Internet-based Morphable Model

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

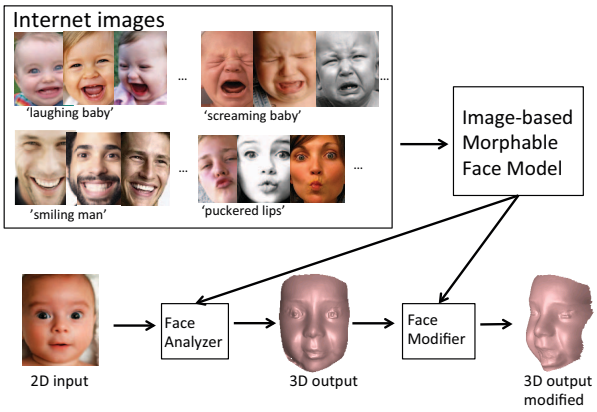

Figure 1. Overview of the method. We construct a morphable model directly from Internet photos, the model is then used for single view reconstruction from any new input image (Face Analyzer) and further for shape modification (Face Modifier), e.g., from neutral to smile in 3D.

Abstract

In this paper we present a new concept of building a morphable model directly from photos on the Internet. Morphable models have shown very impressive results more than a decade ago, and could potentially have a huge impact on all aspects of face modeling and recognition. One of the challenges, however, is to capture and register 3D laser scans of large number of people and facial expressions. Nowadays, there are enormous amounts of face photos on the Internet, large portion of which has semantic labels. We propose a framework to build a morphable model directly from photos, the framework includes dense registration of Internet photos, as well as, new single view shape reconstruction and modification algorithms.

1. Introduction

In their landmark 1999 paper [13], Blanz and Vetter [13] introduced *morphable models*, a powerful tool for modeling 3D face shape and deformations that can be fit from a single photo. Their key idea was to constrain face reconstruction to lie within the linear span of 200 previously captured and aligned 3D face scans. This approach dramatically reduces the degrees of freedom of the face reconstruction problem, and enabled extremely impressive results. Beyond reconstruction, the morphable model framework provides two key benefits: first, a point-to-point correspondence between the reconstruction and all other models, enabling "morphing", and second, modeling underlying *transformations* between types of faces (male to female, smile to frown, thin to flesh, etc.).

Morphable models could potentially have an unprecedented impact on face recognition, tracking and detection algorithms, e.g., Face Recognition Grand Challenge (FRGC) concluded [1] that, if available, 3D models dramatically increase recognition performance due to their invariance to lighting, viewpoint, and occlusion. More than a decade later, however, morphable models have yet to achieve their initial promise; while we've seen face detection and recognition enjoy widespread deployment into consumer cameras and photo sharing technology, morphable models have yet to achieve similar impact. Still, face detection and recognition methods operate by training on a very large number of photos to achieve robust performance and often fail, e.g., for non-frontal views, extreme lighting, children, unusual expressions, or other cases that fall outside of the training set. In the research community, the number of follow-on research papers on morphable models has declined in recent years. We believe the key reasons are due to three primary limitations:

1. Range: reconstructions must fall in the linear span of the database, and existing 3D scan databases are too limited to capture the full range of human expression, ethnicity, aging, and other factors that affect shape.

2. Scale: while it's relatively simple to train a face detector on 10,000 examples, acquiring, cleaning, and aligning the 3D models needed for morphable models is a painstaking and cumbersome task. [1].

[1]Future improvements to Kinect and similar scanning systems, may help though.

3. Automation: the approach requires significant manual work both to initialize the reconstruction and to create the database.

In this paper we'd like to address these limitations and introduce a new framework for computing a morphable model. Rather than limiting ourselves to laser scans/shapes captured in a laboratory we propose to leverage the Internet. The vast amounts of photos of people already on the Internet can potentially capture many of the degrees of freedom of the human face. For example, a Google image search for "smiling babies" yields 100 million hits, and similarly large numbers for "frowning babies," "pouting babies," etc. In contrast, consider the logistical challenges of trying to acquire 3D scans of many babies in different expressions. Different search terms yield face photos of any desired age, country, ethnicity, etc. We present a new completely automatic face modeling approach that computes a 3D face basis *directly* from a large photo collection of photos from the Internet (rather than first acquiring a database of 3D face scans and deriving a basis), and consequently enables reconstruction from a single view, and morphing of the reconstruction, e.g., to different facial expressions.

Given a collection of photos our method automatically computes pixel-wise correspondence from every photo in the collection to a single reference (which is also computed by the method). This enables putting in correspondence all the photos in the collection. The key idea of the paper is that once the photos are aligned it is possible to derive a 3D shape basis directly from the collection, and further to estimate 3D shape from any single image and modify its shape, e.g., to different facial expressions. In particular, we show that the matrix of aligned intensities is a rank $4K$ matrix under the Lambertian reflectance model assumption and can be factored into K 3D basis shapes, as well as lighting and shape coefficients per image using SVD. We demonstrate the effectiveness of this method on challenging images taken "in the wild", including images of human faces with varying facial expression taken under arbitrary lighting and pose, and show shape reconstructions and modifications that are produced completely automatically.

1.1. Overview

Section 2 summarizes related approaches. Section 3 describes how we align collected Internet photos and derive a shape basis, we next describe the factorization method that allows reconstructing shape from a single image in Section 4 and modifying the shape to perform different facial expressions from a *single* image in Section 5. Experimental evaluations are presented in Section 6 and conclusions in Section 7.

2. Related Work

Despite a large literature on face modeling, still it is very challenging to estimate 3D shape of a face from a single image and in particular with facial expressions, taken under unconstrained conditions. 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.) [27, 11, 14, 25, 6]. Recently, [21] showed that it is possible to use ideas from rigid photometric stereo [8, 18] for shape estimation by combining photos from a large Internet collection, they recovered a single shape which agrees with the majority of photos in the collection (even though the photos included different facial expressions).

Because single view reconstruction problem is ill-posed, all existing methods depend heavily on prior models. Blanz and Vetter [13] showed the potential for very high quality 3D face modeling from a single image by expressing 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 expressions and subtle details that vary from one individual to the next. There are three publicly available implementations of morphable models [3, 5, 4] to which we compare in the results section. Similarly,[24] reconstruct a shape by combining patches from a database of depths, [7] proposed general (non face specific) priors on depth and albedo for shape from shading. [19] produce single view reconstructions of scenes assuming availability of Kinect data and is not designed to work on faces, [2] learns transformation from image features to 3D structure to infer a scene's structure as a combination of planes (is not applicable to faces). Kemelmacher and Basri [20] use a shape-from-shading approach that requires only a single template face as a prior, thus the geometry varies significantly depending on which template is used. Most of the approaches require some kind of manual initialization. To summarize, all approaches require availability of high resolution depth data and apart from [13] are not designed to establish dense correspondence, thus cannot later modify the shape not the image. In our work we build a shape basis directly from the photos, moreover we establish dense correspondence between the shape basis and the input image enabling modification of input's shape and texture.

3. Basis construction from Internet photos

We begin by describing our data collection approach, and then show how to compute pixel-wise correspondence between every photo in the collection to a common reference (we call it "global correspondence"), and finally we show how given a new previously unknown input photo we

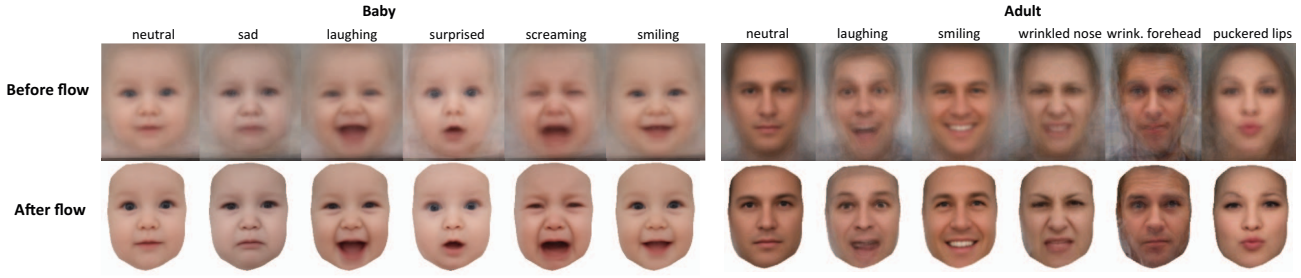

Figure 2. Averages of Internet photos, divided to clusters based on search terms. Each image represents an average of roughly 300 photos. Top: Photos are first aligned using rigid transformation (based on location of the eyes, nose and mouth), bottom: averages after pixel-wise alignment (much sharper).

automatically align it to the rest of the collection. This registration procedure allows us to construct an image basis that will be used in the next sections for shape estimation and modification from a single image.

3.1. Data Collection

Our major motivation in constructing an Internet based image basis is to address variations in facial shape that are challenging to capture with 3D scanning devices. Photos of young kids and babies are an example of this. We used image search queries like "smiling baby", "crying baby", "screaming baby" and so forth to collect sets of photos (we call them clusters) divided by semantic labels. Similarly we collected photos of adults making different facial expressions. As a result we collected around 4000 photos with roughly 300 photos per cluster.

3.2. Pixel-wise correspondence

The goal of this part is to obtain dense correspondence between any two photos in the collection. Let's assume for simplicity that we are given two images (I_1 and I_2) of the same person making different facial expressions while the pose and lighting are fixed. Finding dense correspondence between such two photos means looking for transformations $u(x, y)$ and $v(x, y)$ such that the distance between $I_1(x + u(x, y), y + v(x, y))$ and $I_2(x, y)$ is minimized, i.e., assuming brightness constancy is held (corresponding pixels have similar intensities).

This problem boils down to traditional optical flow estimation. It becomes a much more difficult problem for unconstrained photos, in particular taken with arbitrary lighting, due to violation of the brightness constancy. Recently, however, Kemelmacher and Seitz proposed a method they called "Collection Flow" (CF) [22] where they showed that given a large photo collection of same person, e.g., photos of a celebrity downloaded from the Internet, it is possible to leverage the collection for lighting invariant flow estimation. The key idea is to replace direct flow estimation $I_1 \rightarrow I_2$ with sub-flows $I_1 \rightarrow I_1'$ and $I_2' \rightarrow I_2$ that can be eventually combined, where I_1' and I_2' are low rank projections of the input images to the space of all photos. The projection was done in a way to capture the lighting (low frequency component) of the image while normalizing for the facial expression (high frequency component). Intuitively, flow was computed from an image to the collection average (modified to include the lighting of the particular image). This process alternates between improving the normalized images and flow estimation and converges rather quickly. Please see [22] for further details.

In our method, we follow the ideas presented in [22] but propose to apply collection flow on each cluster independently and then estimate flow between each cluster's average (computed by the method) to a global average (chosen as one of the clusters). There are two reasons to apply CF per cluster vs. the full collection. First, we found that CF performs well in case the photos have either similar identity (same gender, age and person) and varying facial expression (as in the original paper) or with roughly similar facial expression but varying identity. The performance degraded in case both factors change. Second, since the clusters have semantic meanings having separate flows per cluster enables morphing capabilities. Note that even though the facial expression is roughly the same per cluster it can still vary quite significantly across individuals (and since some of the expressions are not clearly defined with a particular search term, e.g., crying can be with closed mouth or open mouth) as can be seen in Fig. 1. Below we formalize the alignment process.

Given a photo I_1 in cluster i we'd like to estimate flow to photo I_2 in cluster j. We first run collection flow on cluster i to get the flow $I_1 \rightarrow A_i^1$ where A_i^1 is the average of cluster i illuminated by lighting L_1 of the input image, and similarly for I_2 we find flow $I_2 \rightarrow A_j^2$. This process is performed in parallel for all photos in the all clusters. We can then warp the images I_1 and I_2 to their respective cluster averages. The output is pixel wise correspondence from each cluster's photo to its average. Figure 2 shows the averages of clusters before (top) and after (bottom) the collection flow process. Note how much sharper the facial features look, indicating good correspondence. The method

is completely automatic. The averages also reveal major differences between expressions, e.g., "crying babies" and "laughing babies" both have an open mouth and narrower eyes (compared to neutral) however the extent to which the eyes are closed or the shape of the mouth causes a dramatic expression change.

To obtain correspondence across clusters we warp all images to their respective cluster average, from the warped images we construct an $f \times p$ matrix M_i (for cluster i) where f is the number of images and p number of pixels in each image. Each image is represented by a row in M_i, and similarly for cluster j. We then project the average of cluster i, A_i onto the global average A_g (chosen as the first cluster) obtaining A_i^g and estimate optical flow between A_i^g and A_g. This step is done to match color and lighting of the target cluster, in all cases the rank of the projection is rank-4 as in [22].

Once correspondence was obtained all images are warped to a common global reference and matrix M that contains warped images from all the clusters is constructed.

3.3. Registration of a new input photo

Given an input image, we'd like to align it to the rest of the collection. For this we estimate the distance between the HoG representation of the input image and each of the images in the collection. We choose the cluster number to which the image belongs by measuring to which cluster the majority of nearest neighbor images belong. Given the cluster number (say i) the image is projected to M_i and low-rank version of the input image is computed, and further optical flow between the low rank version and the input image is computed. This produces a correspondence between the input image and its cluster. All the subflows are computed using Ce Lui's implementation [23].

4. Single view reconstruction algorithm

Given the $f \times p$ matrix of *warped images* M, let's consider the intensity change at a particular pixel across the images. The change in intensity can be caused by difference in lighting or surface normal (due to facial expression– even though we aligned for 2D flow still there is a possible change in surface normals), and texture, e.g., freckles. Let an image be represented as $I(x,y) = L^T S(x,y)$ assuming Lambertian reflectance model, where $S \in \mathbb{R}^{4\times p}, L \in \mathbb{R}^{4\times 1}$, following [10]. Let us further assume that shape in an image can be represented by a linear combination of a set of basis shapes, i.e., $I(x,y) = L^T \sum_{i=1}^{k} S_i(x,y)$. Given that image representation, the rank of M should be $4K$. In the next part we will show how to factorize the matrix to enable recovery of the shape basis, lighting coefficients, and how to combine the shape basis to enable single view reconstruction.

The intuition behind this representation is that we use the images set to produce a set of basis shapes each of size $4 \times p$, that spans the shapes of the faces captured in images. This idea comes from classic photometric stereo [26, 9, 18] where it was shown that it's possible to factorize a set of images to lighting and shape (normals+albedo), it was further shown in [21] that it's possible to reconstruct the average shapes of a person's face by factorization of images of the same individual but with different facial expressions (the shape usually has a common expression–averaged expression of the dataset). In our work, we derive a basis of shapes that can represent the flow-warped collection. We further show how to recover the basis coefficients, and use them to reconstruct a facial shape *per* image. The main question, however, is how to separate the coefficients from the lighting representation? To this end, we propose a double-SVD approach, which includes rank constraints due to the lighting representation. We were inspired by Bregler's non-rigid shape factorization [15], however there it was done for a completely different problem–separating pose and shape parameters. [28] created a basis that spans flow and normals given photos of different people, with the *same expression (neutral)* to use for recognition. Photos were not pre-aligned using optical flow, and thus additional constraints, e.g., symmetry, were needed. [16] create a basis that spans deformations due to flow but consider a controlled video sequence that is taken with 3 colored lights (thus every facial expression in every frame can be reconstructed using rigid photometric stereo). These two works did not present reconstructon results. We are not aware of any other approach that considered separation of lighting and non-rigid deformation coefficients.

4.1. Factorization to deformation and lighting

We factorize M using Singular Value Decomposition, $M = UDV^T$ and take the rank-$4K$ approximation to get $M = PB$ where $P = U\sqrt{D}$ is $f \times 4K$ and $B = \sqrt{D}V^T$ is $4K \times p$. In the absence of ambiguities, P should contain a combination of low order coefficients of the lighting and coefficients that combine the shape basis B. In general, however there is a $4K \times 4K$ ambiguity since M can be represented also by $M = PA^{-1}AB$ which needs to be resolved. We will discuss that ambiguity at the end of the section. The question is how to factor the matrix P to recover the basis coefficients. Let us look closer on the representation of M given our assumptions:

$$M = PB = \begin{bmatrix} \cdots \\ c_{i1}L_i \ldots c_{ik}L_i \\ \cdots \end{bmatrix} \begin{bmatrix} B_1 \\ \vdots \\ B_k \end{bmatrix} \quad (1)$$

where $L_i \in R^{1\times 4}$ represent the lighting coefficients in an image i, and c_{ij} where $j = 1, \cdots, k$ are the basis coeffi-

cients, $B \in \mathbb{R}^{4k \times p}$ is the basis.

To recover the coefficients we propose the following idea. Given that repetitive structure in each row of P, we can reshape each row to a $4 \times k$ rank-1 matrix $P'_i \in \mathbb{R}^{k \times 4}$, i.e.

$$P'_i = \begin{bmatrix} c_{i1} L_i \\ \vdots \\ c_{ik} L_i \end{bmatrix} = \begin{bmatrix} c_{i1} \\ \vdots \\ c_{ik} \end{bmatrix} \begin{bmatrix} L_{i1} \dots L_{i4} \end{bmatrix} \quad (2)$$

Factorizing P'_i with SVD and taking the rank-1 projection will give us the coefficients and lighting. This assumes specific structure of rows of P. To obtain such structure we need to solve for gauge ambiguity. It was shown in [12] that classic photometric stereo factorization can recover lighting and shape up to 3×3 Generalized Bas-Relief ambiguity, and for arbitrary lighting approximated with first order spherical harmonics up to a 4×4 Lorentz transformation [9]. In our case, generally the ambiguity can be $K \times K$. Instead of directly looking for the ambiguity matrix we propose Algorithm 4.1 that leverages desired structure of matrices P and B for ambiguity-aware deformation coefficient recovery. The key idea is to calculate best rank-1 approximation of every row in P and then re-estimate the basis shapes B according to the approximation. This process is iterated until convergence, which typically happens after $5 - 6$ iterations.

Data: M is $f \times p$ matrix of images;
$M = UDV'$;
denote $P = U\sqrt{D}$ and $B = \sqrt{D}V^T$;
Result: P and B for which the rank-1 condition holds
while *until convergence* **do**
  for *every image* i **do**
    $P(i,:)$ is $1 \times 4K$;
    reshape $P(i,:)$ to $4 \times K$ matrix P';
    run SVD on $P' = udv^T$;
    set $\alpha = u(:,1)d(1,1)$;
    set $l = v(:,1)$;
    reshape αl^T to $1 \times 4K$ vector;
    $P(i,:) = [\alpha l^T]_{1 \times 4K}$;
  end
  Estimate B s.t. $\min ||M - PB||^2$;
end

Algorithm 1: Modify P and B to hold rank-1 condition

Once P and B are estimated, we can determine the true lighting l and basis coefficients α (in Alg. 4.1) per image. Given the basis coefficients the sought shape is $S = \sum_{j=1}^{K} c_{ij} B_j$. Given the shape matrix, the surface can be recovered by integration (see for example section 2.5 in [21]), in our case we recover a shape matrix per color channel and integrate the three shape matrices together (instead of one equation per pixel we have 3 equations). Once the depth is reconstructed, it is still in the 2d state of the global reference and therefore inverse flow should be applied to transform the shape from reference to the original expression of the input image. The inverse flow is obtained from the flows between cluster averages, as in Section 3.2.

5. Synthesis of novel 2D and 3D views

Once depth per image is reconstructed and correspondence between every image to every other image in the dataset obtained, it becomes possible to transform between different faces and expressions. Specifically, we can change the expression of a person from a single image by transforming it using the flow between the clusters. To synthesize view of cluster j from image in cluster i we project the photo (aligned by flow to the cluster average) onto the rank-4 cluster i basis and also onto the rank-4 cluster j basis, yielding a pair of illumination-aligned projections. Subtracting these two projections yields a difference image between clusters i and j, under the same illumination as the input photo, and adding it to the input photo yields a texture change. We also apply the flow difference, between cluster i and j, warped to the coordinate system of the input photo.

6. Results

We show results and evaluations in this section. In the paper we mostly show reconstructions of babies/young kids because these are the most challenging for any single view method, it demonstrates our point that it is very difficult to collect a database of 3D scans that work for any person in the world, using photos however it's a much easier task. In Fig. 4 we show many input photos and the corresponding reconstructions automatically obtained using our method (for each photo we show three views of the reconstruction). We intentionally show un-textured surfaces to show the real reconstruction–texture often hides problems in reconstruction. Note the dramatic difference in facial expressions which is captured in the reconstruction (going from laughing to screaming to sad and so forth), the change in identity, ethnicity and gender, variety of lighting conditions, etc. The method works on completely unconstrained photos.

In Figure 3 we demonstrate how based on the dense correspondence that is found between every photo to every cluster in the collection, it is possible to make automatic modifications to the input photo to achieve change in the facial expression. Figure 5 further shows that, similarly, flow can be used to modify the 3D shape.

We have compared our single view estimation method to all the available methods we found: 1) Image-Metrics [4] ("PortableYou"): the user is asked to choose a gender of the person, and then the process is completely automatic. 2)

Vizago [5]: the user is asked to specify gender and manually click on 12 points (3 on the contour of the face–chin and sides, 2 on ears, 3 on the nose, 2 on eyes and 2 corners of the mouth). 3) Kemelmacher and Basri [20] on the YaleB dataset. Below we discuss these comparisons in more detail. Figure 7 shows reconstruction results of Vizago [5] and Image Metrics [4], both are implementations of the morphable model method. By observing the profile views we see that the shapes are mostly of an average adult and do not capture the facial expression. The shapes are also typically shown textured and therefore it is harder to see the underlying shape. These two results are typical. We have also experimented with FaceGen [3] which produces similar results.

In Figure 6 we compare to the single view method of [20], we ran our algorithm on the YaleB [17] dataset (image from YaleB was an input to our method–we didn't use any of the extra data available with YaleB images, e.g., fiducials, lighting, etc.) and present several typical reconstructions (3rd column). Second column presents the ground truth shape–estimated by taking all the photos of the same person given in YaleB and running calibrated photometric stereo (known lighting directions per image) [26]. Column 4 shows the depth map difference between our single view reconstruction and photometric stereo, below each difference there is the mean error and standard deviation in percents. We used exactly the same measure as in [20]'s Figure 7, i.e., 100 $z_{gt} - z_{rec}/z_{gt}$. We get comparable results (or slightly better), their typical error is $6-7\%$ while ours is 4%. Note that [20] is not designed to work with facial expressions and its performance degrades when the input photo is less similar to the reference template. Due to this we only present results on adults with neutral expression.

We have also tried non-face specific methods such as [19] and [2] and both do not perform well on unconstrained face images, we therefore do not include a comparison.

Please refer to the supplementary material for more examples.

7. Conclusions

We believe that morphable models have a huge potential to advance unconstrained face modeling, however most existing methods heavily depend on priors that are challenging to construct, e.g., aligned 3D scans, limited to single expression. The key idea of this paper, is to find a way to leverage photographs (which already exist on the Internet) for construction of a morphable model basis. To this end, we showed that if photos can be divided to "clusters" based on semantic labels (e.g., "smiling", "sad"), we can 1) get dense pixel-wise correspondence between any pair of photos in different clusters that represent facial expressions, e.g., smiling photo to sad photo, and 2) use this correspondence to analyze the space of warped images, i.e., factor to lighting and deformation. This enabled a new single view shape reconstruction and modification method, with exciting results on very challenging photos, e.g., faces with extreme expressions taken in uncontrolled conditions.

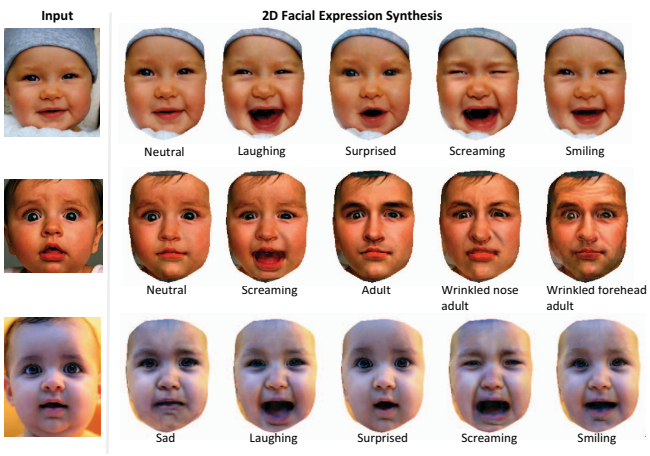

Figure 3. Given a single input image (left), the method can automatically synthesize the same person in different facial expressions using the derived morphable model.

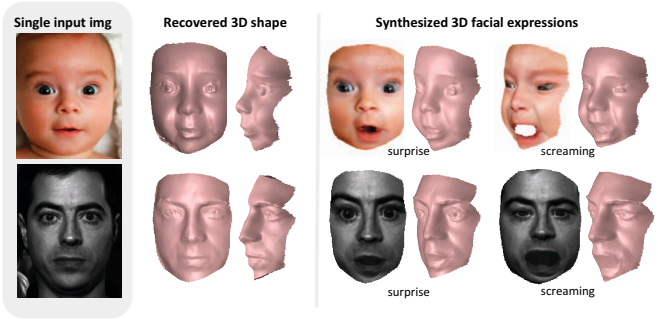

Figure 5. Creating 3D facial expressions from a single view.

References

[1] Frgc (face recognition grand challenge) by nist. 1

[2] http://make3d.cs.cornell.edu/. 2, 6

[3] http://www.facegen.com/. 2, 6

[4] http://www.image-metrics.com/. 2, 5, 6, 8

[5] http://www.vizago.ch/. 2, 6, 8

[6] 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

[7] J. T. Barron and J. Malik. Shape, albedo, and illumination from a single image of an unknown object. *CVPR*, 2012. 2

[8] R. Basri, D. Jacobs, and I. Kemelmacher. Photometric stereo with general, unknown lighting. *International Journal of Computer Vision*, pages 239–257, 2007. 2

[9] R. Basri, D. Jacobs, and I. Kemelmacher. Photometric stereo with general, unknown lighting. *International Journal of Computer Vision*, 72(3):239–257, 2007. 4, 5

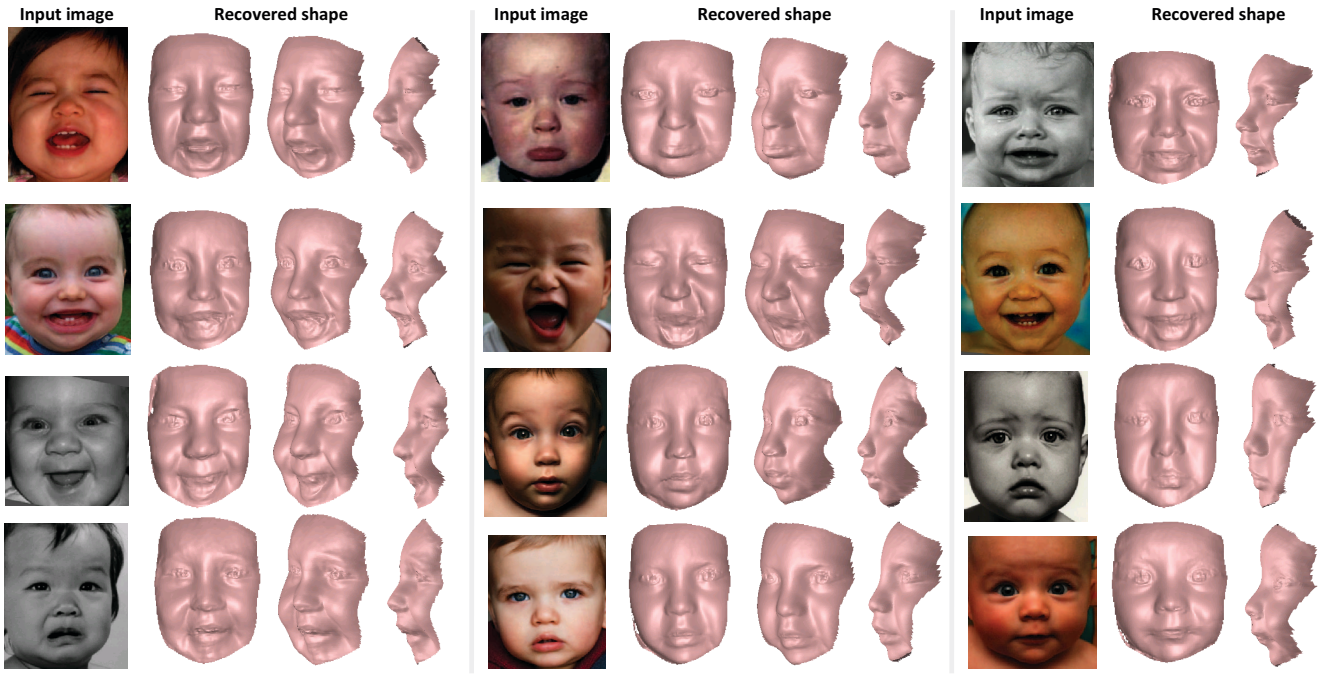

Figure 4. Single view reconstructions automatically generated using our method. Note how the dramatic changes due to facial expressions, identity, pose, lighting, etc.

[10] R. Basri and D. W. Jacobs. Lambertian reflectance and linear subspaces. *PAMI*, 25(2):218–233, 2003. 4

[11] 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

[12] P. N. Belhumeur, D. J. Kriegman, and A. L. Yuille. The bas-relief ambiguity. *International Journal of Computer Vision*, 35(1):33–44, 1999. 5

[13] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *SIGGRAPH*, pages 187–194, 1999. 1, 2

[14] 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

[15] C. Bregler, A. Hertzmann, and H. Biermann. Recovering non-rigid 3d shape from image streams. In *CVPR*, 2000. 4

[16] J. M. Buenaposada Biencinto, E. Muñoz, and L. Baumela Molina. A model of brightness variations due to illumination changes and non-rigid motion using spherical harmonics. 2008. 4

[17] A. Georghiades, P. Belhumeur, and D. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE PAMI*, 2001. 6

[18] H. Hayakawa. Photometric stereo under a light source with arbitrary motion. *JOSA A*, 11(11):3079–3089, 1994. 2, 4

[19] K. Karsch, C. Liu, and S. B. Kang. Depth extraction from video using non-parametric sampling. In *ECCV*, 2012. 2, 6

[20] I. Kemelmacher-Shlizerman and R. Basri. 3d face reconstruction from a single image using a single reference face shape. *IEEE PAMI*, 2010. 2, 6, 8

[21] I. Kemelmacher-Shlizerman and S. M. Seitz. Face reconstruction in the wild. In *ICCV*, 2011. 2, 4, 5

[22] I. Kemelmacher-Shlizerman and S. M. Seitz. Collection flow. In *CVPR*, pages 1792–1799, 2012. 3, 4

[23] C. Liu. *Beyond Pixels: Exploring New Representations and Applications for Motion Analysis*. PhD thesis, MIT, 2009. 4

[24] T.Hassner and R. Basri. Example based 3d reconstruction from single 2d images. In *Beyond Patches Workshop CVPR*, 2006. 2

[25] T. Weise, B. Leibe, and L. V. Gool. Fast 3d scanning with automatic motion compensation. In *CVPR*, June 2007. 2

[26] R. J. Woodham. Multiple light source optical ow. *ICCV*, page 4246, 1990. 4, 6

[27] 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

[28] S. K. Zhou, R. Chellappa, and D. W. Jacobs. Characterization of human faces under illumination variations using rank, integrability, and symmetry constraints. In *Computer Vision-ECCV 2004*, pages 588–601. Springer, 2004. 4

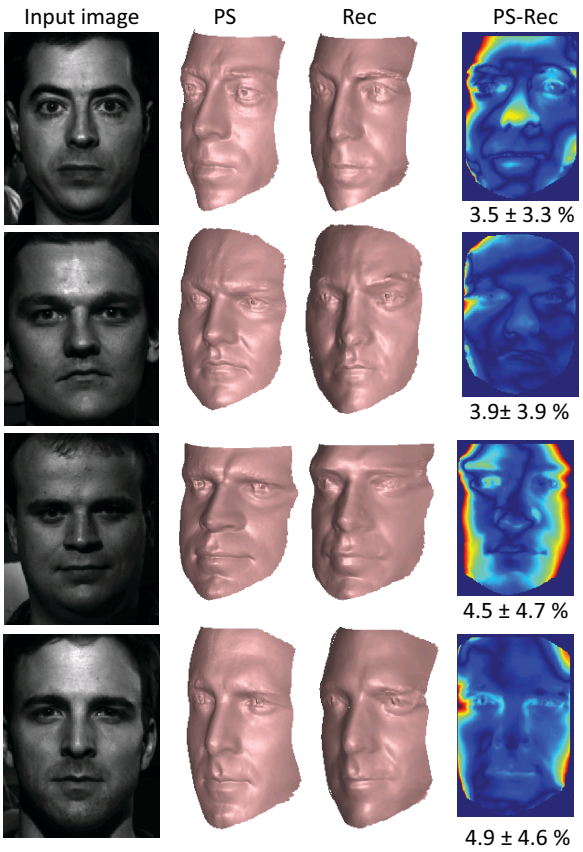

Figure 6. Comparison to [20] on YaleB dataset. We use exactly the same measure as in Fig.7 of [20] and show several typical reconstructions on this dataset compared to calibrated photometric stereo (known lighting). We get comparable results, typical reconstruction error in [20] is around $6-7\%$ while ours in typically 4%.

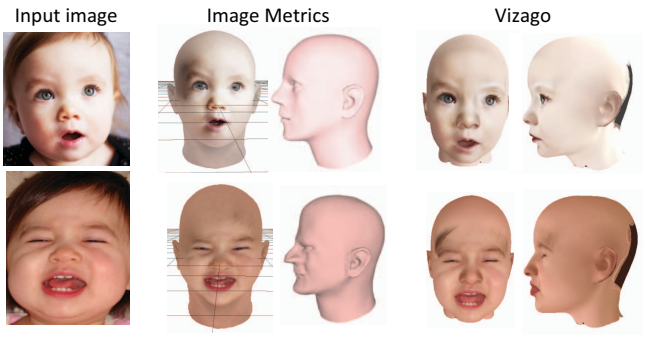

Figure 7. Typical reconstruction with morphable model methods [5, 4] on baby pictures. The shape does not account for facial expressions, and looks close to the average person model (note the profile views of the reconstruction). Compare with Figure 4 that includes our results.