortion to zero.

Runtime We ran our algorithm on a single PC with a 12 core i7 CPU, 16GB of memory and four NVIDIA GTX 1080 Ti graphics cards. For a typical online video, the preprocessing and structure from motion plus visual hull in Figure 2(A) takes 40 minutes. Extracting 2D query strands in Figure 2(B) takes 3 minutes. The head shape reconstruction and texture extraction takes 3 minutes to run. Hair database retrieval and deformation in Figure 2(D) is 40 minutes with 500 candidates. The 3D orientation local deformation and final hair strand generation from the reconstructed head takes 2 min.

8.2 Results and Comparisons

Figure 7 and Figure 8 show the results together with the reference frames from the videos. We can see that the reconstructions are good for a variety of lighting conditions, diverse people and hairstyles. See also the supplementary videos.

Next, we compared our results to the state-of-the-art hair in-the-wild reconstruction methods [Chai et al. 2016; Hu et al. 2017; Zhang

et al. 2017]. More view comparisons are shown in the supplementary video.[1] We performed qualitative, quantitative and user study comparisons below.

Figure 9 shows comparisons for single-view methods. We picked a frontal frame from each of the 5 video clips of celebrities as input. We compared our untextured results with [Chai et al. 2016] and textured results with [Hu et al. 2017]. Note that our 3D models captured more personalized hairstyles; for example in Adele's case (the 1st row), [Chai et al. 2016] produced a short hairstyle, while Adele has a long hairstyle. Compared to [Hu et al. 2017], where each hairstyle has a flat back, our results show more variety.

In [Zhang et al. 2017], frontal, left, and right views are manually chosen from the same video clip. Since we do not have the back view in our video frames and the back view is necessary for the four-view reconstruction method, the authors were allowed to use any back view image they could find to reconstruct (the authors did not reconstruct Adele; back view photos can be found in the supplementary video). In our algorithm, we did not use the back view photo of the person. Our results are similar to [Zhang et al. 2017]; however ours are closer to the input; this can be seen by looking at the result of Justin (the 4th row) produced by [Zhang et al. 2017] which has a larger volume.

We did a **quantitative comparison** by projecting the reconstructed hair as lines onto the images, computing the intersection-over-union rate to the ground truth hair mask (manually labeled, but not used in our training or testing of the hair classifiers) per frame. We show the average IOUs over all the frames of each subject in Table 1. A larger IOU means that the reconstructed hair approximates the input better. We used the same camera pose of each frame estimated from structure from motion, where a perspective camera model is assumed, to project both the results from [Zhang et al. 2017] and our results. In total, our reconstruction results get an average IOU rate of around 0.8, while the four-view reconstruction method gets an average IOU of around 0.6. We showed the projection and ground truth hair mask of some example frames in Figure 10. Our results resemble the hairstyles better in all the frames. Note that since the back view image used in [Zhang et al. 2017] does not necessarily come from the same person, the inconsistency between

---

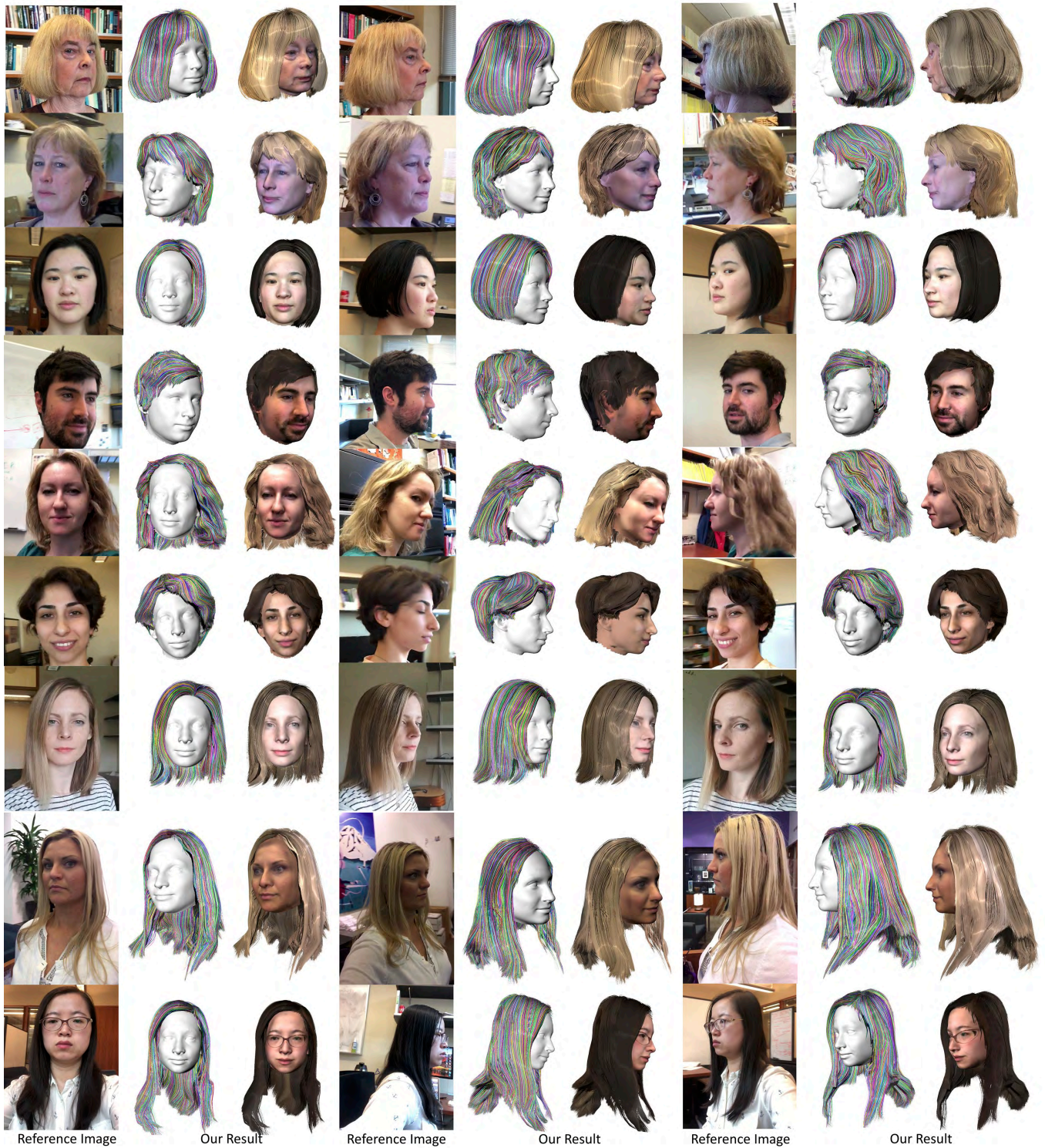[1]We thank the authors of those papers for helping creating comparison results.

Fig. 7. Reconstruction results from mobile selfie videos of different people in different environments.
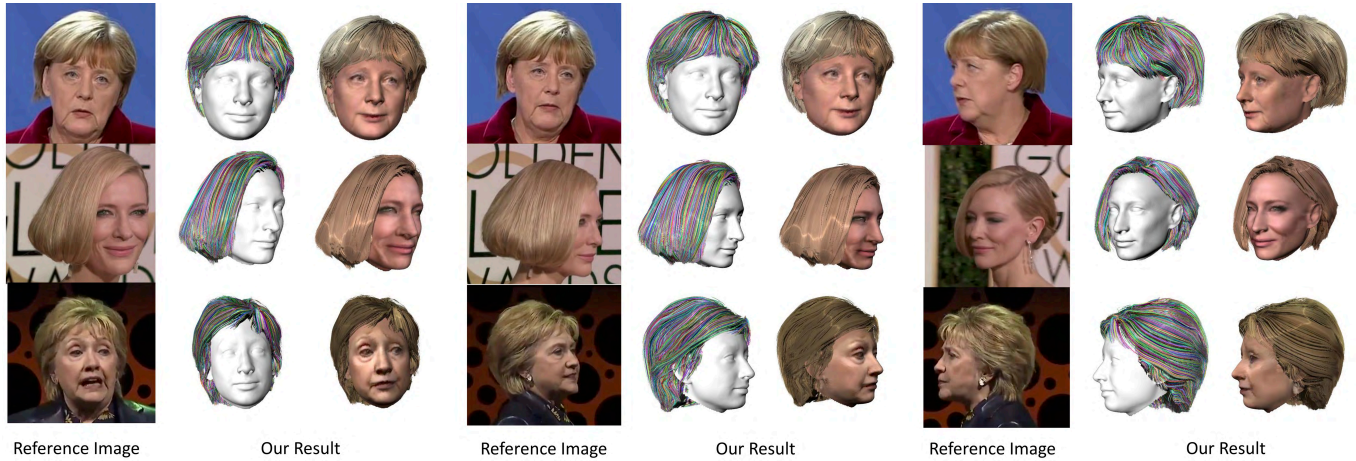
Fig. 8. Example results of our method. From top to bottom, the view coverage for Angela Merkel's video is 15 degree to −75 degree, 67 to −75 degree for Cate Blanchett and 60 to −74 degree for Hillary Clinton. Note that we can even create a natural looking result for Angela Merkel with a small view coverage.
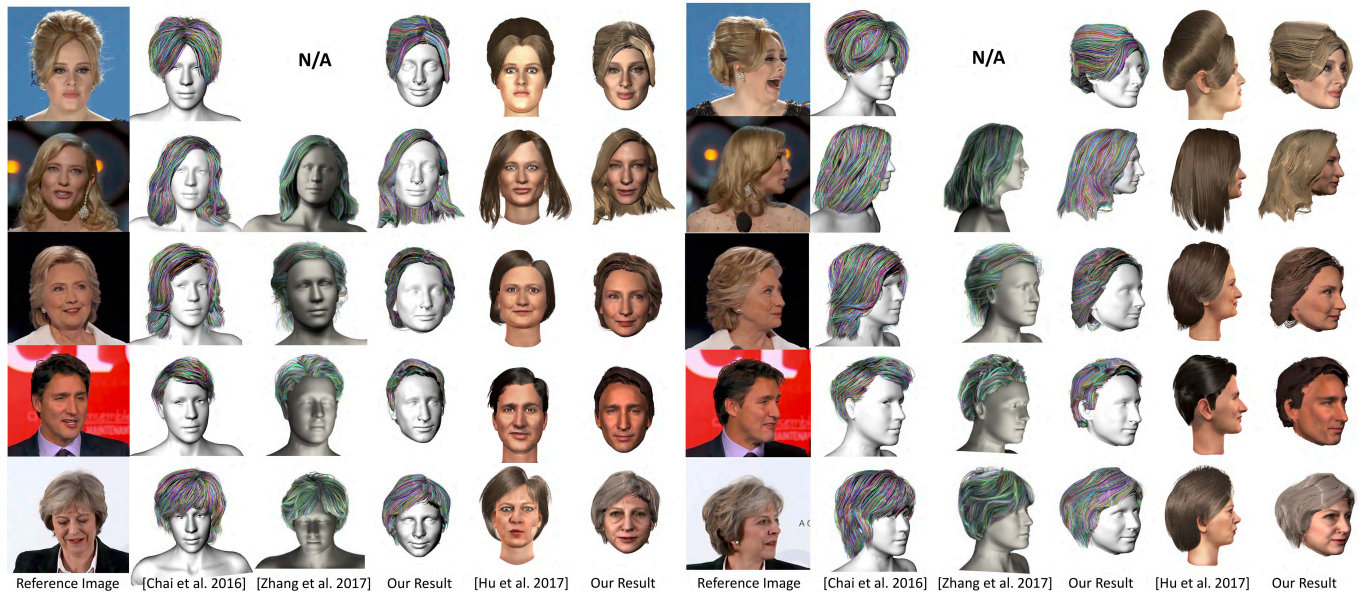

Fig. 9. This figure shows our results compared to the state-of-the-art methods. For each subject, we show the results in frontal and side views. For each view, the first column shows a reference frame from the video, then we show in the order of the untextured results from [Chai et al. 2016], [Zhang et al. 2017], our method and the textured results from [Hu et al. 2017], our method. Note how our result captures more personalized hair details, as also indicated by human studies and quantitative comparisons. **More view comparisons are provided in the supplementary video.**

four views might affect the final results. Also, in [Zhang et al. 2017], the authors assumed an orthographic camera model, which might account for some of the difference.

User Study We performed Amazon Mechanical Turk studies to compare our results to other methods. We showed two results side by side with the ground truth image in different views and asked which shape was more similar to the input, ignoring the face, shoulder and rendering qualities. For each subject, we did 3 groups of studies comparing the frontal, left, and right views. To

Table 1. IOU accuracy between the projected reconstructed hair and the hair segmentation (manually labeled ground truth).

Subject	Frames	[Zhang et al. 2017]	Ours
Hillary	266	0.6295 ± 0.0411	**0.8294 ± 0.0446**
Theresa	252	0.6598 ± 0.0258	**0.8111 ± 0.0216**
Cate	255	0.5991 ± 0.0474	**0.7749 ± 0.0669**
Justin	307	0.4787 ± 0.0882	**0.8028 ± 0.0187**

Fig. 10. This figure shows four example frames comparing the silhouettes of the reconstructed hairstyles to the hair segmentation results. The red mask is the annotated groundtruth hair mask over the image frame. The green mask shows the projected silhouettes from our method over the image and the blue mask shows the projected silhouettes from [Zhang et al. 2017].

Table 2. The ratio of preference to our results over total compared to [Zhang et al. 2017] based on Amazon Mechanical Turk tests.

Subject	frontal	left	right	total
Hillary	39/40	27/40	27/40	93/120
Theresa	13/40	26/40	27/40	66/120
Cate	30/40	26/40	32/40	88/120
Justin	27/40	37/40	38/40	102/120

Table 3. The ratio of preference to our results over total compared to [Hu et al. 2017] based on Amazon Mechanical Turk test.

Subject	0°	15°	90°	total
Adele	29/40	32/40	38/40	99/120
Hillary	35/40	38/40	36/40	109/120
Theresa	35/40	37/40	39/40	111/120
Cate	40/40	40/40	40/40	120/120
Justin	39/40	35/40	32/40	106/120

remove bias, we switched the order of the two results and did 3 more groups of studies. Each view was rated by 40 Turkers, giving a total of 120 different Turkers for each subject. We reported the rate of preference to our results over total in Table 2. Our results achieved an average preference rate of 72.7% in all the study groups. Similarly, we did a comparison of the textured results to the avatar digitalization work [Hu et al. 2017] showing the 0°, 15° and 90° views. The ratio of preferences is reported in Table 3. In total, our results were considered better by 90.8% of the Turkers.

Robustness: To evaluate the robustness of our hair segmentation classifier, we counted the failure frames for each input video: 2 frames for Adele, 3 and 0 frames for Cate Blanchette, 1 and 20 frames for Hillary Clinton, 0 frames for Justin Trudeau, 0 frames for Theresa May, and 4 frames for Angela Merkel (2 videos each for Cate and Hillary). For selfie video inputs, the numbers are 16 frames, 4 frames, 3 frames, 5 frames, 10 frames, 4 frames, 3 frames, 5 frames and 0 frames for each individual from top to bottom of Figure 7. We had an average successful segmented frame rate of

98.7%. Generally we observed more failure segmentation frames on selfie videos due to the motion blur when the subject was switching hands and more uncontrolled lighting compared to celebrity videos. We ran simulation experiments to test how robust our system is against failure segmentation frames. We gradually decreased the number of frames (frames selected randomly) used to generate the rough hair shape X_h and calculated the average IOU between the projected X_h and the ground truth hair label on four celebrity videos. The result is plotted in Fig. 11. Our system is quite robust even when half of the frames are dropped; however, as the number of failure frames increases we cannot guarantee a good global shape, which will lead to a poor reconstruction result.

3D vs. 2D: We further compared retrieving the best matching hairstyle in 3D vs retrieving in 2D. For each subject, we used the same frames as we used in 3D retrieval as query input. Instead of reprojecting each 2D strand back to 3D using per-frame depth information, we projected the candidates from the hairstyle database
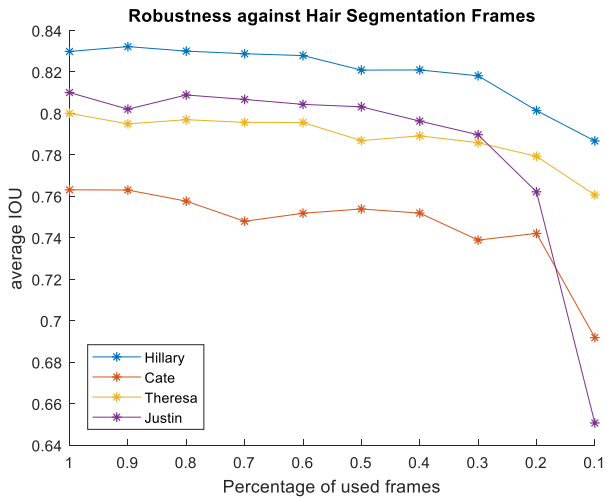
Fig. 11. This figure shows how similar the rough hair shape X_h to groundtruth is as the number of good hair segmentation frames decreases.

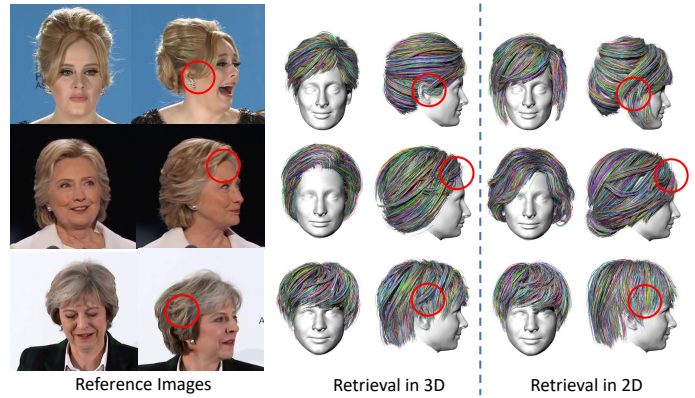Reference Images   Retrieval in 3D   Retrieval in 2D

Fig. 12. This figure shows the comparison for retrieving in 3D and 2D. The left two columns show two reference frames from the input video. The two columns in the middle show the frontal and side view of the best matching hairstyle from the database from 3D retrieval. The right two columns show the best matching result from 2D retrieval. The 2D retrieved results are generally similar to frontal views, however, do not look as similar as the 3D retrieval results in side views.

after pruning (Section 7) to each frame and computed the query distance similar to Equation 1, but in 2D. We show the best matching result before any deformation in Fig. 12. There are three main reasons that we chose to retrieve in 3D instead of 2D. 1) The multi-frames from the same video have a lot of overlapping regions, which caused redundancy in the query input when retrieving using each 2D frame. 2) Projecting the hairstyles from the database increased the computational cost compared to projecting the sparse strands back to 3D. We might pre-generate projected 2D views to a set of pre-defined poses as in the single-view method [Chai et al. 2016], however, since our input video does not have fixed views, we would need to define a larger pose space than the single-view input. 3) The retrieved results are usually biased towards the frontal view, because the hair strands in relatively frontal views are seen in more frames than the side views as shown in Fig. 12. For example, in Adele's comparison, the two hair strips in the front contributed more to the retrieval, making the best matching result less like the input subject from the side view. By projecting the strands to 3D, we allow the information from different views to contribute equally to the retrieval, as well as improving computational cost.

9 LIMITATIONS AND APPLICATIONS

9.1 Limitations

Our method cannot work on highly dynamic hairstyles due to the high non-rigidity of the hair volumes across the in-the-wild videos. See the example video frames of Olivia Culpo in Figure 13(a). The human hand interaction and body occlusion make the segmentation difficult. Explicitly modeling long hair dynamics is beyond the scope of the paper, and we assume small dynamics from the input videos. We also could not reconstruct very curly hairstyles and complicated hairstyles such as braids.

Our method also fails on videos where the background is too complicated as shown in Figure 13(b). The other people or crowds

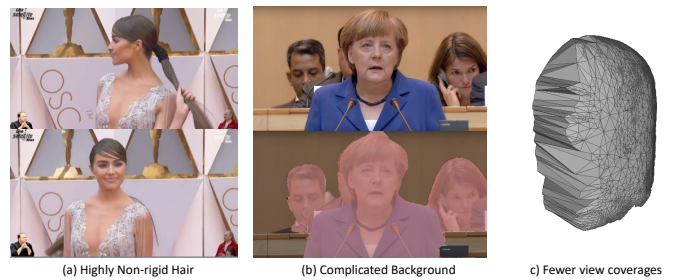(a) Highly Non-rigid Hair   (b) Complicated Background   c) Fewer view coverages

Fig. 13. Limitations of our algorithm. In (a) we show example video frames of highly non-rigid hairstyle. In (b) we show an example video frame with a complicated background. In (c) we show the back of a deformed hair mesh towards a visual hull from a small view coverage input.

in the background make it hard to estimate the head silhouette of the person and will lead to incorrect correspondences when running structure-from-motion.

For the low-confidence view corrections, we require an input video covering a view range of at least 90 degrees. Fewer views will cause the visual hull to be extremely distorted as shown in Figure 13(c), where our deformable registration will fail with a large fitting error. Note that as the view coverage decreases, this problem will be reduced to a single-view reconstruction problem.

9.2 Applications

Our reconstructed models can be now used for a variety of applications as shown in Figure 14(a)(b); we can also change the overall color of the hairstyle, making it darker or lighter, or morphing it to another hairstyle.

Since the hair roots of all our hair models are transfered from the generic shape used to compute the head model, we can assume
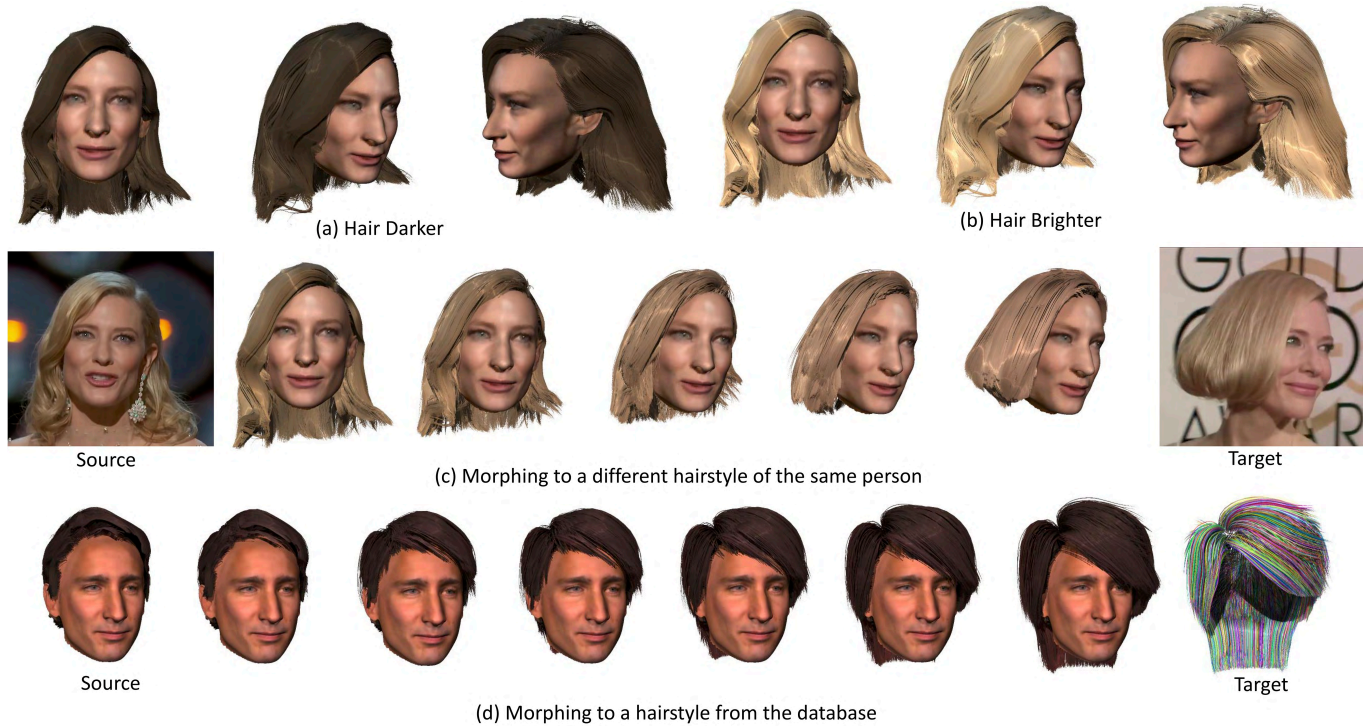
(a) Hair Darker

(b) Hair Brighter

Source

(c) Morphing to a different hairstyle of the same person

Target

Source

(d) Morphing to a hairstyle from the database

Target

Fig. 14. Hairstyle change examples. We show a darker and lighter version of Cate Blanchett's hairstyle in (a)(b). (c) shows the hair morphing intermediate results from two different hairstyles of the same person. (d) shows the hair morphing from the person's reconstructed hairstyle to a given hairstyle from the dataset.

that hair root points and hair strands are in correspondence for the same person. We resampled all the hair strands with the same number of vertices (50 in our implementations), which can be used for applications such as personalized strand-level hairstyle morphing.

In Figure 14(c), we show the hair morphing result of Cate Blanchett in two hairstyles from two different videos. The intermediate results are created by a one-to-one strand interpolation of the source and target hair strands. We can also morph the reconstructed hairstyle to a given hairstyle from the dataset as shown in Figure 14(d). The given hairstyle was re-grown from its 3D orientation field using the same set of scalp points to create correspondences to the original hairstyle. We trimmed the hair strands that intersect with the face during interpolation.

10 DISCUSSION AND FUTURE WORK

We have described a method that takes as input a video of a person's head in the wild and outputs a detailed 3D hair strand model combined with a reconstructed 3D head to produce a full head model. This method is fully automatic and shows that a head model with higher fidelity can be recovered by combining information from video frames. Our method is not restricted to specific views and head poses, making the full head reconstruction from in-the-wild videos possible. We showed our results on several celebrities as well as mobile selfie videos and compared our work to the most recent state of the art.

However, there are still a number of limitations and possible extensions to explore. One direction is to refine the rough hair mesh estimation for non-rigid hairstyles. Currently in our input, the person's head moves gently and our rough hair mesh is generated from the visual hull, which is only an approximation of the real hair volume, since the hair is non-rigid. We might explore using the ARkit of a smart phone to provide extra depth information to align the hair volume densely across frames. Also, we currently rely on the rigid camera poses estimated using structure from motion with SIFT features to connect all the views. In the future, we want to use facial features and hair specific features to increase the robustness of the frame alignment.

In our paper, we aim at a more challenging problem which is reconstructing 3D hair models from in-the-wild data, so some results are still not highly detailed, and we cannot handle complicated hairstyles such as braids and highly curled hairs.

We created face textures from video frames, which caused artifacts due to hair occlusions, decoration occlusions and low resolution of the faces. In the future, we can use generate photo-realistic textures as in [Saito et al. 2016]. Finally, a future extension to incorporate a facial blend shape model or estimate per frame facial dynamics as in [Suwajanakorn et al. 2014] to create a fully animatable model would be desirable.

ACKNOWLEDGMENTS

REFERENCES

Oleg Alexander, Graham Fyffe, Jay Busch, Xueming Yu, Ryosuke Ichikari, Andrew Jones, Paul Debevec, Jorge Jimenez, Etienne Danvoye, Bernardo Antionazzi, et al. 2013. Digital Ira: creating a real-time photoreal digital actor. In *ACM SIGGRAPH 2013 Posters*. ACM, 1.

B. Allen, B. Curless, and Z. Popović. 2003. The space of human body shapes: reconstruction and parameterization from range scans. In *ACM Transactions on Graphics (TOG)*, Vol. 22. ACM, 587–594.

Thabo Beeler, Bernd Bickel, Paul Beardsley, Bob Sumner, and Markus Gross. 2010. High-quality single-shot capture of facial geometry. *ACM Transactions on Graphics (TOG)* 29, 4 (2010), 40.

Volker Blanz and Thomas Vetter. 1999. A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*. ACM Press/Addison-Wesley Publishing Co., 187–194.

Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. 2013. Facewarehouse: a 3d facial expression database for visual computing. (2013).

Chen Cao, Hongzhi Wu, Yanlin Weng, Tianjia Shao, and Kun Zhou. 2016. Real-time facial animation with image-based dynamic avatars. *ACM Transactions on Graphics (TOG)* 35, 4 (2016), 126.

Menglei Chai, Linjie Luo, Kalyan Sunkavalli, Nathan Carr, Sunil Hadap, and Kun Zhou. 2015. High-quality hair modeling from a single portrait photo. *ACM Transactions on Graphics (TOG)* 34, 6 (2015), 204.

Menglei Chai, Tianjia Shao, Hongzhi Wu, Yanlin Weng, and Kun Zhou. 2016. Autohair: Fully automatic hair modeling from a single image. *ACM Transactions on Graphics (TOG)* 35, 4 (2016), 116.

Menglei Chai, Lvdi Wang, Yanlin Weng, Xiaogang Jin, and Kun Zhou. 2013. Dynamic hair manipulation in images and videos. *ACM Transactions on Graphics (TOG)* 32, 4 (2013), 75.

Menglei Chai, Lvdi Wang, Yanlin Weng, Yizhou Yu, Baining Guo, and Kun Zhou. 2012. Single-view hair modeling for portrait manipulation. *ACM Transactions on Graphics (TOG)* 31, 4 (2012), 116.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. 2016. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv preprint arXiv:1606.00915* (2016).

Paul Debevec. 2012. The light stages and their applications to photoreal digital actors. *SIGGRAPH Asia* (2012).

Hongbo Fu, Yichen Wei, Chiew-Lan Tai, and Long Quan. 2007. Sketching hairstyles. In *Proceedings of the 4th Eurographics workshop on Sketch-based interfaces and modeling*. ACM, 31–36.

Michael Goesele, Noah Snavely, Brian Curless, Hugues Hoppe, and Steven M Seitz. 2007. Multi-view stereo for community photo collections. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. IEEE, 1–8.

Hairbobo. 2017. Hairbobo. http://www.hairbobo.com/faxingtupian. (Sept. 2017).

Liwen Hu, Chongyang Ma, Linjie Luo, and Hao Li. 2014a. Robust hair capture using simulated examples. *ACM Transactions on Graphics (TOG)* 33, 4 (2014), 126.

Liwen Hu, Chongyang Ma, Linjie Luo, and Hao Li. 2015. Single-view hair modeling using a hairstyle database. *ACM Transactions on Graphics (TOG)* 34, 4 (2015), 125.

Liwen Hu, Chongyang Ma, Linjie Luo, Li-Yi Wei, and Hao Li. 2014b. Capturing braided hairstyles. *ACM Transactions on Graphics (TOG)* 33, 6 (2014), 225.

Liwen Hu, Shunsuke Saito, Lingyu Wei, Koki Nagano, Jaewoo Seo, Jens Fursund, Iman Sadeghi, Carrie Sun, Yen-Chun Chen, and Hao Li. 2017. Avatar Digitization from a Single Image for Real-time Rendering. *ACM Trans. Graph.* 36, 6, Article 195 (Nov. 2017), 14 pages. https://doi.org/10.1145/3130800.31310887

Alexandru Eugen Ichim, Sofien Bouaziz, and Mark Pauly. 2015. Dynamic 3D avatar creation from hand-held video input. *ACM Transactions on Graphics (TOG)* 34, 4 (2015), 45.

Wenzel Jakob, Jonathan T Moon, and Steve Marschner. 2009. Capturing hair assemblies fiber by fiber. In *ACM Transactions on Graphics (TOG)*, Vol. 28. ACM, 164.

Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 675–678.

Michael Kazhdan and Hugues Hoppe. 2013. Screened poisson surface reconstruction. *ACM Transactions on Graphics (TOG)* 32, 3 (2013), 29.

Ira Kemelmacher-Shlizerman. 2016. Transfiguring portraits. *ACM Transactions on Graphics (TOG)* 35, 4 (2016), 94.

Ira Kemelmacher-Shlizerman and Ronen Basri. 2011. 3d face reconstruction from a single image using a single reference face shape. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 33, 2 (2011), 394–405.

Ira Kemelmacher-Shlizerman and Steven M Seitz. 2011. Face reconstruction in the wild. In *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 1746–1753.

Aldo Laurentini. 1994. The visual hull concept for silhouette-based image understanding. *IEEE Transactions on pattern analysis and machine intelligence* 16, 2 (1994), 150–162.

Shu Liang, Ira Kemelmacher-Shlizerman, and Linda G Shapiro. 2014. 3d face hallucination from a single depth frame. In *3D Vision (3DV), 2014 2nd international conference on*, Vol. 1. IEEE, 31–38.

Shu Liang, Linda G Shapiro, and Ira Kemelmacher-Shlizerman. 2016. Head reconstruction from internet photos. In *European Conference on Computer Vision*. Springer, 360–374.

Sifei Liu, Jianping Shi, Ji Liang, and Ming-Hsuan Yang. 2017. Face Parsing via Recurrent Propagation. *arXiv preprint arXiv:1708.01936* (2017).

Sifei Liu, Jimei Yang, Chang Huang, and Ming-Hsuan Yang. 2015. Multi-objective convolutional learning for face labeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3451–3459.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3431–3440.

Linjie Luo, Hao Li, and Szymon Rusinkiewicz. 2013. Structure-aware hair capture. *ACM Transactions on Graphics (TOG)* 32, 4 (2013), 76.

Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2012. Hierarchical face parsing via deep learning. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2480–2487.

Fabio Maninchedda, Christian Häne, Bastien Jacquet, Amaël Delaunoy, and Marc Pollefeys. 2016. Semantic 3D Reconstruction of Heads. In *European Conference on Computer Vision*. Springer, 667–683.

Richard A Newcombe, Dieter Fox, and Steven M Seitz. 2015. DynamicFusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 343–352.

Sylvain Paris, Hector M Briceño, and François X Sillion. 2004. Capture of hair geometry from multiple images. In *ACM Transactions on Graphics (TOG)*, Vol. 23. ACM, 712–719.

Sylvain Paris, Will Chang, Oleg I Kozhushnyan, Wojciech Jarosz, Wojciech Matusik, Matthias Zwicker, and Frédo Durand. 2008. Hair photobooth: geometric and photometric acquisition of real hairstyles. In *ACM Transactions on Graphics (TOG)*, Vol. 27. ACM, 30.

Elad Richardson, Matan Sela, and Ron Kimmel. 2016. 3D face reconstruction by learning from synthetic data. In *3D Vision (3DV), 2016 Fourth International Conference on*. IEEE, 460–469.

Elad Richardson, Matan Sela, Roy Or-El, and Ron Kimmel. 2017. Learning detailed face reconstruction from a single image. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 5553–5562.

Shunsuke Saito, Lingyu Wei, Liwen Hu, Koki Nagano, and Hao Li. 2016. Photorealistic Facial Texture Inference Using Deep Neural Networks. *arXiv preprint arXiv:1612.00523* (2016).

Olga Sorkine, Daniel Cohen-Or, Yaron Lipman, Marc Alexa, Christian Rössl, and H-P Seidel. 2004. Laplacian surface editing. In *Proceedings of the 2004 Eurographics/ACM SIGGRAPH symposium on Geometry processing*. ACM, 175–184.

Supasorn Suwajanakorn, Ira Kemelmacher-Shlizerman, and Steven M Seitz. 2014. Total moving face reconstruction. In *Computer Vision–ECCV 2014*.

Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. 2015. What Makes Tom Hanks Look Like Tom Hanks. In *Proceedings of the IEEE International Conference on Computer Vision*. 3952–3960.

Justus Thies, Michael Zollhoefer, Matthias Niessner, Levi Valgaerts, Marc Stamminger, and Christian Theobalt. 2015. Real-time Expression Transfer for Facial Reenactment. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)* (2015).

Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gérard Medioni. 2017. Regressing robust and discriminative 3D morphable models with a very deep neural network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 1493–1502.

Nuttapon Vanakittistien, Attawith Sudsang, and Nuttapong Chentanez. 2016. 3D hair model from small set of images. In *Proceedings of the 9th International Conference on Motion in Games*. ACM, 85–90.

Dan Wang, Xiujuan Chai, Hongming Zhang, Hong Chang, Wei Zeng, and Shiguang Shan. 2011. A novel coarse-to-fine hair segmentation method. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*. IEEE, 233–238.

Kelly Ward, Florence Bertails, Tae-Yong Kim, Stephen R Marschner, Marie-Paule Cani, and Ming C Lin. 2007. A survey on hair modeling: Styling, simulation, and rendering. *IEEE Transactions on Visualization and Computer Graphics* 13, 2 (2007).

Yanlin Weng, Lvdi Wang, Xiao Li, Menglei Chai, and Kun Zhou. 2013. Hair interpolation for portrait morphing. In *Computer Graphics Forum*, Vol. 32. Wiley Online Library, 79–84.

Changchang Wu. 2011. VisualSFM: A visual structure from motion system. (2011).

Xuehan Xiong and Fernando De la Torre. 2013. Supervised Descent Method and its Applications to Face Alignment. In *IEEE Conference on Computer Vision and Pattern*

Recognition (CVPR).

Yaser Yacoob and Larry S Davis. 2006. Detection and analysis of hair. *IEEE transactions on pattern analysis and machine intelligence* 28, 7 (2006), 1164–1169.

Meng Zhang, Menglei Chai, Hongzhi Wu, Hao Yang, and Kun Zhou. 2017. A data-driven approach to four-view image-based hair modeling. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 156.

Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr. 2015. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*. 1529–1537.

Michael Zollhöfer, Matthias Nießner, Shahram Izadi, Christoph Rehmann, Christopher Zach, Matthew Fisher, Chenglei Wu, Andrew Fitzgibbon, Charles Loop, Christian Theobalt, et al. 2014. Real-time non-rigid reconstruction using an rgb-d camera. *ACM Transactions on Graphics, TOG* (2014).

Appendices

Hair segmentation is an important part of face parsing and full head reconstruction. [Yacoob and Davis 2006] detected hair based on the position relationship between face and hair and a simple color model. [Wang et al. 2011] proposed a coarse-to-fine hair segmentation method that starts from a coarse candidate region and performs graph-cuts to segment the hair. A CNN-based face parsing method [Luo et al. 2012] hierarchically combined several detectors to detect face components. [Liu et al. 2017, 2015] proposed multi-objective learning frameworks that could parse facial components as well as hair regions, but this model requires facial landmarks as prior inputs and can only handle simple hairstyles. To allow robust hair segmentation on various hairstyles, [Chai et al. 2016] trained a deep network specifically for the hair regions. However, their method requires pre-alignment of the face to detect the hair region. In recent years, fully convolutional networks (FCN) [Long et al. 2015] have been widely used for pixel-level segmentation. We adopted the FCN model for robust hair segmentation and trained a network to segment hair regions across various poses.

We collected 15, 977 hair salon images from a hairstyle design website [Hairbobo 2017] [2] and 3, 923 Internet images of celebrities from Google image search in various head poses and different hairstyles total of 19, 900. Each pixel in those images was labeled manually as hair or non-hair. To preserve continuity of the hair region, hair accessories were labeled as hair. The training images were not cropped nor aligned to increase robustness.

A   HAIR SEGMENTATION CLASSIFIER

We trained the hair segmentation classifier using a fully convolutional network with 13, 900 images out of all the collected photos, and the remaining were used for testing. Specifically, the FCN-32s model [Long et al. 2015] was used, and it was fine-tuned with PASCAL VOC data from the ILSVRC-trained VGG-16 model. To only detect the hair category, we changed the output number of the last convolution layer to one and add a sigmoid layer to get scores between 0 and 1. The output score represents the probability that the pixel belongs to hair. In our implementation, we resized all our input images to 500×500. With FCN-32s, we downsampled the output with a factor of 32. We later used bilinear interpolation to upsample the output heatmap to obtain the final segmentation result. We

fine-tuned the pretrained FCN-32s with the following parameters: minibatch size 1, learning rate 10^{-9}, momentum 0.99, and weight decay 0.0005. We froze all the layers except the last score layer in the first 100, 000 iterations. Then we fine-tuned all the layers in the next 100, 000 iterations.

We tested our segmenter on the 6, 000 test images. The hair segmentation network was implemented with Caffe [Jia et al. 2014] and C++ and ran on a NVIDIA GTX 1080 GPU with an inference time of 150ms for a 500×500 input image. The accuracy on the test images reached 0.9613, and the IOU rate reached 0.8585.

Using automatic hair segmentation (that works well across views, even back views) is the key to enabling a fully automatic hair modeling system. Our algorithm is capable of segmenting the hair region successfully in different views and head poses. However, it still fails to segment some hairstyles correctly when the hair color is too close to the background or when the image has a large motion blur.

We compared our classifier to two methods: [Liu et al. 2017] and DeepLab [Chen et al. 2016] (which was compared to in [Chai et al. 2016], but [Chai et al. 2016] does not provide code so we compared with DeepLab). We fine tuned and ran DeepLab on all of our test images and got a pixel accuracy of 0.8892 and IOU rate of 0.7173. For [Liu et al. 2017], we used the pre-trained model to run on only 41.3% of our test images, since it requires pre-detection of the face and fails on back and side views. The pixel accuracy for the 41.3% test images was 0.8462, and IOU rate was 0.5573.

B   HAIR DIRECTIONAL CLASSIFIER

For training, the hair areas were manaully divided into subregions based on their general growing trends. One of four directional labels was assigned in the labeling stage: $[0, 0.5\pi)$, $[0.5\pi, \pi)$, $[\pi, 1.5\pi)$, $[1.5\pi, 2\pi)$. Hair accessories and hair occlusions were labeled as undetermined region, and background pixels were labeled as background. We trained a modified VGG16 network as proposed in [Chai et al. 2016] on the hair regions to automatically predict the directional labels for each pixel, using a multi-class approach with 6 classes (4 directions, 1 background, 1 undetermined).

To train our hair directional classifier, we cropped and extracted the hair region, resized each image to 256×256 and downsampled 8 times with a bilinear filter. The output result was then upsampled to the original image size with bilinear interpolation and then followed by a CRF for per-pixel labels. In the training stage, we utilized the same set of 13, 900 images and augmented the dataset to 20, 000 images by image rotation, translation, and mirroring.

We implemented the classifier in the same environment as the segmenter and set the minibatch size to 32 and the initial learning rate to 0.0001 with exponential decay. Our network converged after 50k steps. The inference time was 59ms for a 256×256 input image. We ran the directional classifier on the same test set of 6, 000 images and got an accuracy of 0.9425.

Our classifier typically fails to generate a correct direction label for some small regions on the side of the face. However, in our pipeline, since we have video sequences, we can still get a correct direction from a different view. The availability of many views compensates for individual failure cases.

---

[2] http://www.hairbobo.com/faxingtupian