Face Reconstruction in the Wild

Ira Kemelmacher-Shlizerman
University of Washington
kemelmi@cs.washington.edu

Steven M. Seitz
University of Washington and Google Inc.
seitz@cs.washington.edu

Abstract

We address the problem of reconstructing 3D face models from large unstructured photo collections, e.g., obtained by Google image search or from personal photo collections in iPhoto. This problem is extremely challenging due to the high degree of variability in pose, illumination, facial expression, non-rigid changes in face shape and reflectance over time and occlusions. In light of this extreme variability, no single reconstruction can be consistent with all of the images. Instead, we define as the goal of reconstruction to recover a model that is locally consistent *with the image set. I.e., each local region of the model is consistent with a large set of photos, resulting in a model that captures the dominant trends in the input data for different parts of the face. Our approach leverages multi-image shading, but unlike traditional photometric stereo approaches, allows for changes in viewpoint and shape. We optimize over pose, shape, and lighting in an iterative approach that seeks to minimize the rank of the transformed images. This approach produces high quality shape models for a wide range of celebrities from photos available on the Internet.*

1. Introduction

An Internet image search for a celebrity yields thousands of photos. Similarly, our personal photo collections contain thousands of photos of faces. In this paper we consider the problem of computing 3D face reconstructions from such collections. This problem is extraordinarily challenging due to the following factors (some of which are illustrated in Figure 1):

- **expression** often varies significantly (e.g., smile, neutral, surprise, etc.)
- **exposure**, color balance, resolution, zoom levels, and other imaging parameters are different from one photo to the next
- **illumination** is unknown and varying; some photos are taken indoors, other outdoors, some with flash, some not
- **pose** typically varies between frontal and profile

Figure 1. (a) A few example photos (out of hundreds) we used for the reconstruction of George W. Bush and our reconstruction. (b) Photos of Bill Clinton and our reconstruction. Note the significant variability in facial expression, lighting and pose.

- **aging** results in very significant changes in face shape over time
- **facial hair**, make-up, glasses may come and go, changing reflectance and geometry

Given such extreme variations, how can we even *define* the "shape" of a person's face? Indeed, there is no single shape, but rather a large space of different face shapes for each individual. Nonetheless, we seek to obtain a *canonical* shape that captures that person's characteristics as well as possible. The two major contributions of this paper are to define a notion of a canonical shape, and to introduce a practical 3D reconstruction algorithm that produces high quality results for a broad range of Internet and personal photo collections.

One natural notion of a canonical shape is the mean over the image set. The disadvantage of a mean, however, is that it produces an over-smoothed result that wipes away details of the model. A second possibility is a mode, i.e., the most commonly occurring shape in the dataset. While this approach works better than a mean, we find that it is too restrictive, as all parts of the face must be in the same configuration (very few photos may match). Our solution is to solve for a shape that is *locally similar* to as many photos

as possible. Intuitively, the idea is to capture the dominant mouth shape in the collection, the dominant eyes, and so forth. In practice we do the decomposition on a *point-by-point* basis, defining a different image set for each point on the model. For each point, the goal is to select an image set that captures that point in approximately the same rigid configuration, but under different illumination conditions. We then exploit the observed variations in shading to solve for the local surface shape and then integrate out a 3D model for the whole face. This analysis occurs in the context of a more general approach that allows for pose variations by solving for camera pose and warping images to frontal. The recovered shape is then fed back into the pose/warping process, and the process is repeated until convergence.

Despite a large literature of prior work on face analysis and reconstruction, very few approaches work on the unstructured face collections proposed here (Google, iPhoto, etc.). The last few years have seen tremendous progress in rigid 3D scene modeling from the Internet and other unstructured collections [12, 1] using structure-from-motion and stereo techniques. These techniques, however, are not applicable for face reconstruction due to the non-rigid nature of faces and lack of dense correspondence and camera calibration information. Indeed, most state of the art techniques for high quality face reconstruction require a subject to come to a lab to be scanned with special equipment (e.g., laser, stereo, lighting rigs, etc.) [15, 5, 8, 13, 3]. In contrast, we propose to exploit multi-image shading cues for face reconstruction–leveraging the lighting variation in large collections.

One class of techniques that *can* handle very general face images are *single-view methods*. For example, Blanz and Vetter [6] showed the potential for very high quality 3D face modeling from a single image more than a decade ago. The idea is to express novel faces as a linear combination of a database of 3D laser-scanned faces. This approach works extremely well when the target is roughly within the linear span of the database (in their case, 200 young adults), but is not well suited for capturing facial shape with facial expressions and subtle details that vary from one individual to the next. It also requires manual initialization—a fully automatic solution has proven elusive. Kemelmacher and Basri [11] also enable single-image reconstruction using a shape-from-shading approach that requires only a single template face as a prior. This approach can yield good looking results, but the geometry varies significantly depending on which image and template is used. This approach also requires manual work to register the image with the template. Because the single-view problem is ill-posed, these methods depend heavily on prior models. It is therefore interesting to explore how far we can go with the images alone (without strong prior assumptions on shape).

Our use of multi-image shading cues builds on classic results in photometric stereo [14]. But while prior work in photometric stereo required rigid geometry, fixed reflectance over time, and fixed camera pose, we relax all of these assumptions in order to operate on general unstructured image collections.

In summary, we offer two primary contributions over prior work.

1. We pose the reconstruction problem of recovering a *locally consistent* shape that models each part of the face using a different subset of images.

2. We introduce the first fully-automatic face reconstruction technique capable of operating on general face photo sets, such as those found via Internet search or in personal photo collections like iPhoto.

2. The Algorithm

In this section we assume that we are given a large collection of photos of a single person under different illuminations. In addition, the facial pose, expression, and geometry may vary from photo to photo. In this section, we describe the steps of our 3D reconstruction approach which involves detecting fiducials, solving for pose, warping to frontal pose, recovering an initial shape and lighting based on photometric stereo, and refining the model using local view selection. We iterate this entire procedure until convergence.

Our image formation model assumes weak perspective projection, relatively distant illumination, and Lambertian reflectance.

2.1. Pose normalization

To account for variations in face orientation over the image set, we warp each image to a canonical, frontal pose. To this end, we first estimate pose by detecting fiducial points on the face (see Section 3 for details), and use the positions of these fiducials from a template 3D face to recover 3D rotation, translation, and scale for each photo.

The relation between points q on the image and points on the template Q is given by

$$q = sRQ + t. \tag{1}$$

To recover s, R, and t, we first subtract the centroid from both point sets to get $p = q - \overline{q}$ and $P = Q - \overline{Q}$, then estimate a 2×3 linear transformation $A = pP^T(PP^T)^{-1}$ and translation $t = \overline{q} - A\overline{Q}$. To recover an estimate of the rotation and scale we let the third row of A be the cross product between the first two rows and by taking its SVD, $A' = UDV^T$, we estimate the closest rotation in terms of Frobenius norm $R = UV^T$. Two of the singular values of A' are identical, and this is our estimate of scale. We then

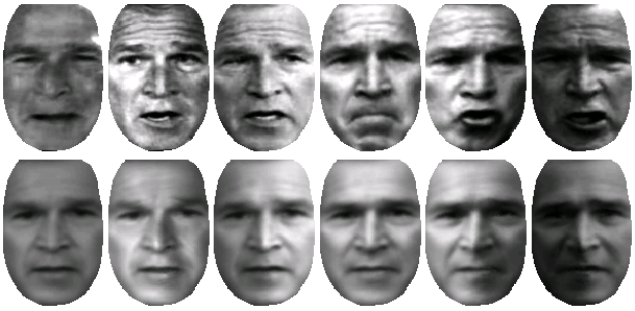

Figure 2. Expression normalization by low-rank approximation. The first row shows the warped images, the 2nd row shows the the low rank approximated images. Note how the lighting is mostly preserved, but the facial expression is normalized.

estimate the yaw, pitch and roll angles from the rotation matrix. Given the estimated pose we transform the template to the orientation of the face in the image, the image is back-projected onto the shape, and then a frontal view of the face is rendered. This results in a collection of faces where every face is in approximately frontal position as can be seen in Fig. 2 (1st row).

2.2. Initial lighting and shape estimation

In this section we assume that all images in the collection are warped to a canonical position. We begin by estimating per-image illumination, using an uncalibrated photometric stereo approach (e.g., [4, 2]). From the warped images we construct an $n \times p$ matrix M, where n denotes the number of images and p is the number of pixels in each image. Each image is represented by a row in M. We factorize M using Singular Value Decomposition, $M = UDV^T$ and take the rank-4 approximation to get $M = LS$ where $L = U\sqrt{D}$ is $n \times 4$ and $S = \sqrt{D}V^T$ is $4 \times p$. In the absence of ambiguities, L should contain the low order coefficients of the lighting and S the albedo, and components of the surface normals at each point on the surface scaled by the albedo (or the first four principal components of the shape). In general, however there is a 4×4 ambiguity since M can be represented also by $M = LA^{-1}AS$ which needs to be resolved to make the shape integrable. This however is not needed for our algorithm until we will want to integrate the surface normals to obtain the surface. We will discuss that ambiguity at the end of section.

In Figure 2 we show several warped photos and their low rank approximations. We can see that the facial expression is normalized (all faces appear in neutral expression), the lighting in each image is roughly correct, and the appearance of the person is clearly recognizable. Hence, the classical photometric stereo procedure already has a *built-in normalization for facial expression*, due to the rank-4 approximation. We shall see, however, that fine details are lost in the resulting reconstruction, and we propose techniques to recover them in the next section.

2.3. Local surface normal estimation

Applying classical photometric stereo to the entire image collection produces an over-smoothed shape; this is not surprising, as we are effectively averaging over all expressions. We achieve better results by using different images to reconstruct different parts of the face. Intuitively, we might use a set of images in which the mouth is fixed in a smiling pose to model the mouth, and a different set in which the eyes are open to reconstruct the eye region. In practice, the image selection is automated, and operates on a point-by-point basis.

The local selection of images works as follows. For each point on the face we first calculate how well the images fit the initial shape estimate, i.e., we calculate the distance $|M_j - LS_j|^2$ where M_j is a $n \times 1$ vector representing the intensities of a pixel in all images (column of M), S_j is 4×1 and L is $n \times 4$. We then normalize the distance and choose a subset of images for which the distance is less than a threshold, making sure that the number of images is larger than 4 and that the condition number of $L_{k\times 4}$ is not high (k represents the number of chosen images) to prevent degenerate lighting conditions. The resulting set of images is then used to recover S_j by minimizing the following functional

$$\min_{S_j} ||M_{k\times 1} - L_{k\times 4}S_j|| + S_j^T G S_j. \qquad (2)$$

The first term represents the lighting consistency relation and the second term acts as a Tikhonov regularization term which avoids poor conditioning. The matrix G in the regularization term is chosen to be $G = \text{diag}(-1, 1, 1, 1)$. This choice is motivated by the fact that each column S_j should have the first four principle components for this point which are $(\rho, \rho n_x, \rho n_y, \rho n_z)$ where ρ is the albedo and (n_x, n_y, n_z) is the surface normal. Since the normal should have unit length the following relation should hold $S_j^T G S_j = 0$; this is exactly our regularization term. The solution to this minimization problem is

$$S_j = (L_{k\times 4}^T L_{k\times 4} + G)^{-1} L^T M_{k\times 1}. \qquad (3)$$

This solution provides the surface normals of the face.

Figure 3 shows examples of photos used to reconstruct different points on the shape. It is interesting to see that the photos chosen for each of the points have similar geometry at that point across the images, e.g., the images for the mouth point (bottom right) all have roughly neutral facial expression. This locally adaptive approach has a number of advantages. Note that areas that are consistent across many photos (approximately rigid areas of the face) will select a very large subset of photos, leading to a very robust shape in these areas. For areas that are non-rigid, the algorithm tends to select a subset of photos which mutually agree, leading to

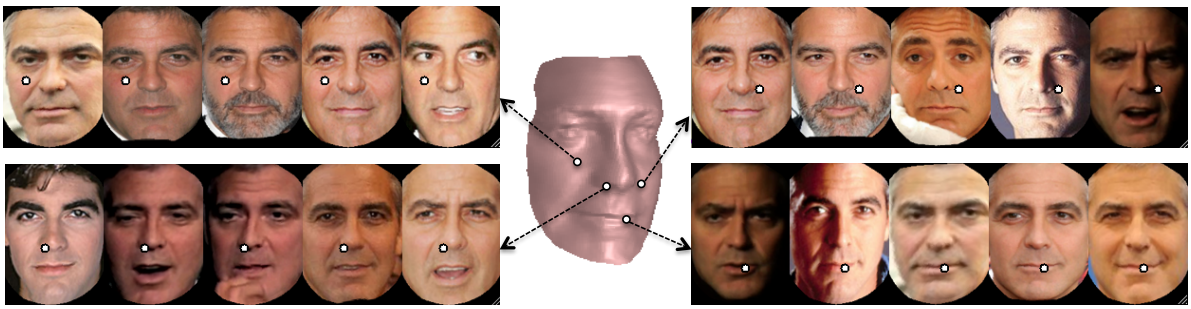

Figure 3. The shape of George Clooney, and example of photos that were used for reconstructions of the different points.

a consistent shape (rather than a ghosted average over many shapes) in many cases. Of course this works only up to a point–when there is no clear winner, ghosting effects can still occur (see the results section for examples). Another advantage of the local approach is that it helps to prevent outliers from biasing the results, e.g., sun glasses, occlusions, misalignments, wrong identity, etc.

2.4. Ambiguity recovery

It was shown [2] that the ambiguity between the lighting L matrix and the S matrix can be resolved up to the generalized bas-relief transformation by applying integrability constraints. From our experiments, this procedure is unstable and does not perform well when the image set contains expression and shape variations. Instead, we use the template face to resolve the ambiguity. In particular we take the estimated shape matrix S, re-estimate the lighting by $\min_L ||M-LS||^2$, and choose the images which are fit well by this low rank approximation: $||M - LS|| < \epsilon$. We then solve for $\min_A ||S_t - AS||^2$ where A is the 4×4 ambiguity and S_t includes the albedo and surface normal components of the template shape. This results in recovery of the lighting coefficients LA^{-1} and the surface normals AS.

2.5. Integration

Given the shape matrix, it is straightforward to recover the surface of the object by integrating the surface normals (e.g. as in [4]). The albedo is the first row of S and, by normalizing its other three rows, the surface normals are recovered. To reconstruct the depth map $z(x, y)$, the surface normal at each point is represented as the gradient of the depth map $(n_1, n_2, n_3)^T = (z_x, z_y, -1)^T / \sqrt{z_x^2 + z_y^2 + 1}$ which results in $z_x = -n_1/n_3$ and $z_y = -n_2/n_3$. In combination with the forward difference representation of the gradient, this can be formulated as a linear system of equations with $z(x, y)$ as its unknowns and solved with least squares.

This procedure results in a surface reconstructed up to the generalized bas-relief transformation, i.e., $Z_{\text{true}} = aX + bY + cZ$ where Z is the output of integrability. We recover these 3 unknowns using the template shape to provide the shear and scale of the depth map.

2.6. Iterations of the algorithm

Once the surface is reconstructed we re-estimate the 3D pose of the face using the reconstructed shape as in section 2.1 and re-warp the images using that shape. Note that since the reconstruction is aligned with the template, we can determine where the fiducials are on the reconstruction, by assuming their 2D positions are fixed in the canonical pose (their 2D positions are fixed, but their depths may vary between the template and the reconstruction). The re-warped images are then used as an input to the local selection of image subsets for surface normals reconstruction. We have observed that the initial pose estimation with the template model is typically quite accurate, with artifacts appearing in the warped images mainly around the nose. In the second iteration, after rewarping with the reconstructed shape, those artifacts disappear. The stopping criteria for this procedure is when the rank of the matrix of warped images does not change significantly. More precisely, we measured rank as the sum of the fifth and higher singular values.

2.7. Summary of the algorithm

The full algorithm is as follows:

1. find fiducials in each image, initialize shape to template
2. solve for pose, warp each image to frontal using current shape estimate
3. solve for initial shape and lighting using photometric stereo
4. apply local view selection to re-estimate normals
5. resolve GBR and integrate a shape
6. until convergence, goto Step 2.

3. Experiments

In this section we describe our experiments. We first describe the automatic preprocessing framework of the images in the set, and evaluation with synthetic data. We then show experiments with real unconstrained data: 1) taken from the Labeled Faces in Wild (LFW) [10] database, 2) collections

Figure 4. Synthetic experiment: (a) Example images rendered using ground truth models. (b) mean and variance of 1) all images in neutral expression but not pose corrected, 2) in neutral expression but pose corrected and 3) all 7 expressions and pose corrected. (c) Plot of the distance between reconstruction and neutral ground-truth, as a function of percent of non neutral photos in the image set.

of photos of celebrities that we collected from Google image search, and 3) on a personal photo collection. Finally we compare our results with standard photometric stereo and a single view approach.

3.1. Preprocessing of the images

To obtain a collection of photos of the same person we first scanned all the images using Picasa, and extracted the photos that were recognized by Picasa as the person. Then for each photo in the dataset we first apply a face detector [7] followed by a fiducial points detector [9] that finds the left and right corners of each eye, the two nostrils, the tip of the nose and the left and right corners of the mouth. We eliminate photos that have low detection confidence measure (less than 0.5), and photos for which the fiducial points detector confidence is less than -3. We also gamma correct the images to account for possible non-linear camera response. The initial template shape that we use in our method is the neutral face model from the space-time faces dataset [15].

3.2. Evaluation with synthetic data

To evaluate our method we rendered a 3D dataset that included 7 facial expressions, 5 poses (frontal, azimuth $10°, 20°$, elevation $-10°, 10°$) and 21 lighting directions from the shapes in [15]. Example images and expressions can be seen in Fig. 4 (a). All images were corrected for pose as described in Sec. 2.1, and the ambiguity was estimated using the neutral shape. Fig 4 (b) shows mean and variance of set of images where the face appears in 1) neutral expression and without pose correction, 2) neutral expression and pose corrected and 3) all 7 expressions and pose corrected. We can see that when the pose is not corrected there are significant artifacts in the nose area, and when all expressions are averaged together artifacts appear around the nose, mouth and eyebrows. We also conducted an experiment to measure the impact of adding non-frontal photos to reconstruction accuracy, and found that, using a ground-truth template model, the use of non-frontal images had only a modest detrimental impact: a reconstruction error of $0.76 \pm 0.48\%$ from frontal images vs. $0.93 \pm 0.68\%$ for frontal plus non-frontal. The slight decrease in performance is likely due to occlusions (not modeled by the warping process) and resampling artifacts.

Fig 4 (c) shows results of an experiment of adding different amounts of random photos with non neutral expression to a set of photos with neutral expression. We plot the mean and standard deviation of the distance between the reconstruction and the neutral ground-truth model vs. the percent of non-neutral "outlier" photos in the set. The distance is measured by $100|z_{rec} - z_{gt}|/z_{gt}$ per point on the surface. We can see that a model very close to neutral is reconstructed up to a 25% outlier rate, after which the local view selection becomes more influenced by non-neutral photos. The accuracy of the reconstruction from only neutral and frontal photos and neutral and pose corrected photos does not vary much–$0.76 \pm 0.48\%$ vs. $0.93 \pm 0.68\%$.

3.3. Real images

We ran our method on the four people from the LFW dataset with the largest number of photos: George W. Bush (530 photos), Colin Powell (236), Tony Blair (144), and Donald Rumsfeld (121). The resolution of the photos in LFW is quite small—the size of each photo is 250×250 with the face dimensions around 150×200. Figure 6 shows our reconstructions. The reconstructions look convincing, note especially the noses of George W. Bush and Colin Powell. Donald Rumsfeld and Colin Powell appear with eye glasses in all photos, and we can see the round areas around the eyes in the reconstructed shape. For Tony Blair the average shape turned out to have a smile.

We also experimented with larger sets by collecting images from Google image search. We searched for names of celebrities (Bill Clinton, George Clooney, Kevin Spacey, Tom Hanks) and downloaded around 850 photos per celebrity. For Clooney and Spacey we also downloaded

a video interview from YouTube and sampled one frame per second to add around 300 more photos to the collection. The number of photos used per person is stated in Table 1. Figure 7 shows the reconstructed shapes. Figure 8 shows more renderings of the shapes compared to photos of the person in similar position. Note how well details of each individual were reconstructed, e.g., the noses and the areas around the eyes. Figure 10 shows the improvement in shape reconstruction when iterating the method. Initially the input images are warped using the template model (0th iteration) and then re-warped using the estimated shape. We observe major improvements in shape reconstruction in iterations #1 where the rough shape is reconstructed and #2 where the nose area gets sharper, and some minor improvements in the subsequent iterations.

We have also run our algorithm on personal photo collections, Figure 6 (right) shows the reconstruction result (we omit more only for lack of space).

3.4. Comparison of the local approach to global filtering

To evaluate the impact of local view selection, we now compare to alternative filtering methods that operate "globally", on an entire image at a time. Figure 5 (a) shows the reconstruction of George Clooney's face shape using all the photos in the dataset (i.e., no filtering). This result is clearly oversmoothed. A simple filtering approach is to choose only "frontal" views; we consider only faces that appear within the range of ± 5 degrees from the frontal position (for each of the estimated yaw and pitch angles) which results in reconstruction (b). And we can further filter the set by taking into account only photos that are fit well by the low rank approximation (for the entire image) which results in reconstruction (c). Table 1 presents the number of images that remain for each of these approaches. While these simple "global" filtering steps provide some improvements, the resulting model is not very realistic. In particular, note how the nose is far too sharp and the upper lip protrudes unrealistically in (a) - (c), compared to the more natural reconstruction using the local approach (see Fig. 8) to see how (d) compares to real photos.

3.5. Comparison to single view

Finally we have also compared our reconstruction to a leading single-view modeling technique by [11]. Figure 9 shows two photos of Clinton with single view reconstructions. We also show the average of all single view reconstructions from every photo in the dataset. While the reconstructions are reasonable, different input photos lead to different shapes. Note that the average shape does not look very much like Clinton. The advantage of a multi-view approach is that we can integrate over many images to yield a single consistent model.

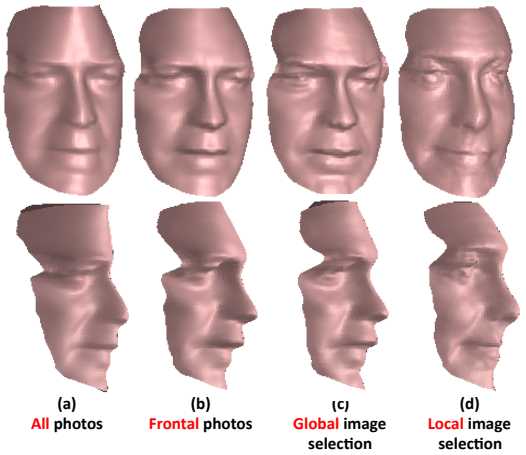

Figure 5. Comparison between reconstructions of George Clooney: (a) reconstructed shape from all the aligned and warped photos (b) from subset of photos where the face appears in frontal pose (c) after global image selection (d) after per pixel (local) image selection. See the number of images used in Table 1.

	Clooney	Clinton	Spacey	Hanks
Total	643	373	678	698
Frontal	183	125	140	183
Global	79	52	93	163
Local	72±21	52±16	68±18	85±23

Table 1. The total number of images, number of images remained after filtering by pose (frontal), after global image selection process and after local image selection process. Each column represents a different person. For the local image selection process we present the mean and standard deviation over the pixels.

Figure 8. Rendering of the shapes in different viewpoints and photos (not used in reconstruction) of the person in similar viewpoints.

4. Discussion

We presented the first method capable of generating 3D face models automatically from unstructured photo collections, overcoming vast variability in facial pose, expression, lighting, exposure, facial hair, and aging effects. Instead

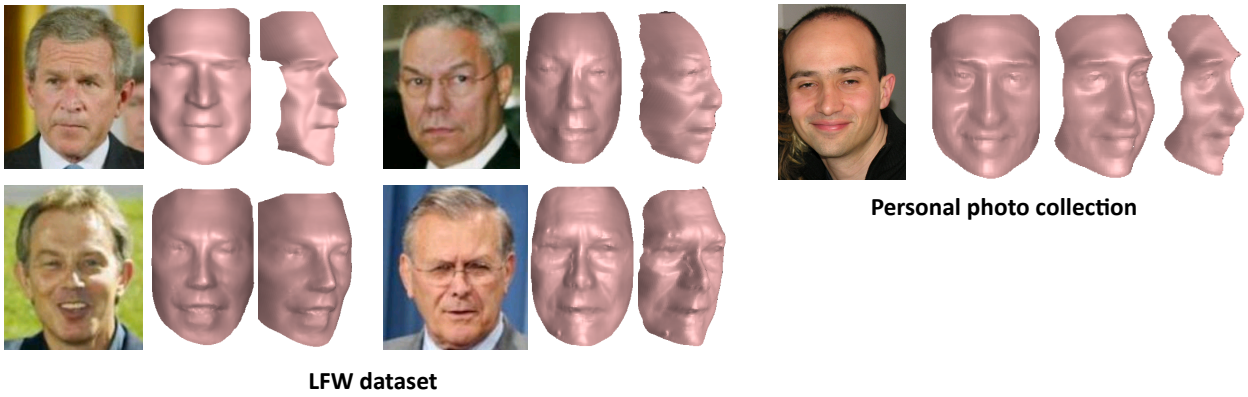

Figure 6. Results on people with the largest number of photos from the LFW dataset. In spite of the relatively small number of photos per person and low resolution of the photos the reconstructions look convincing–note the reconstruction of the nose of George W. Bush and Colin Powell and the smile of Tony Blair. The glasses on Powell and Rumsfeld are baked into the reconstruction. On the right we show reconstruction result of a face from a personal photo collection.

Figure 7. Results on photos collected from Google search. For each person (each row) we present the average image created by averaging images that were used for reconstruction, the image that was used to texture map the shape and the texture mapped shape, and finally three viewpoints of the reconstructed shape. Observe how well the profile (last column, not used for the reconstruction) fits the model.

of explicitly modeling all of these effects, we take the approach of normalizing the images to account for rotation, and selecting a subset of the images, on a pixel-by-pixel basis, in which the geometry is locally similar. A key point in

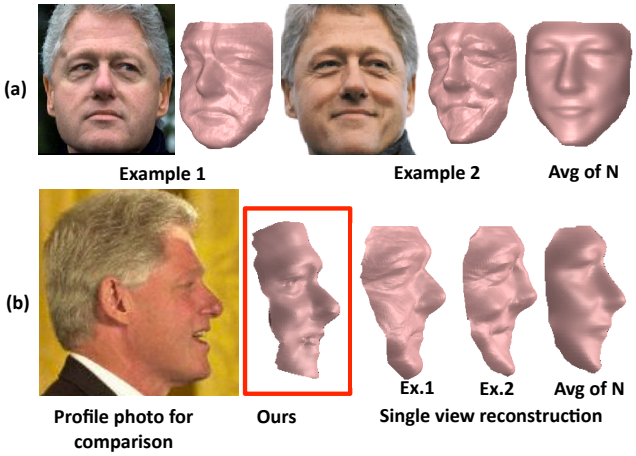

Figure 9. Comparison to single view reconstructions provided by [11]. (a) Two photos of Bill Clinton, the shape reconstruction using the single view method (from each of the images separately), and average of single view reconstructions from all the images in the set. (b) Photo of Clinton's profile, and profile renderings of our reconstruction, two single view reconstructions and average of all the single view reconstructions.

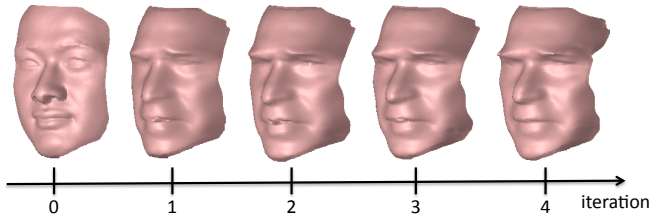

Figure 10. The shapes reconstructed in each iteration of the method. The template shape is shown as the 0th iteration.

this paper is that by leveraging large photo collections we can achieve a high quality and *consistent* reconstruction for all the images in a dataset.

There are a number of potential improvements and challenges going forward. Our reconstructions are not metrically correct, as we rely on a template to resolve the scale ambiguities. It would be interesting to exploit other cues such as knowledge of pose and profile contours to resolve this ambiguity. Note that our approach does not exploit a face-specific model or prior in the reconstruction process (a face template is only used to align faces in preparation for photometric stereo, and as a post process to resolve the global scale). While we wanted to see how far we could go *without* a face prior, there is a wealth of prior data and knowledge of face geometry which could be leveraged in future systems to fill in missing details and remove noise and artifacts.

Acknowledgements

This work was supported in part by National Science Foundation grant IIS-0811878, the University of Washington Animation Research Labs, Adobe, Google, and Microsoft.

References

[1] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski. Building rome in a day. *International Conference on Computer Vision*, 2009. 2

[2] R. E. A.L. Yuille, D. Snow and P. Belhumeur. Determining generative models for objects under varying illumination: Shape and albedo from multiple images using svd and integrability. *International Journal on Computer Vision*, 35:203–222, 1999. 3, 4

[3] O. Alexander, M. Rogers, W. Lambeth, J.-Y. Chiang, W.-C. Ma, C.-C. Wang, and P. Debevec. The digital emily project: Achieving a photorealistic digital actor. *IEEE Computer Graphics and Applications*, 30:20–31, 2010. 2

[4] R. Basri, D. Jacobs, and I. Kemelmacher. Photometric stereo with general, unknown lighting. *International Journal of Computer Vision*, pages 239–257, 2007. 3, 4

[5] T. Beeler, B. Bickel, P. Beardsley, B. Sumner, and M. Gross. High-quality single-shot capture of facial geometry. *ACM Trans. on Graphics (Proc. SIGGRAPH)*, 29(3), 2010. 2

[6] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *SIGGRAPH*, pages 187–194, 1999. 2

[7] L. Bourdev and J. Brandt. Robust object detection via soft cascade. *Conference on Computer Vision and Pattern Recognition*, 2005. 5

[8] D. Bradley, W. Heidrich, T. Popa, and A. Sheffer. High resolution passive facial performance capture. *ACM Trans. on Graphics (Proc. SIGGRAPH)*, 29(3), 2010. 2

[9] M. Everingham, J. Sivic, and A. Zisserman. "Hello! My name is... Buffy" – automatic naming of characters in TV video. In *BMVC*, 2006. 5

[10] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, 2007. 4

[11] I. Kemelmacher-Shlizerman and R. Basri. 3d face reconstruction from a single image using a single reference face shape. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010. 2, 6, 8

[12] N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism: Exploring image collections in 3d. *SIGGRAPH*, 2006. 2

[13] T. Weise, B. Leibe, and L. V. Gool. Fast 3d scanning with automatic motion compensation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'07)*, June 2007. 2

[14] R. Woodham. Photometric method for determining surface orientation from multiple images. *Optical Engineerings*, 19:139–144, 1980. 2

[15] L. Zhang, N. Snavely, B. Curless, and S. M. Seitz. Spacetime faces: high resolution capture for modeling and animation. In *SIGGRAPH*, pages 548–558, 2004. 2, 5