Total Moving Face Reconstruction

Supasorn Suwajanakorn, Ira Kemelmacher-Shlizerman, and Steven M. Seitz

University of Washington, USA

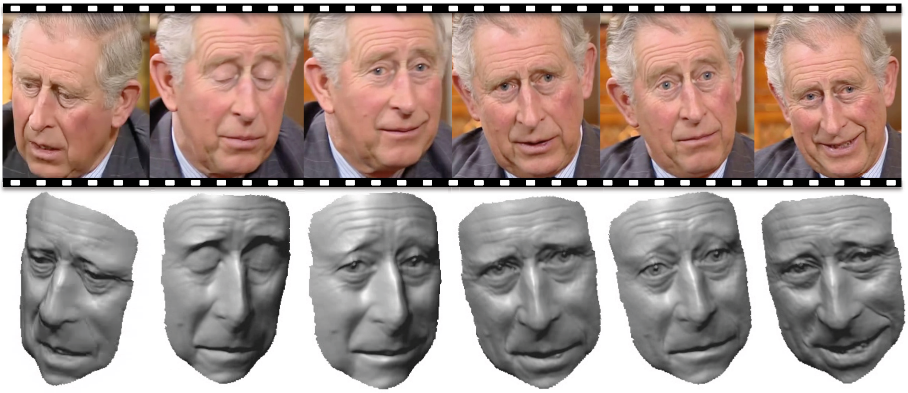

Fig. 1. Given a YouTube video of a person's face our method estimates high detail geometry (full 3D flow and pose) in each video frame completely automatically

Abstract. We present an approach that takes a single video of a person's face and reconstructs a high detail 3D shape for each video frame. We target videos taken under uncontrolled and uncalibrated imaging conditions, such as youtube videos of celebrities. In the heart of this work is a new dense 3D flow estimation method coupled with shape from shading. Unlike related works we do not assume availability of a blend shape model, nor require the person to participate in a training/capturing process. Instead we leverage the large amounts of photos that are available per individual in personal or internet photo collections. We show results for a variety of video sequences that include various lighting conditions, head poses, and facial expressions.

Keywords: 3D reconstruction, faces, non-rigid reconstruction.

1 Introduction

Reconstructing the time-varying geometry of a person's face from a video is extremely challenging. Indeed, the highly nonrigid nature of the human face, coupled with our ability to discern even minute facial details and geometry flaws,

D. Fleet et al. (Eds.): ECCV 2014, Part IV, LNCS 8692, pp. 796–812, 2014.

make it very difficult to achieve high quality results. Operating on free-form video captured "in the wild" adds another level of complexity; only a handful of such results have been reported in the literature [14,13,23,20,17].

Rather than reconstruct the input video in isolation, suppose that we had access to a large collection of other photos of the same person captured at different times, with varying pose, expression and lighting. Indeed, most people are captured in numerous photos and videos over their lifetimes; we propose to leverage the *total corpus of available imagery* of the same person to help reconstruct his/her face in an input video. We call this problem *total* moving face reconstruction.

Virtually all modern 3D face tracking and video reconstruction approaches leverage an assumption that the human face is well represented by a linear combination of *blend shapes*, e.g., Morphable models [9,10,41], AAMs [22,19], and Nonrigid Sfm [14,13,23,20]. The advantage of the blend-shape model is that it makes the problem more constrained, as the number of parameters (blend shapes and/or coefficients) is less than the number of measurements (pixels in the video). The main disadvantage is the low-rank model limits expressiveness and the ability to capture fine details.

Instead, our approach is based on deriving a person-specific face model (from all available imagery), and fitting it to each image in the video using a novel 3D optical flow approach coupled with shading cues. The combination of flow and shading enables capturing even minute shape variations (e.g., dimples, wrinkles, pimples, etc.) over the sequence.

We leverage the corpus of images to compute a person-specific face model that captures both the average 3D shape and the illumination-dependent appearance subspace. One key property of this model is that it enables appearance matching of any new image, and solving for dense correspondence via a 3D optical flow approach, yielding more precise alignment and robust 3D tracking than are possible by matching sparse fiducials, e.g., [17]. Another key property is that our use of previously captured photos enables accurate reconstruction even under degenerate motions (e.g., no head rotation) that foil nonrigid structure-from-motion methods [14,13,23,20]. Finally, we incorporate shading cues to obtain higher resolution details than are possible to capture with any other method.

2 Related Work

High quality time-varying 3D face geometry capture is extremely challenging due to highly non rigid nature of the human face–ultimately we would like to capture wrinkles, eye and muscle movement, dimples, detailed mouth expressions, eye lid details, and so forth. All these together form our perception of a person's face and are highly important for further face analysis.

Early methods in 3D facial performance capture use marker-based motion capture systems, e.g., [26], that track a sparse set of markers on a person's face. This requires the person to spend hours in a lab, and tracks only a sparse set of points. In contrast, modern high detail reconstruction methods use multi-view stereo approaches on input coming from multiple high resolution synchronized cameras which does not require markers but assumes calibration of the cameras and controlled lighting [7,8,11] or uncontrolled lighting[42]. Structured light [45] and light stages [16,2,3,25] provide the ability to use multiple synchronized and calibrated lights for reconstruction.

Recently, RGBD cameras were proven to be extremely successful in face and expression tracking [10,41,33]. The idea is to fit raw depth camera output to a deformable facial expression model (blend shapes) created by an artist for facial expression retargeting, puppeteering, and high quality face tracking. Similarly, [17] showed that it is possible to achieve high quality tracking via 3D regression and fitting to a blend shape model extracted from large number of face shapes captured via kinect fusion method [18]. These methods achieve very impressive face tracking results, however 1) require the person to participate in the training stage or be present in front of a depth camera, and 2) assume that face shapes can be represented by a linear combination of blend shapes. Representing face shapes using linear combinations of laser scans of other people's faces and artist created blend shapes goes back to the classical work by Blanz and Vetter [9] as well as more recent works by [21,40]. These however only enable capture of large scale deformations and tend to miss the fine details (wrinkles, dimples, etc.) that distinguish individuals.

Non-rigid structure from motion methods enable reconstruction from a single video by creating a linear basis for the non rigid motion that appears in the particular video; correspondence between the frames is typically given [14,13,20] or estimated via optical flow [23]. The major drawback of these methods is that the basis is extracted from the video itself which not only limits the ability to capture fine details, but also requires head pose to change significantly throughout the video to enable basis reconstruction.

Most related to our work are single view methods, particularly [30,28,27,29]. These methods can produce detailed reconstructions, but do not estimate how the scene deforms over time. Similar to scene flow methods [39], we reconstruct a dense 3D flow field; key differences include our illumination invariance model, and that we compute 3D to 2D correspondence rather than 3D to 3D. Furthermore, recent scene flow methods either assume availability of a stereo pair of photos taken in the same rigid configuration (e.g., same expression) [37,38] or rigid motion throughout the video [4]. The most relevant to our work is [24], who also operate on monocular video and leverage motion and shading cues to reconstruct a moving face model. However, whereas we simply fit a rigid 3D model independently to each frame, their technical approach involves several additional steps including blend-shape coefficient fitting, keyframe selection, feature-point refinement, multi frame optical flow, and temporal shape filtering (we filter only pose, not shape or flow). We believe the success of our much simpler approach stems from our 3D flow model ([24] move mesh vertices only parallel to the image plane), and our use of *all available imagery* to build an illumination-invariant appearance model. Most importantly, their approach requires a prior, lab-captured model of each actor (requiring a stereo rig and manual work), and hence is *not* applicable to videos of celebrities or other content in personal photo and video collections.

In this paper we target high detail reconstruction from a single video captured *in the wild*, i.e., under uncontrolled imaging conditions. Instead of requiring the person to be scanned in the lab or participate in the reconstruction process (as many other methods require [24,18,10,41,33]), we leverage whatever existing imagery is available online or in personal photo collections. This enables applying our approach on YouTube videos of celebrities (e.g., video of Prince Charles[1] as in Figure 1), for which we produce arguably the best reconstructions to date.

3 Overview of the Method

Given a video of a person, we seek to reconstruct a moving 3D model of his/her face that captures apparent motion and fine-scale shape details as well as possible. Specifically, we compute a 3D reconstruction that optimally fits both the *image motion* and *shading* in each frame. Because the problem is not fully constrained (we have only one view of the deforming face at each time instant), we leverage *all available imagery* of the person's face (e.g., photos on the Internet or in personal collections) to compute a reference model of that person (Section 4), capturing both their average shape and appearance under a subspace of illuminations. The reference model is used to constrain the gross shape of the sought reconstruction.

To compute the 3D facial deformation in each frame, we formulate a novel 3D optical flow problem (Section 5.1) that computes dense correspondence between the 3D model and each video frame, and optimally deform the reference mesh to fit. Similarly, to capture wrinkles and other high frequency structures, we introduce a novel approach to deform the reference mesh so that, when rendered, the mesh shading fits the image shading as accurately as possible.

We note that our method does not guarantee an accurate fit to ground-truth geometry, as the shape of the face may change in each frame and single-image cues are not sufficient for this purpose. Rather, we seek to produce a reasonably convincing model (leveraging all available imagery) which optimally fits the image information in each frame.

4 Average Shape and Appearance from All Available Imagery

While a person's face shape may be slightly different at each time instant, their rough shape (e.g., distance between eyes, nose length, overall geometry), tends to be consistent over time. Hence, we leverage all available imagery (photos and/or video frames) to reconstruct a shape and appearance model of the person that captures their average shape and appearance under a subspace of illuminations.

[1] `http://www.youtube.com/watch?v=s89KEI2AfBU`

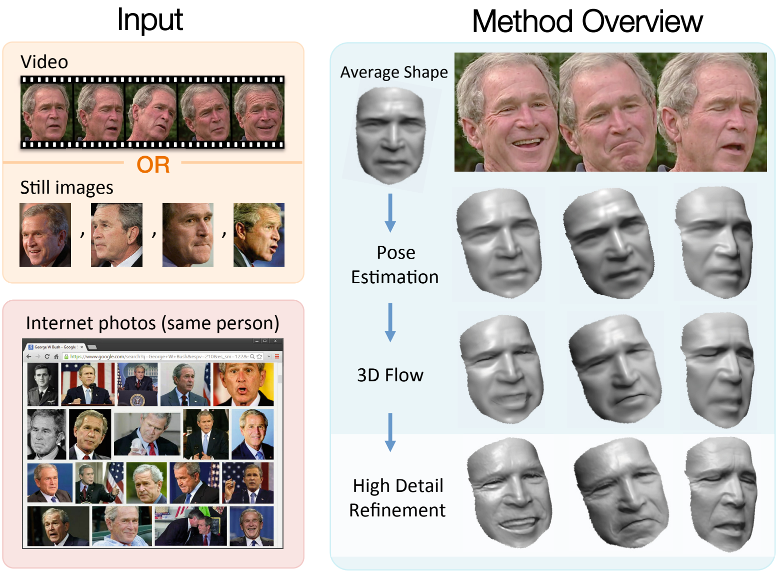

Fig. 2. Overview of our method. Given a video sequence we estimate 3D pose (average shape is rotated to the input pose for each of the 3 examples), followed by estimate of dense 3D flow of the average model to fit the input expression, and final refinement using shading cues (note the appearance of teeth, details in eyes, and so forth.)

In principle, this shape could be acquired in a number of different ways, e.g., a laser scan, kinect fusion model, stereo reconstruction, ohotometric stereo, etc. Given registered or rendered imagery of the same person under many different illuminations, we can construct an illumination subspace by projecting onto the first four singular vectors [5].

In practice, such 3D data with registered imagery is seldom available. Hence, we leverage Kemelmacher et al's Face Reconstruction in the Wild approach [31] to obtain an average shape and appearance model (rank-4 linear basis of the aligned image set). In practice, we find that aligning the images using Collection Flow [32] prior to reconstruction yields slightly sharper reconstructions. We will assume that as a result of this process we have obtained an average shape of the person v_{avg}, texture basis I_{avg}, and initial 3D pose estimate P_0.

5 Total Moving Reconstruction

We now describe our approach for reconstructing a moving 3D face shape by deforming an average model to fit the motion and shading cues in each video frame. The face in any given frame may have unknown and possibly changing lighting conditions, arbitrary facial expressions, and varying head orientation (even profile or other highly non-frontal poses are supported–see supplementary video).

Key to our approach is a metric based on *photo consistency*, i.e., comparing mesh renderings with input video frames. This capability depends critically on being able to match the illumination and shading in each input frame to that of the rendered mesh, a property achieved by our appearance subspace representation (Section 4).

We recover shape in two steps: first, we deform the average shape to fit the image motion, and second, we deform the resulting shape to fit the shading cues in each frame. We now formulate each problem in turn.

5.1 3D Flow Objective

Given an average shape, we seek a 3D flow field mapping it to the reconstructed shape in a given input image (video frame). Denote by $\mathbf{v} := (x, y, z)^\top$ a vertex on the average mesh we wish to deform, and $\boldsymbol{f}(\mathbf{v}) \in \mathbb{R}^3$ is the desired per vertex 3D flow (3D displacement to the reconstruction). As the average shape is provided as a depth map $\boldsymbol{d}(u, v)$, vertices are connected to form triangle meshes over 4 neighbor pixels in a regular 4-connected grid of the depth map and flow $\boldsymbol{f}(\mathbf{v})$ can also be parametrized on 2D image plane as $\boldsymbol{f}(u, v) = \boldsymbol{f}(u, v, \boldsymbol{d}(u, v))$. $I(u, v)$ gives the input image intensity at pixel (u, v), and denote $C(\mathbf{v})$ to be the intensity of vertex $\mathbf{v}$ in the rendering of the average shape from the viewpoint of the input image. Define the camera function as $\mathbb{P}\ : \mathbb{R}^3 \to \mathbb{R}^2$ which takes a vertex as input and applies a rigid transformation and weak-perspective projection to produce 2D point on the image plane. We therefore cast 3D flow as an optimization problem with the following objective:

$$E_{flow3d}(\boldsymbol{f}) = \sum_{\mathbf{v}} |I(\mathbb{P}(\mathbf{v} + \boldsymbol{f}(\mathbf{v}))) - C(\mathbf{v})|^2 + \alpha \left(|\nabla \boldsymbol{f}_x|^2 + |\nabla \boldsymbol{f}_y|^2 + |\nabla \boldsymbol{f}_z|^2\right) \tag{1}$$

where $|\nabla \boldsymbol{f}_x|^2 = \left(\frac{\partial \boldsymbol{f}_x}{\partial u}\right)^2 + \left(\frac{\partial \boldsymbol{f}_x}{\partial v}\right)^2$ is the gradient magnitude of the x component of flow parametrized on 2D image plane and $|\nabla \boldsymbol{f}_y|^2$, $|\nabla \boldsymbol{f}_z|^2$ along y and z and are defined similarly. $\alpha > 0$ is the smoothness weight that serves as a regularization parameter. We will describe how to optimize this function shortly.

5.2 Shape-from-Shading Objective

Applying the estimated 3D flow field $\boldsymbol{f}$ yields a new mesh $\mathbf{v}' = \mathbf{v} + \boldsymbol{f}$ that deforms the average shape to match the input image. While the resulting reconstruction captures dense nonrigid correspondence, it does not model the impact of the deformation on surface normals and their resulting shading effects. Hence, we introduce a second step to optimize the reconstruction to best fit the *shading* of the input image, by iteratively deforming the mesh vertices and re-rendering.

Specifically, we optimize for new z-coordinate $z(\mathbf{v}')$ of each vertex $\mathbf{v}'$ by minimizing the sum of photometric and position error terms:

$$E_{shading}(z) = \sum_{\mathbf{v}'} |I\left(\mathbb{P}(\mathbf{v}')\right) - \boldsymbol{l}^\top \boldsymbol{h}_{\mathbf{v}'}\left(z(\mathbf{v}')\right)|^2 + \beta |z(\mathbf{v}') - \mathbf{v}'_z|^2 \tag{2}$$

$\mathbf{v}'_z$ is the original z-coordinate of $\mathbf{v}'$ after 3D flow, $\boldsymbol{h}_{\mathbf{v}'}$ is a 4D spherical harmonics approximation to surface reflectance at new vertex mesh $(\mathbf{v}'_x, \mathbf{v}'_y, z(\mathbf{v}'))$ and $\boldsymbol{l}$ is a 4D vector of spherical harmonics coefficients. β is a regularization weight for the second position error term that constrains final z to be close to the original shape. We describe in detail each of the optimization steps in the following subsections.

6 Optimization

We now describe our optimization approach for computing 3D flow and shading-based mesh refinement. Our approach requires an initial estimate of 3D head pose and lighting (described in Section 6.3).

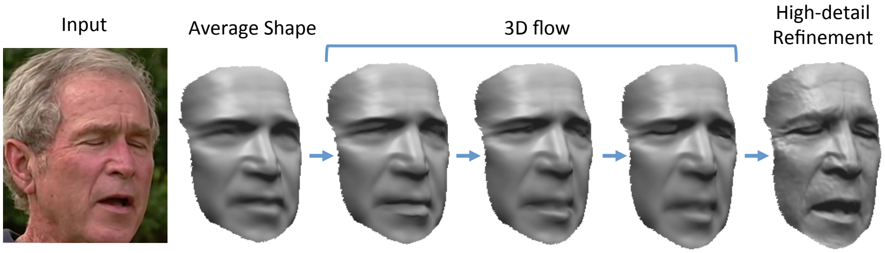

Fig. 3. 3D flow convergence example. The optimization starts from an average model of Bush with closed mouth, the mouth opens with 3D flow estimation iterations and gets refined at the shading step. This computation is done independently for each single frame in the video (temporal constraint is applied only at the rigid pose estimation step).

6.1 3D Flow Estimation

Minimizing Eq. 1 is a non-linear optimization task even if we assume weak-perspective projection with L2 norm because $I\left(\mathbb{P}(\mathbf{v} + \boldsymbol{f}(\mathbf{v}))\right)$ is generally non-linear in the image coordinate. To optimize this objective, we use Levenberg-Marquardt (LM) implemented in the Ceres Solver [1]. This requires a calculation of the Jacobian matrix in which the variables are x, y, and z for each flow value. To compute the derivatives of $I\left(\mathbb{P}(\mathbf{v} + \boldsymbol{f}(\mathbf{v}))\right)$ with respect to each flow component x, y and z, let us denote $\mathbb{P}(\mathbf{v} + \boldsymbol{f}(\mathbf{v})) = (u, v)^\top$ and $\boldsymbol{f}(\mathbf{v}) = (x, y, z)^\top$. By applying the chain rule with respect to x we get:

$$\frac{\partial}{\partial x} I\left(\mathbb{P}(\mathbf{v} + \boldsymbol{f}(\mathbf{v}))\right) = I_u \frac{\partial u}{\partial x} + I_v \frac{\partial v}{\partial x} \tag{3}$$

where I_u and I_v denote image derivatives along the horizontal and vertical axis and are computed using the 5-point derivative filter $\frac{1}{12}[-1 \; 8 \; 0 \; -8 \; 1]$. Let us further define the camera function as

$$\mathbb{P}(\boldsymbol{q}) = \pi(\mathbf{R}_{3\times 3}\boldsymbol{q} + \mathbf{T}_{3\times 1}) \tag{4}$$

$$\pi(\boldsymbol{r}) = (f \cdot \boldsymbol{r}_x/\bar{z}, f \cdot \boldsymbol{r}_y/\bar{z})^\top \tag{5}$$

where $\mathbf{R}_{3\times 3}$ is a rotation matrix and $\mathbf{T}_{3\times 1}$ is a translation vector. π is a weak-perspective projection with $\bar{z}$ being the constant average of vertex z-coordinate; $\frac{\partial u}{\partial x}$ and $\frac{\partial u}{\partial y}$ are evaluated using automatic differentiation. This provides a derivative with respect to x, derivatives with respect to y and z are computed similarly.

To differentiate the smoothness term, we approximate the partial derivatives of $\nabla \boldsymbol{f}_x, \nabla \boldsymbol{f}_y, \nabla \boldsymbol{f}_z$ by forward differences (i.e., re-parametrize flow on 2D image plane $\frac{\partial \boldsymbol{f}_x}{\partial u} = \boldsymbol{f}_x(u+1, v) - \boldsymbol{f}_x(u, v), \frac{\partial \boldsymbol{f}_x}{\partial v} = \boldsymbol{f}_x(u, v+1) - \boldsymbol{f}_x(u, v)$), and then take the derivatives. Similar computation is done for y and z components.

We implement this in a coarse-to-fine multi-resolution scheme [15] to deal with large flow displacements, i.e., we construct a Gaussian pyramid of the input image with down sampling rate of 0.75, and use the output flow in a coarser level as an initialization for the next finer level.

6.2 Shading-Based Refinement

We deform the average mesh to fit the input face according to the estimated 3D flow and use this new mesh as initialization to shading based mesh refinement. The idea is to capture high frequency details, e.g., wrinkles, folds, etc. We assume Lambertian reflectance and use the 1st order spherical harmonics (SH) approximation to Lambertian reflectance [6] to model the relationship between surface normals and image intensities. From Eq. 2, we define the SH approximation to surface reflectance at each new vertex $\mathbf{w} = (\mathbf{v}'_x, \mathbf{v}'_y, z)$ as

$$\boldsymbol{h}_{\mathbf{v}'} = \left(1, \frac{(\mathbf{w}_u - \mathbf{w}) \times (\mathbf{w}_v - \mathbf{w})}{\|(\mathbf{w}_u - \mathbf{w}) \times (\mathbf{w}_v - \mathbf{w})\|}\right)^\top \tag{6}$$

where $\mathbf{w}_u$ and $\mathbf{w}_v$ are vertices adjacent to $\mathbf{w}$ in the mesh structure along the positive horizontal and vertical directions. We estimate the SH coefficients $\boldsymbol{l}$ by finding the best coefficients that fit the deformed mesh after 3D flow to the input via:

$$\min_{\boldsymbol{l}} \sum_{\mathbf{v}'} |I\left(\mathbb{P}(\mathbf{v}')\right) - \boldsymbol{l}^\top \boldsymbol{h}_{\mathbf{v}'}\left(\mathbf{v}'_z\right)|^2. \tag{7}$$

To finally optimize Eq. 2, we pre-compute $I\left(\mathbb{P}(\mathbf{v}')\right)$ and further linearize by precomputing the normalizing factor $\|(\mathbf{w}_u - \mathbf{w}) \times (\mathbf{w}_v - \mathbf{w})\|$ as suggested in [30] using the deformed mesh. The resulting formulation becomes linear in z and solved efficiently using linear least squares optimization.

6.3 Pose and Lighting

Faces in input frames/photos may appear in an arbitrary 3D pose, and often in highly non-frontal poses, e.g., 90 degrees out of plane rotation. To estimate 3D

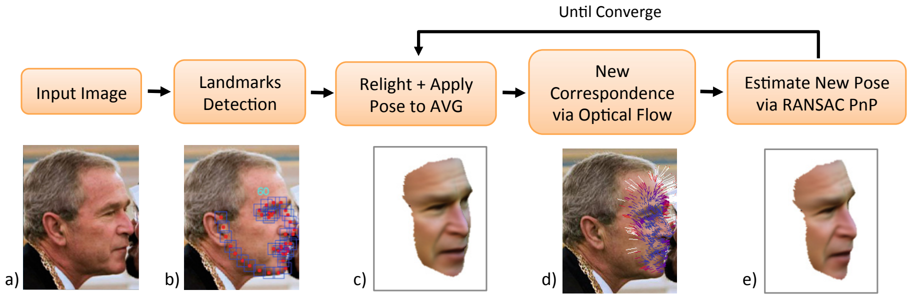
Until Converge
Input Image
Landmarks Detection
Relight + Apply Pose to AVG
New Correspondence via Optical Flow
Estimate New Pose via RANSAC PnP
a)
b)
c)
d)
e)

Fig. 4. Pose refinement algorithm. (a) non-frontal photo–challenging for current methods, (b) landmarks detection and (c) pose estimation using landmarks (slightly off) which is used to initialize our refinement. (d) optical flow matching between an average model rendering in the initial pose and input image. (e) final pose estimation result using PnP on dense point sets chosen via RANSAC.

flow we first need to compute the 3D rigid transformation $P = [\mathbf{R} \mid \mathbf{T}]$ that takes the average mesh $\boldsymbol{v}$ and transforms it to the position of the face in the image. While we obtain an initial estimate from the warping process in Section 4, it is performed using a 3D reference model of a different individual (see [31] for more details), thus pose estimation error increases with larger angles of rotation, e.g., due to difference in nose shape across people. We propose the following process (Alg. 1) to recover accurate face pose in a single photo, and we further show how to leverage temporal information in videos to achieve accurate pose estimates. We solve the Perspective-n-Point problem (PnP) using OpenCV's implementation of Levenberg-Marquardt [12]. Following the optimization in Alg. 1 we get high quality pose estimates for challenging poses. To achieve temporal coherence across the video, we refine the individual pose estimates using nearby frames.

```
Data: $P_0 = P_{ref}$ initialize pose from Sec. 4;
$I$: input image;
$A_P^L$: rendering of an average shape $v_{avg}$ with texture in pose $P$ and lighting $L$;
$i = 0$;
Result: 3D pose $P$
while until convergence do
    estimate lighting $L_i$ of input $I$ using process described in Sec. 4;
    render $v_{avg}$ in pose $P_i$ and input lighting $L_i$;
    run 2D optical flow between $A_{P_i}^{L_i}$ and $I$;
    generate 3D-to-2D correspondences from $v_{avg}$ to $I$ through 2D flow ;
    solve PnP using RANSAC on subset of correspondences:;
    solve PnP on all inliers to compute new estimate of pose $P_{i+1}$;
end
```

Algorithm 1. Out of plane pose estimation in a single photo.

Specifically, we use each frame's 12 neighbors and their corresponding poses for refinement, as follows. We compute bi-directional 2D optical flow between every consecutive pair of frames, then we concatenate them to produce flows between frame j and all its neighbors. Once these flows are available, we project 3D points of v_{avg} onto the image plane using pose estimate of frame $j+1$ then follow 2D flow from frame $j+1$ to j to produce dense 3D-to-2D correspondences between v_{avg} and the image pixels of frame j. Then we solve RANSAC PnP problem as in Alg. 1 to get another pose estimate for frame j. Performing this for all its neighbors will produce 12 additional estimates for frame j which are averaged together using quarternion average for rotations and linear average for translations. While we did not rigorously evaluate our method in comparison to state of the art [44,46,36], we have found that our pose estimation is comparable to these methods and gives temporally smooth, drift-free pose estimates, as can be observed from the accompanying videos. This process is completely automatic and the same for all video sequences.

7 Results

We evaluate the performance of our approach on a variety of videos downloaded from the Internet. Figure 7 shows example frames from four different videos (Tom Hanks, George Bush, Arnold Schwarzenegger, and Thaksin Shinawatra) downloaded from YouTube.com[2] and the corresponding per-frame 3D shape reconstructions obtained using our algorithm. On the left of Figure 7, we also present the average shapes (that are used to initialize the 3D flow estimation) for each person; these were obtained using [31]. The level of detail in the reconstructions is remarkable; the algorithm succeeds in capturing very fine details such as wrinkles and subtle expressions. Note the change in facial expression (compared to the average shape) in each frame, e.g., mouth opening, eyes close and open, wrinkles appear and disappear, detail in eye region, and so forth. The approach is robust to very large changes in pose, providing high quality results even for profile views (e.g., supplementary video of Tom Hanks). The stability of our results without any temporal smoothing other than pose filtering is evidence for the strength of the photo-based illumination subspace approach. Specifically, the illumination matching process makes the flow more accurate and thus stable. We strongly encourage the readers to watch the accompanying videos where we show per frame reconstructions for full length videos. Specifically, the lengths are: Tom Hanks: 20s (591 frames), George Bush 20s (610 frames), Arnold Schwarzenegger 24s (719frames), and Thaksin Shinawatra 20s (600 frames). Note that unlike non-rigid SfM methods [23], our reconstruction quality is independent of input video length (we can produce good results from

[2] URLs of input videos:
Hanks: `https://www.youtube.com/watch?v=emLpj38huDA`
Bush: `https://www.youtube.com/watch?v=BJbUXw87j0A`
Schwarzenegger: `https://www.youtube.com/watch?v=wH8VtPG-okI`
Shinawatra: `https://www.youtube.com/watch?v=dZdhr1WcYEM`

even a single frame). And since we estimate pose independently in each frame (and then average) by matching to an illumination-matched reference, the approach is not susceptible to drift problems that plague many tracking methods. We show long and short sequences to illustrate the quality of the reconstruction under a large variety of imaging conditions, non rigid motion, pose and lighting.

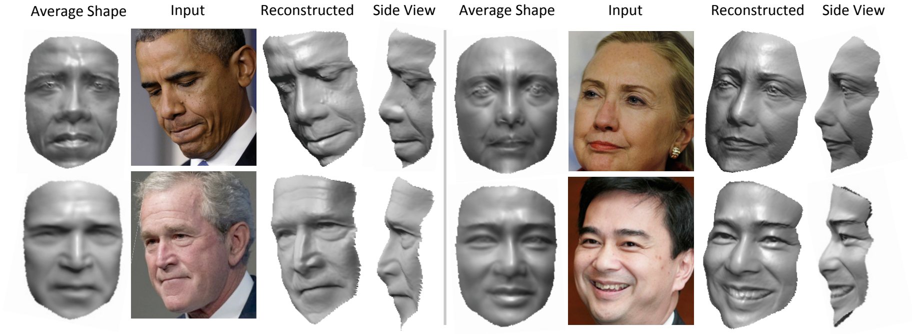

Fig. 5. Example results on still images in non-frontal views. Single view methods typically fail on such extreme poses.

In addition to handling videos, we can also estimate 3D shapes from single still images, and we show a number of results in Figure 5. We chose to show photos of faces that appear in highly non-frontal poses, these are typically the hardest cases for any state of the art single view method. The algorithm's ability to handle such extreme poses stems from our use of a person-specific template and appearance model that can be relit to match the input photo. In contrast, most other face tracking methods use generic face models which produce less reliable pose estimates, particularly for non-frontal poses.

Implementation Details. We use the Ceres solver [1] for optimization in the 3D flow estimation stage with $\alpha = 0.03$. For pose refinement we used the 2D optical flow code of [34] with the following parameters: α=0.02, ratio=0.75, nOuterFPIterations=4, nSORIterations=40. The regularization weight in shading-based refinement step is $\beta = 2$ for all videos. The running times are 35s for pose estimation (incl. 15s for temporal refinement), 70s for 3D flow, and 0.1s for shading, for a 350 × 350 frame size (face size 220 × 260 pixels).

Limitations. While we found our method to be extremely robust to a variety of lighting conditions, individuals and poses, there are a number of limitations that we would like to discuss. The first are due to the use of spherical harmonics approximation to reflectance modeling, and the Lambertian assumption. In Figure 6 we present a number of frames where (a) the person rotates the head and specularities appear on the forehead, and (b) cast shadows appear around the nose area. These are not covered by our reflectance model and therefore the algorithm will produce slightly erroneous results in the specular and shadow vertices.

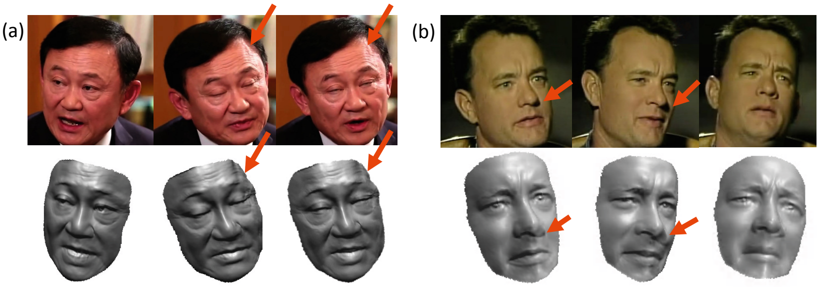

Fig. 6. Limitations of our reconstruction due to (a) specular highlight (b) cast shadows. We show a few frames from a video where the method introduces artifacts on the forehead in case of specularities or near the nose in case of cast shadows. This is due to violations of the Lambertian assumption. The full video and per frame reconstruction is shown in the accompanying video at 30fps.

Comparisons. We provide qualitative comparisons to calibrated results captured in the lab using range sensing and multi view stereo. We run our algorithm on data from [8] and compare their capture with our reconstruction in Figure 8, note the resemblance to the model captured by [8] (acquired by stereo setup) and our single view reconstruction. The base shape was acquired using the method described in Sec. 4 on 100 renders under different random lightings of frame 390 (neutral expression). The input photos are renderings of frames 80 and 340 in the dataset provided by [8]. We have also compared with Kinect Fusion [35] and present the results in Figure 9. The input to our reconstruction is a single frame; to obtain the Kinect Fusion result the person had to stay still while the depth camera captures him from a number of different viewpoints. To preserve consistency we ran our method on the direct RGB stream of kinect camera (lower in resolution than typical videos). We also compare to single view reconstruction methods, see results in the supp. material. The comparisons are qualitative since our method currently does not **guarantee** an accurate fit to ground-truth geometry due to gauge and bas-relief ambiguities. Any monocular uncalibrated approach will have this limitation, unless they assume a 3D model of the person a priori, e.g., [8,10,17,24,41] or sufficient 3D head rotation [23]. Rather, we seek to produce a reasonably accurate model (leveraging all available imagery) which optimally fits the image information in each frame. It is our future work to conduct a quantitative evaluation once time-varying 3D datasets exist with the level of detail we are attempting to capture and extend our work to handle shadows and specularities, and account for non-uniform albedo as introduced by earlier work [43].

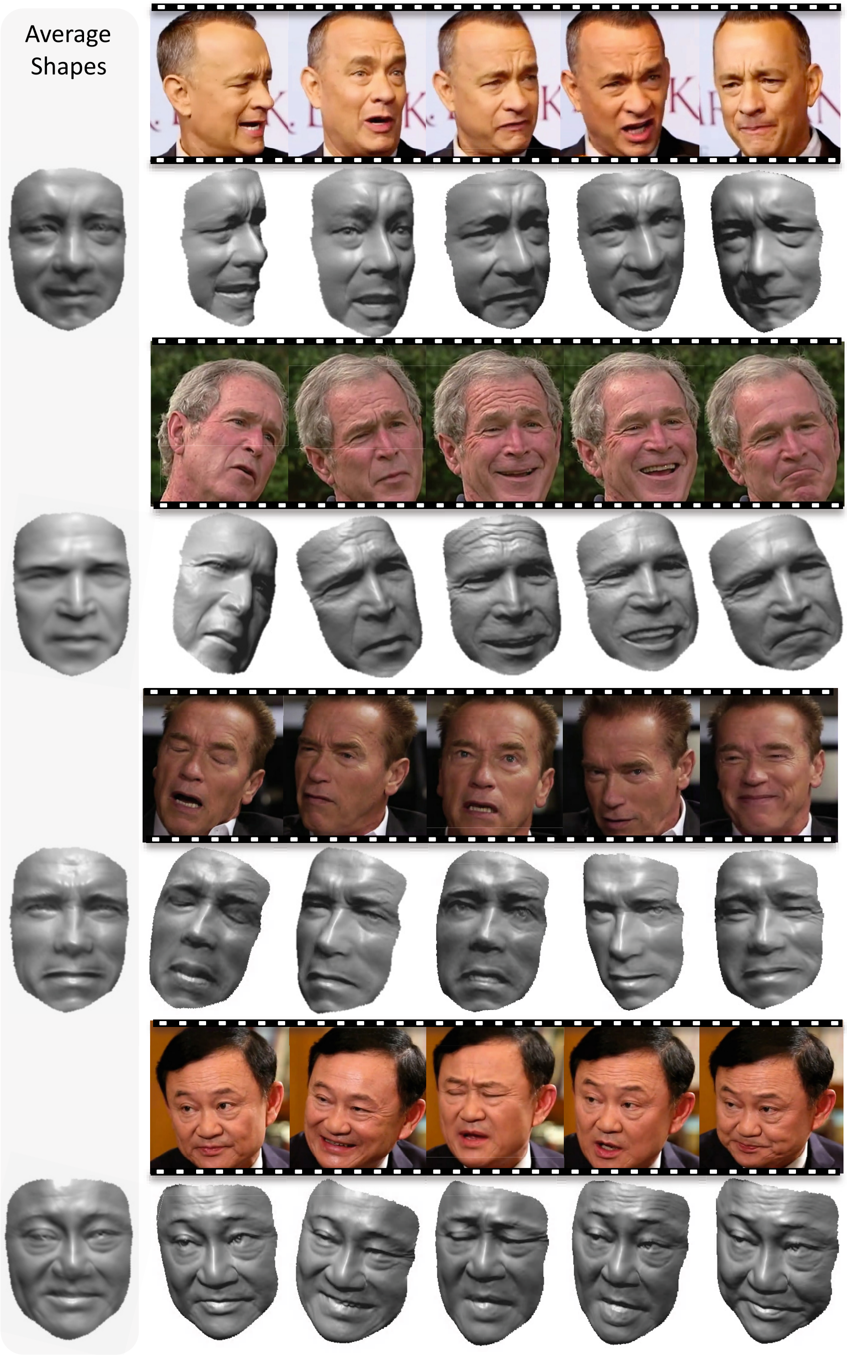

Fig. 7. Results of our reconstruction on four video sequences. Average shape per individual are presented on the left. The video reconstruction results illustrate variety in facial expressions, head pose, appearance of wrinkles, eye detail, and even partial teeth. Take a look at the full videos in the supplementary material for the full experience!

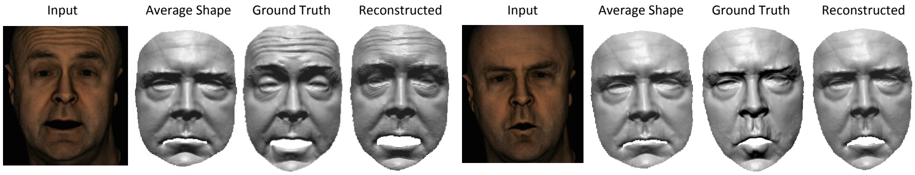

Fig. 8. Comparison to ground truth meshes [8]. Given the input photo (left) we show our reconstruction and the original shape captured by [8] for this particlar expression.

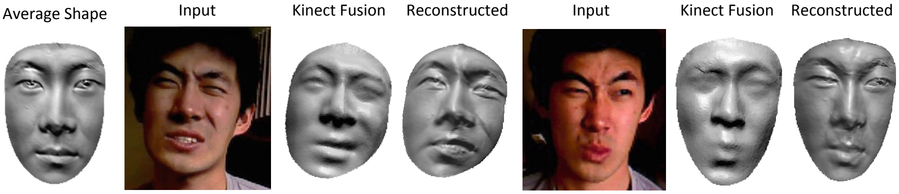

Fig. 9. Comparison to KinectFusion [35]. Two input photos, our reconstructions and results obtained using Kinect Fusion. The input photos are of lower quality than typical video sequences (taken from RGB kinect stream).

Acknowledgements. We thank Google and Intel for supporting this research.

All the examples are viewed best as videos, so we strongly encourage you to watch the supplementary video!

References

1. Agarwal, S., Mierle, K., et al.: Ceres solver, `https://code.google.com/p/ceres-solver/`
2. Alexander, O., Fyffe, G., Busch, J., Yu, X., Ichikari, R., Jones, A., Debevec, P., Jimenez, J., Danvoye, E., Antionazzi, B., et al.: Digital ira: Creating a real-time photoreal digital actor. In: ACM SIGGRAPH 2013 Posters, p. 1. ACM (2013)
3. Alexander, O., Rogers, M., Lambeth, W., Chiang, M., Debevec, P.: The digital emily project: photoreal facial modeling and animation. In: ACM SIGGRAPH 2009 Courses, p. 12. ACM (2009)
4. Basha, T., Moses, Y., Kiryati, N.: Multi-view scene flow estimation: A view centered variational approach. International Journal of Computer Vision 101(1), 6–21 (2013)
5. Basri, R., Jacobs, D., Kemelmacher, I.: Photometric stereo with general, unknown lighting. International Journal of Computer Vision 72(3), 239–257 (2007)
6. Basri, R., Jacobs, D.W.: W Jacobs. Lambertian reflectance and linear subspaces. IEEE Transactions on Pattern Analysis and Machine Intelligence 25(2), 218–233 (2003)
7. Beeler, T., Bickel, B., Beardsley, P., Sumner, B., Gross, M.: High-quality single-shot capture of facial geometry. ACM Transactions on Graphics (TOG) 29(4), 40 (2010)

8. Beeler, T., Hahn, F., Bradley, D., Bickel, B., Beardsley, P., Gotsman, C., Sumner, R.W., Gross, M.: High-quality passive facial performance capture using anchor frames. In: ACM Transactions on Graphics (TOG), vol. 30, p. 75. ACM (2011)
9. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques, pp. 187–194. ACM Press/Addison-Wesley Publishing Co. (1999)
10. Bouaziz, S., Wang, Y., Pauly, M.: Online modeling for realtime facial animation. ACM Transactions on Graphics (TOG) 32(4), 40 (2013)
11. Bradley, D., Heidrich, W., Popa, T., Sheffer, A.: High resolution passive facial performance capture. ACM Transactions on Graphics (TOG) 29(4), 41 (2010)
12. Bradski, G.: Dr. Dobb's Journal of Software Tools
13. Brand, M.: A direct method for 3d factorization of nonrigid motion observed in 2d. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 2, pp. 122–128. IEEE (2005)
14. Bregler, C., Hertzmann, A., Biermann, H.: Recovering non-rigid 3d shape from image streams. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2000, vol. 2, pp. 690–696. IEEE (2000)
15. Brox, T., Bruhn, A., Papenberg, N., Weickert, J.: High accuracy optical flow estimation based on a theory for warping. In: Pajdla, T., Matas, J. (eds.) ECCV 2004. LNCS, vol. 3024, pp. 25–36. Springer, Heidelberg (2004)
16. Alexander, O., Fyffe, G., Busch, J., Yu, X., Ichikari, R., Jones, A., Debevec, P., Jimenez, J., Danvoye, E., Antionazzi, B.: Digital ira: Creating a real-time photoreal digital actor
17. Cao, C., Weng, Y., Lin, S.: andK. Zhou. 3d shape regression for real-time facial animation. ACM TOG (Proc. SIGGRAPH) 32(4), 41 (2013)
18. Cao, C., Weng, Y., Zhou, S., Tong, Y., Zhou, K.: Facewarehouse: A 3d facial expression database for visual computing (2013)
19. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. In: Burkhardt, H., Neumann, B. (eds.) ECCV 1998. LNCS, vol. 1407, pp. 484–498. Springer, Heidelberg (1998)
20. Dai, Y., Li, H., He, M.: A simple prior-free method for non-rigid structure-from-motion factorization. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2018–2025. IEEE (2012)
21. Dale, K., Sunkavalli, K., Johnson, M.K., Vlasic, D., Matusik, W., Pfister, H.: Video face replacement. In: ACM Transactions on Graphics (TOG), vol. 30, p. 130. ACM (2011)
22. Ezzat, T., Poggio, T.: Facial analysis and synthesis using image-based models. In: Proceedings of the Second International Conference on Automatic Face and Gesture Recognition, 1996, pp. 116–121. IEEE (1996)
23. Garg, R., Roussos, A., Agapito, L.: Dense variational reconstruction of non-rigid surfaces from monocular video. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1272–1279. IEEE (2013)
24. Garrido, P., Valgaerts, L., Wu, C., Theobalt, C.: Reconstructing detailed dynamic face geometry from monocular video. ACM Transactions on Graphics (TOG) 32(6), 158 (2013)
25. Ghosh, A., Fyffe, G., Tunwattanapong, B., Busch, J., Yu, X., Debevec, P.: Multi-view face capture using polarized spherical gradient illumination. ACM Transactions on Graphics (TOG) 30(6), 129 (2011)

26. Guenter, B., Grimm, C., Wood, D., Malvar, H., Pighin, F.: Making faces. In: Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques, pp. 55–66. ACM (1998)
27. Hassner, T.: Viewing real-world faces in 3d. In: ICCV (2013)
28. Hassner, T., Basri, R.: Example based 3d reconstruction from single 2d images. In: Conference on Computer Vision and Pattern Recognition Workshop, CVPRW 2006, pp. 15–15. IEEE (2006)
29. Kemelmacher-Shlizerman, I.: Internet based morphable model. In: International Conference on Computer Vision, ICCV (2013)
30. Kemelmacher-Shlizerman, I., Basri, R.: face reconstruction from a single image using a single reference face shape. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(2), 394–405 (2011)
31. Kemelmacher-Shlizerman, I., Seitz, S.M.: Face reconstruction in the wild. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp. 1746–1753. IEEE (2011)
32. Kemelmacher-Shlizerman, I., Seitz, S.M.: Collection flow. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1792–1799. IEEE (2012)
33. Li, H., Weise, T., Pauly, M.: Example-based facial rigging. ACM Transactions on Graphics (TOG) 29(4), 32 (2010)
34. Liu, C.: Beyond Pixels: Exploring New Representations and Applications for Motion Analysis. PhD thesis. MIT (2009)
35. Newcombe, R.A., Davison, A.J., Izadi, S., Kohli, P., Hilliges, O., Shotton, J., Molyneaux, D., Hodges, S., Kim, D., Fitzgibbon, A.: Kinectfusion: Real-time dense surface mapping and tracking. In: 10th IEEE International Symposium on Mixed and Augmented Reality, ISMAR 2011, pp. 127–136. IEEE (2011)
36. Saragih, J.M., Lucey, S., Cohn, J.F.: Face alignment through subspace constrained mean-shifts. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 1034–1041. IEEE (2009)
37. Valgaerts, L., Bruhn, A., Zimmer, H., Weickert, J., Stoll, C., Theobalt, C.: Joint estimation of motion, structure and geometry from stereo sequences. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 568–581. Springer, Heidelberg (2010)
38. Valgaerts, L., Wu, C., Bruhn, A., Seidel, H.-P., Theobalt, C.: Lightweight binocular facial performance capture under uncontrolled lighting. ACM Trans. Graph. 31(6), 187 (2012)
39. Vedula, S., Baker, S., Rander, P., Collins, R., Kanade, T.: Three-dimensional scene flow. In: The Proceedings of the Seventh IEEE International Conference on Computer Vision 1999, vol. 2, pp. 722–729. IEEE (1999)
40. Vlasic, D., Brand, M., Pfister, H., Popović, J.: Face transfer with multilinear models. In: ACM Transactions on Graphics (TOG), vol. 24, pp. 426–433. ACM (2005)
41. Weise, T., Bouaziz, S., Li, H., Pauly, M.: Realtime performance-based facial animation. ACM Transactions on Graphics (TOG) 30(4), 77 (2011)
42. Wu, C., Stoll, C., Valgaerts, L., Theobalt, C.: On-set performance capture of multiple actors with a stereo camera. ACM Transactions on Graphics (TOG) 32(6) (2013)
43. Wu, C., Varanasi, K., Liu, Y., Seidel, H.-P., Theobalt, C.: Shading-based dynamic shape refinement from multi-view video under general illumination. In: International Conference on Computer Vision ICCV (2011)

44. Xiong, X., De la Torre, F.: Supervised descent method and its applications to face alignment. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 532–539. IEEE (2013)
45. Zhang, L., Snavely, N., Curless, B., Seitz, S.M.: Spacetime faces: High-resolution capture for˜ modeling and animation. In: Data-Driven 3D Facial Animation, pp. 248–276. Springer, Heidelberg (2007)
46. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2879–2886. IEEE (2012)