Expression Flow for 3D-Aware Face Component Transfer

Fei Yang[1] Jue Wang[2] Eli Shechtman[2] Lubomir Bourdev[2] Dimitri Metaxas[1]

[1]Rutgers University [2]Adobe Systems

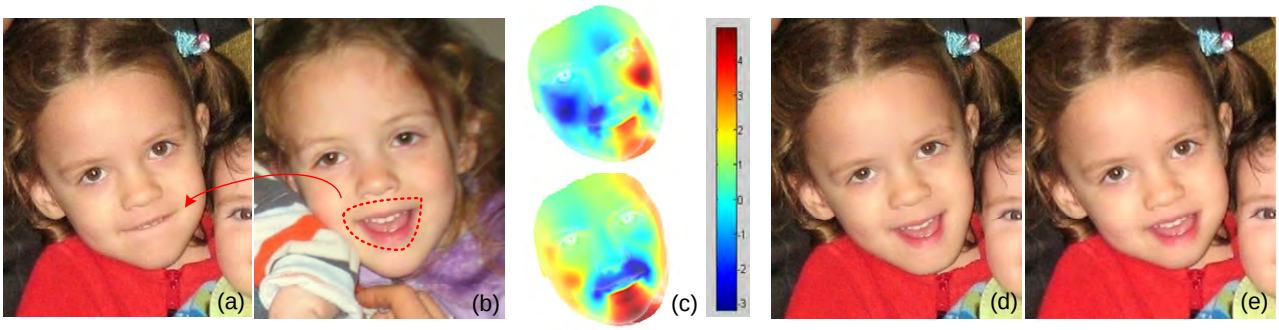

Figure 1: *Example of applying the proposed expression flow for face component transfer. (a) and (b) are input images, and the user wants to replace the closed mouth in (a) with the open mouth in (b). (c). Expression flow generated by our system, which warps the entire face in (a) to accommodate the new mouth shape. Top: horizontal flow field, bottom: vertical flow filed. (d) Final composite generated by our system. (e). Composite generated using 2D alignment and blending. Note the unnaturally short distance between the mouth and the chin.*

Abstract

We address the problem of correcting an undesirable expression on a face photo by transferring local facial components, such as a smiling mouth, from another face photo of the same person which has the desired expression. Direct copying and blending using existing compositing tools results in semantically unnatural composites, since expression is a global effect and the local component in one expression is often incompatible with the shape and other components of the face in another expression. To solve this problem we present Expression Flow, a 2D flow field which can warp the target face globally in a natural way, so that the warped face is compatible with the new facial component to be copied over. To do this, starting with the two input face photos, we jointly construct a pair of 3D face shapes with the same identity but different expressions. The expression flow is computed by projecting the difference between the two 3D shapes back to 2D. It describes how to warp the target face photo to match the expression of the reference photo. User studies suggest that our system is able to generate face composites with much higher fidelity than existing methods.

CR Categories: I.4.9 [IMAGE PROCESSING AND COMPUTER VISION]: Applications

Keywords: facial expression, facial component, face modeling, facial flow, image warping

Links: DL PDF WEB

1 Introduction

Everyone who has the experience of taking photographs of family members and friends knows how hard it is to capture the perfect moment. For one, the camera may not be at the right setting at the right time. Furthermore, there is always a delay between the time one sees a perfect smile in the viewfinder and the time that the image is actually captured, especially for low-end cell phone cameras which have slow response. For these reasons, face images captured by amateur photographers often contain various imperfections. Generally speaking, there are two types of imperfections. The first type is photometric flaws due to improper camera settings, thus the face may appear to be too dark, grainy, or blurry. The second type, which is often more noticeable and severe, is the bad expression of the subject, such as closed eyes, half-open mouth, etc.

With recent advances in image editing, photometric imperfections can be largely improved using modern post-processing tools. For instance, the personal photo enhancement system [Joshi et al. 2010] provides a set of adjustment tools to correct global attributes of the face such as color, exposure, and sharpness. Compared with photometric imperfections, expression artifacts are much harder to correct. Given a non-smiling face photo, one could simply find a smiling photo of the same person from his/her personal album, and use it to replace the whole face using existing methods [Bitouk et al. 2008]. Unfortunately, this global swap also replaces other parts of the face which the user may want to keep. Local component transfer among face images is thus sometimes more preferable.

However, local component transfer between face images with different expressions is a very challenging task. It is well known in the facial expression literature [Faigin 1991] that expressions of emotion engage both signal-intensive areas of the face: the eye region, and the mouth region. For an expression of emotion to appear genuine, both areas need to show a visible and coordinated pattern of activity. This is particularly true of the sincere smile, which in its broad form alters almost all of the facial topography from the lower eyelid downwards to the bottom margin of the face. While general image compositing tools [Agarwala et al. 2004] allow the user to crop a face region and seamlessly blend it into another face, they are incapable of improving the compatibility of the copied com-

ponent and the target face, as the example shown in Figure 1. To replace the closed mouth in Figure 1a with an open one in Figure 1b, a straightforward solution is to crop the mouth region, apply additional alignment adjustments, and seamlessly blend it into the target face. However, the resulting composite is semantically very unnatural (Figure 1e). This is because, when the mouth opens, the shape of the whole lower-half of the face deforms accordingly. To our best knowledge there are no existing tools that automatically handle these deformations for creating realistic facial composites.

We address this problem by presenting *Expression Flow*, a 2D flow field applied on the target image to deform the face in such a way that it becomes compatible with the facial component to be copied over. To compute the expression flow we first reconstruct a 3D face shape for each image using a dataset of other people's face shapes. Unlike traditional 3D fitting which tries to minimize the fitting error on each image, we *jointly* reconstruct a pair of 3D shapes, which have the same identity, but with different expressions that match our input image pair. This is formulated as an optimization problem with the objective to minimize the fitting errors with a person identity constraint. A 3D flow is then computed from the pair of aligned 3D shapes, and projected to 2D to form the 2D expression flow. The shapes are also used to warp the 3D pose of the new component before blending in. Due to the identity constraint, the expression flow reflects changes mainly due to differences of expression, and can deform the face in a natural way, as shown in Figure 1d.

Our expression flow is a hybrid of 3D and 2D methods. On the one hand, we rely on rough 3D shapes to compute the expression difference between faces with different poses. Since typical expression flows contain much lower level of detail (frequencies) than typical appearance details, we found that our rough 3D reconstruction is adequate for the purpose of expression transfer. On the other hand, we rely on 2D methods to warp face images and transfer local details between them. Our system thus has a greater flexibility and a wider application range than previous 3D and 2D expression transfer methods (see Section 2).

Based on the proposed expression flow we develop an efficient face compositing tool. To evaluate the effectiveness and generality of the proposed system, we conducted a comprehensive user study. The results suggest that the face composites created by our system have much higher fidelity than those generated by previous methods.

2 Related Work

Our work is related to previous research on face editing, facial expression mapping, face alignment, 3D shape fitting and image compositing.

Face Image Editing. Face image enhancement has been the subject of extensive work. Earlier approaches use generic face images as training data for applications such as super-resolution [Liu et al. 2007] and attractiveness enhancement by global face warping [Leyvand et al. 2008]. Recently Joshi et al. [2010] proposed a system to adjust global attributes such as tone, sharpness and lighting of a face image using personal priors. Blanz et al. [2004] fit a morphable 3D model to a face image, and then render a new face using the same pose and illumination to replace it. The face swapping system [Bitouk et al. 2008] achieves a similar goal by constructing and using a large face image library. A real-time system for retrieving and replacing a face photo based on expression and pose similarity was shown in [Shlizerman et al. 2010]. All these systems target global face editing. However replacing an entire head or face is often not desired for personal photo editing, global warping does not handle large topology and appearance changes, and generating realistic textured head models and compositing them into existing photos remains a challenging problem. Our method combines global warping and local compositing of face parts for an effective by-example expression editing.

Expression Mapping. There is also a large body of work on transferring expressions between images, which falls into two categories: 3D methods and 2D approaches. 3D approaches, such as the expression synthesis system proposed by Pighin et al. [1998] and the face reanimating system proposed by Blanz et al. [2003], try to create photorealistic textured 3D facial models from photographs or video. Once these models are constructed, they can be used for expression interpolation. However, creating fully textured 3D models is not trivial. In order to achieve photorealism the system has to model all facial components accurately such as the eyes, teeth, ears and hair, which is computationally expensive and unstable. These systems thus can only work with high resolution face images shot in controlled indoor environments, and unlike our system, are not robust enough to be used on day-to-day personal face photos.

2D expression mapping methods [Williams 1990] extract facial features from two images with different expressions, compute the feature difference vectors and use them to guide image warping. Liu et al. [2001] propose an expression ratio image which captures both the geometric changes and the expression details such as wrinkles. However, due to the lack of 3D information, these methods cannot deal with faces from different view points. Most importantly, these methods alone cannot synthesize features that are not in the original image, such as opening a mouth.

Facial Feature Localization. Various techniques have been proposed for facial feature localization on images as well as in video [Decarlo and Metaxas 2000]. Most of them combine local feature detectors with global geometric constraints. The widely-used Active Shape Model [Cootes et al. 1995] learns statistical distributions of feature points, thus allowing shapes to vary only in ways seen in a training set. Active Appearance Models [Cootes et al. 2001] explore image intensity distributions for constraining the face shape. Pictorial structure methods [Felzenszwalb and Huttenlocher 2005] localize features by maximizing the posterior probability for both appearance and shape. Recent work in this field also includes component-based discriminative search [Liang et al. 2008], and a subspace-constrained mean shift method [Saragih et al. 2009].

3D Shape Fitting. Recovering the 3D face shape from a single image is a key component in many 3D-based face processing systems. Blanz and Vetter [1999] optimize the parameters of a 3D morphable model by gradient descent in order to render an image that is as close as possible to the input image. Romdhani and Vetter [2003] extend the inverse compositional image alignment algorithm to 3D morphable models. Shape-from-shading approaches are also applied to 3D face reconstruction [Dovgard and Basri 2004; Shlizerman and Basri 2011]. Kemelmacher-Shlizerman et al. [2010] show how to find similarities in expression under different poses, and use a 3D-aware warping of facial features to compensate for pose differences.

Image Compositing. General image compositing tools such as the photomontage system [Agarwala et al. 2004] and the instant cloning system [Farbman et al. 2009] allow image regions from multiple sources to be seamlessly blended together, either by Poisson blending [Pérez et al. 2003] or using barycentric coordinates. Sunkavalli et al. [2010] propose a harmonization technique which allows more natural composites to be created.

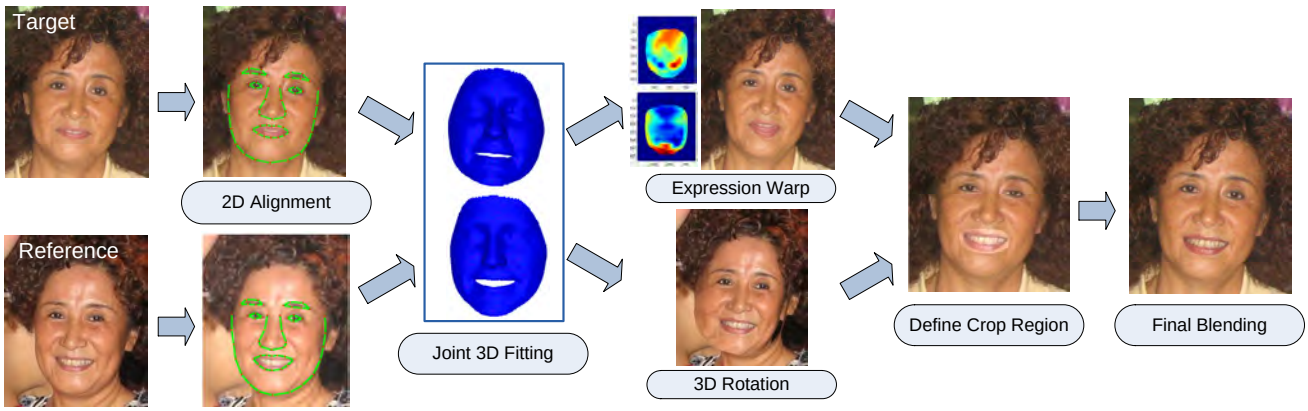

Figure 2: *The flow chart of our system.*

3 Our System

3.1 System Overview

Figure 2 shows the flow chart of the proposed system. Given a target face image which the user wants to improve, and a reference image which contains the desired feature to be copied over, our system first uses computer vision techniques to automatically extract facial feature points on both images. Based on the extracted feature points, we then jointly reconstruct 3D face shapes for both images using a 3D face expression dataset. Our 3D fitting algorithm makes sure that the two shapes have the same identity, thus the main difference between them is due to changes in expression. We then compute a 3D flow by subtracting the two shapes and project it to 2D to create the expression flow. The expression flow is used to warp the target face. We also use the 3D shapes to align in 3D the reference face to the target face. The user then specifies the region of the facial feature to be transferred, which is then seamlessly blended into the target image to create the final composite.

3.2 Single Image Fitting

We first describe how to fit a 3D face shape to a single face image. Given the input image, the facial landmarks are first localized using Active Shape Model (ASM) [Cootes et al. 1995], a robust facial feature localization method. Following Milborrow and Nicolls's approach [2008], we localize 68 feature points, as shown in Figure 2.

We represent the 3D geometry of a face with a shape vector $s = (x_1, y_1, z_1, \cdots, x_n, y_n, z_n)^T$ that contains X, Y, Z coordinates of its n vertices. Following Blanz and Vetter's work [1999], we define a morphable face model using Principal Component Analysis (PCA) on the training dataset. Denote the eigenvectors as v_i, eigenvalues as λ_i, and the mean shape as $\bar{s}$, a new shape can be generated from the PCA model as:

$$s_{new} = \bar{s} + \sum \beta_i v_i = \bar{s} + \mathbf{V} \cdot \beta. \tag{1}$$

The 3D fitting is performed by varying the coefficients β in order to minimize the error between the projections of the pre-defined landmarks on the 3D face geometry, and the 2D feature points detected by ASM. We apply a weak perspective projection model, and define the fitting energy for the kth landmark as:

$$E_k = \frac{1}{2}||R \cdot (\bar{s}^{(k)} + \mathbf{V}^{(k)} \cdot \beta) - X^{(k)}||^2, \tag{2}$$

where R is the 2 by 3 projection matrix, $\mathbf{V}^{(k)}$ is the sub-matrix of $\mathbf{V}$ consisting of the three rows that corresponding to X, Y, Z coordinates of the kth landmark. $X^{(k)} = (x^{(k)}, y^{(k)})^T$ is X, Y coordinates of the kth landmark detected from the face image.

Assuming a Gaussian distribution of the training data, the probability for coefficients β is given by:

$$p(\beta) \sim exp[-\frac{1}{2}\sum(\beta_i/\lambda_i)^2]. \tag{3}$$

Let $\Lambda = diag(\lambda_1^2, \lambda_2^2, \cdots, \lambda_L^2)$. We define the energy of coefficients as:

$$E_{coef} = \frac{1}{2} \cdot \beta^T \Lambda^{-1} \beta. \tag{4}$$

The total energy function to be minimized is thus the combination of the two terms:

$$E = \sum w_k E_k + c \cdot E_{coef}, \tag{5}$$

where c is a parameter controlling the tradeoff between the fitting accuracy and the shape fidelity, which is set to 5×10^6 in our system. w_k is the weight for the kth landmark. In our system we set $w_k = 0.5$ for landmarks of eyebrows, since our training shapes are textureless and these landmarks are hard to be labeled accurately. We empirically set $w_k = 2$ for contour points, $w_k = 3$ for mouth points, and $w_k = 1$ for all other points.

To minimize E, we set $\nabla_\beta E = 0$, which leads to:

$$\beta = P^{-1}Q, \tag{6}$$

where

$$P = \sum w_k (R\mathbf{V}^{(k)})^T R\mathbf{V}^{(k)} + c\Lambda^{-1}, \tag{7}$$

$$Q = \sum w_k (R\mathbf{V}^{(k)})^T (X^{(k)} - R\bar{s}^{(k)}). \tag{8}$$

The above closed-form solution assumes that we know $V^{(k)}$, the 3D vertices corresponding to the k-th landmark. For landmarks located inside the face region we can simply hard-code the corresponding 3D vertex. However, landmarks along the face contour do not have a single corresponding vertex; they must be matched with 3D vertices along the face silhouette. We therefore employ a two-stage optimization approach to find the optimal β. In the first stage we find the correspondences between vertices and landmarks by projecting the vertices onto the image plane, finding their convex

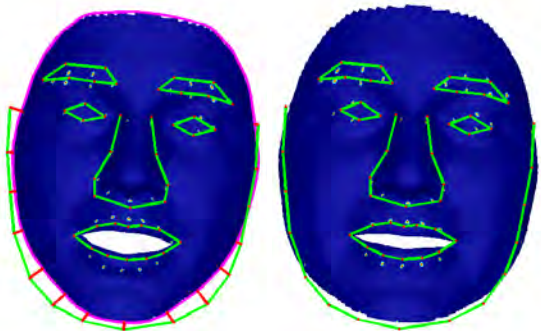

Figure 3: *Fitting a 3D shape to the target image in Figure 2 using our two-stage optimization algorithm. Left: How the shape deforms. Green lines are ASM features lines, the pink line is the projected face contour from face geometry. The short red lines show the contour landmarks projected onto the face contour. Right: fitted face shape after 3 iterations.*

hull and assigning each landmark to the closest point on the convex hull, as shown on Figure 3(left). In the second stage we deform the face shape by minimizing the energy in Equation 5. We repeat the two stages until the shape converges. Figure 3(right) shows the result after three iterations. We can see that the proposed approach minimizes the fitting error. The algorithm is formally described in Algorithm 1.

Algorithm 1 *Single Image Fitting*

Input: facial landmarks $X^{(1),\cdots,(K)}$ and the shape PCA model.

Output: shape s that best fits the landmarks.

1: Set $\beta = 0$.
2: **repeat**
3: Set $s = \bar{s} + \mathbf{V}\beta$.
4: Find projection matrix R from s and $X^{(1),\cdots,(K)}$ by using the least squares method.
5: Project all vertices of s onto the image plane: $s' = P(R, s)$.
6: Find the convex hull of s' as $H(s')$.
7: For contour landmarks X^i, find correspondence using $H(s')$.
8: Solve β in Equation 6.
9: **until** β converges.

3.3 Expression Models

To train the PCA model we use the face expression dataset proposed by Vlasic et al. [2005]. This dataset contains 16 subjects, each performing 5 visemes in 5 different expressions. This dataset is pre-aligned so that the shapes have vertex-to-vertex correspondence.

Building a single PCA model using all training shapes is problematic, since the training shapes vary in both identity and expression. A single PCA might not be expressive enough to capture both types of variations (underfitting), and also does not allow to distinguish between the two. We thus build a PCA model for each expression separately. We could also use more sophisticated nonlinear methods (e.g. manifold [Wang et al. 2004]). However, since that face shapes do not vary dramatically, we have found that this approximation gives desired results.

For a given image, we select the PCA model that gives the minimum reconstruction error using the fitting algorithm described above. The target and reference face therefore may fall into different expression models. We denote the PCA model for the target image as $(\mathbf{V}^t, \Lambda^t, \bar{s}^t)$, and its training shapes as $\mathbf{S}^t = (s_1^t, ..., s_M^t)$. Similarly, we denote the model and its training shapes for the reference image as $(\mathbf{V}^r, \Lambda^r, \bar{s}^r, \mathbf{S}^r)$. The new shapes to be reconstructed from the images are denoted as s^t and s^r.

3.4 Joint Fitting

Using the constructed expression models and the single image fitting approach proposed above, a natural idea is to fit each input image individually, and then try to generate the expression flow by subtracting the two shapes. However, we found that this approach does not work well. The reason is that each 3D shape is a linear combination of all face shapes in the training dataset, which contains faces from multiple human subjects. By fitting the 3D shape individually to each image, we essentially generate 3D shapes that have different virtual identities. The difference between the two shapes is then mainly due to identity difference, not expression difference.

To solve this problem, our approach is to jointly fit 3D shapes to input images so that they have the same identity. To add such a constraint, we re-formulate s^t as a linear combination of the original training shape vectors s_i^t, parameterized with new coefficients γ_i^t $(i = 1, ..., M)$ as:

$$s^t = \bar{s}^t + \mathbf{V}^t\beta^t = \bar{s}^t + \mathbf{S}^t\gamma^t. \tag{9}$$

Similarly, we re-formulate s^r as:

$$s^r = \bar{s}^r + \mathbf{V}^r\beta^r = \bar{s}^r + \mathbf{S}^r\gamma^r. \tag{10}$$

The coefficients γ^t and γ^r describe the face shape of a new person under a certain expression as a linear combination of the shapes of the training subjects under the same expression. They essentially define the virtual identities of the two 3D shapes as a linear combination of training subjects. Since s_i^t and s_i^r correspond to the same human subject, by enforcing $\gamma^t = \gamma^r = \gamma$, we guarantee that s^t and s^r have the same identity. We thus replace γ^t and γ^r with a single γ.

From Equation 9 we have $\beta^t = (\mathbf{V}^t)^T\mathbf{S}^t \cdot \gamma$. Substituting β with γ in Equation 4, the coefficient energy for s^t becomes:

$$E_{coef}^t = \frac{1}{2} \cdot \gamma^T((\mathbf{V}^t)^T\mathbf{S}^t)^T(\Lambda^t)^{-1}((\mathbf{V}^t)^T\mathbf{S}^t)\gamma. \tag{11}$$

Replacing t with r we have the formulation for the coefficient energy E_{coef}^r for s^r. To jointly fit s^t and s^r, we minimize the total energy:

$$E_{total} = \sum w_k(E_k^t + E_k^r) + c \cdot (E_{coef}^t + E_{coef}^r). \tag{12}$$

The optimal γ that minimizes this total energy is:

$$\gamma = (P^t + P^r)^{-1}(Q^t + Q^r), \tag{13}$$

where P^t and P^r have the same formulation as:

$$P = \sum w_k(R\mathbf{S}^{(k)})^T R\mathbf{V}^{(k)} + c(\mathbf{V}^T\mathbf{S})^T\Lambda^{-1}\mathbf{V}^T\mathbf{S}. \tag{14}$$

Substituting $R, \mathbf{S}, \mathbf{V}, \Lambda$ with $R^t, \mathbf{S}^t, \mathbf{V}^t, \Lambda^t$ and $R^r, \mathbf{S}^r, \mathbf{V}^r, \Lambda^r$ gives us P^t and P^r, respectively. Similarly, Q^t and Q^r are defined as:

$$Q = \sum w_k(R\mathbf{S}^{(k)})^T(X^{(k)} - R\bar{s}^{(k)}), \tag{15}$$

and substituting $R, \mathbf{S}, X, \bar{s}$ with $R^t, \mathbf{S}^t, X^t, \bar{s}^t$ and $R^r, \mathbf{S}^r, X^r, \bar{s}^r$ gives us the formulation for Q^t and Q^r.

The joint fitting algorithm is formally described as follows:

Algorithm 2 *Joint Fitting*

Input: facial landmarks of two images, and all PCA models.

Output: shapes s^t and s^r that jointly fit the landmarks on both images.

1: Apply Algorithm 1 to each image to determine their expression models V^t and V^r, as in Section 3.3.
2: Set $\gamma = 0$.
3: **repeat**
4: Set $s^t = \bar{s}^t + \mathbf{S}^t\gamma$ and $s^r = \bar{s}^r + \mathbf{S}^r\gamma$.
5: For each image, apply step 4-7 in Algorithm 1.
6: Solve the common γ in Equation 13.
7: **until** γ converges.

3.5 Computing 2D Flow

We first align the two 3D shapes to remove the pose difference. Since the reconstructed 3D shapes have explicit vertex-to-vertex correspondences, we can compute a 3D difference flow between the two aligned 3D shapes and and project it onto the image plane to create the 2D expression flow. The flow is further smoothed to remove noise. An example of the final expression flow is shown in Figure 4. Figure 4a shows the horizontal flow, where red color means positive movement in X direction (to the right), and blue means negative movement (to the left). This figure essentially describes how the mouth gets wider when the person smiles. Figure 4b shows the vertical flow, where red color means positive movement along Y axis (moving down), and blue means negative movement (moving up). It illustrates that when the person smiles, her jaw gets lower, and the cheeks are lifted.

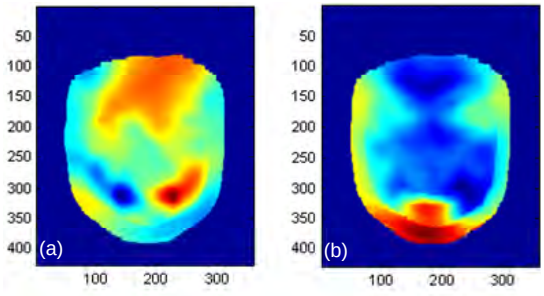

Figure 4: *2D expression flow computed from the example shown in Figure 2. (a) Horizontal flow field. (b) Vertical flow field.*

As shown in Figure 2, by applying the expression flow to the target face, we can warp the target face to have a compatible shape for a larger smile. Similarly, based on the 3D alignment of the two shapes, we can compute a 3D rotation for the reference model, and project it to the image plane to form a 2D alignment flow field, which we call the *alignment flow*. Using the alignment flow, the reference face can be warped to have the same pose as the target face (see Figure 2).

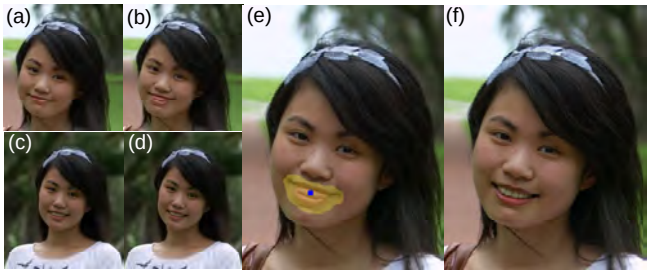

Figure 5: *Automatic crop region generation. (a) Target image. (b) Warped target. (c). Reference image. (d) Warped reference. (e) The user clicks on the mouth region (marked as blue) to specify the region to be replaced. Our system automatically generates the crop region shown as yellow. (f) Final composite after Poisson blending.*

3.6 2D Compositing

After the two input face images are warped to the desired expression and pose, our system provides a set of editing tools to assist the user to generate a high quality composite. As shown in Figure 5, our system employs an interactive feature selection tool, which allows the user to single click a facial feature to generate a crop region that is optimal for blending. This is done by employing a graph cuts image segmentation tool similar to the one proposed in the digital photomontage system [Agarwala et al. 2004]. Specifically, the data term in our graph cuts formulation encourages high gradient regions around the user selected pixels to be included in the crop region. For a pixel p, the likelihood for it being included in the crop region is defined as:

$$C(p) = \alpha \exp(-\frac{D_s(p)}{\sigma_d}) + (1-\alpha)\left(1 - \exp(-\frac{\|\nabla S(p)\|}{\sigma_s})\right), \quad (16)$$

where $D_s(p)$ is the spatial distance from p to the nearest pixel selected by the user, $\|\nabla S(p)\|$ is the gradient magnitude at p, and σ_d, σ_s and α are parameters controlling the shape and weight of each term. $L(p)$is the label of p. The data penalty in the graph cuts formulation is then defined as $C_d(p, L(p)) = 1 - C(p)$ if $L(p) = 1$ (inside the crop region), and $C_d(p, L(p)) = C(p)$ if $L(p) = 0$ (outside the crop region).

We choose to use the "match gradient" formulation in the photomontage system for setting the neighborhood penalty $C_i(p, q, L(p), L(q))$ as:

$$\|\nabla S_{L(p)}(p) - \nabla S_{L(q)}(p)\| + \|\nabla S_{L(p)}(q) - \nabla S_{L(q)}(q)\|, \quad (17)$$

which can lead to fewer blending artifacts. The total energy function which is the sum of the data and neighborhood penalty is then minimized by graph cuts optimization [Boykov et al. 2001].

Once the crop region is computed, we apply additional harmonization steps to make the cropped region more compatible with the target image. The most noticeable artifact we found is that after applying the alignment flow to warp the reference image, it becomes blurry. Blending a blurry region into a sharp image can be very noticeable. To alleviate this problem we first apply the wavelet-based detail enhancement filter proposed in [Fattal 2009] to sharpen the crop region, then blend it into the target image using the Poisson blending method [Pérez et al. 2003].

4 User Assistance

The computer vision components of our system cannot work perfectly well in all cases. For difficult examples, our system requires

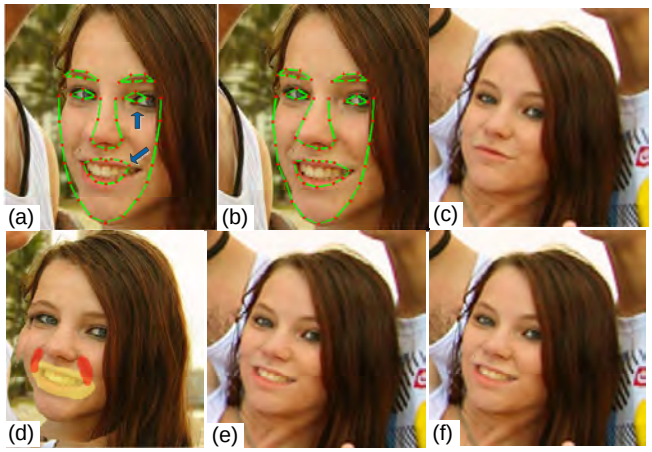

Figure 6: *User assistance modes. (a) Reference image with automatically extracted landmarks. Errors are highlighted by blue arrows. (b) Landmark locations after manual correction. (c) Target image. (d) Automatically computed crop region (yellow) with user correction (red) to add smile folds. (e) Composite without smile folds. (f) Composite with smile folds.*

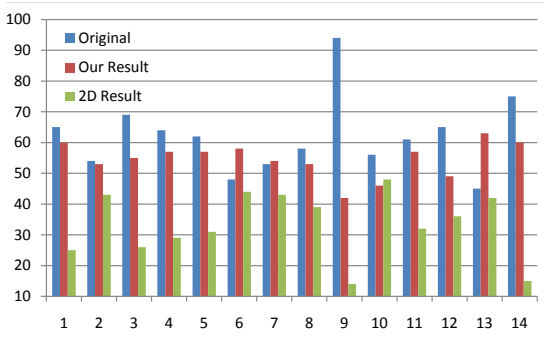

Figure 7: *User study results on comparing the original images, our results and 2D results. Vertical axis is the number of times that a specific category is voted for by the users.*

a small amount of user assistance in order to generate high quality results. The main steps that require user intervention are 2D alignment using ASM and crop region specification.

Our ASM implementation sometimes cannot generate accurate 2D alignment results for side-view faces with large rotation angles. This is a known hard computer vision problem. An example is shown in Figure 6a, where some of the automatically computed landmarks are not accurate, especially for the mouth and the left eye region. Using these landmarks for 3D fitting is then erroneous. In our system we allow the user to manually correct the bad ones, so that accurate 3D fitting can be achieved, as shown in Figure 6b.

The crop region generation tool described in Section 3.6 allows the user to quickly specify a selection mask. However, this method sometimes cannot capture the subtle semantic expression details that the user wants to transfer. Such an example is shown in Figure 6d, where the automatically generated crop region misses the unique smile folds of the subject. The user can manually add the smile folds into the crop region, which leads to a more natural composite shown in Figure 6f.

5 Results and Evaluations

5.1 User Study

To quantitatively and objectively evaluate our system, we conducted a user study using Amazon Mechanical Turk (AMT). Our evaluation dataset contains 14 examples, each including four images: two originals, the result generated by our system and the result generated by a 2D method. The 2D method first applies Lucas-Kanade image registration [Lucas and Kanade 1981] between the two faces using only pixels inside the face region, using the detected fiducial points for initialization, and then uses the rest of our system to create the final composite. This is similar to the state-of-the-art local facial component transfer approaches such as the photomontage system [Agarwala et al. 2004] and the face replacement feature in Photoshop Elements. These examples span across different age groups from small children to elders, as well as different ethnic groups, and include both men and women. For each user and each example, two images out of three (original, our result and 2D result) were randomly chosen to be shown side-by-side, and the user was asked to select the one that appears more natural. Each combination was evaluated by 50 different users, so each result was compared against the originals and the other result both for 50 times. On average the users spent 15 seconds to evaluate each pair.

Figure 7 shows the user study results. As we can see, the original images were typically rated as most natural. This is not a surprise as humans are very sensitive to the slightest imperfection on faces, and we do not expect our results to be more realistic than natural face images. Surprisingly however, in example 6, 7 and 13, our results were actually rated higher than the originals. We believe this is because our results in these examples achieved almost the same level of fidelity as the originals, and the users were essentially rating which face has a more pleasant expression when they did not see noticeable artifacts (see example 7 in Figure 8).

As the data shows, our method was consistently favored by the users against the 2D results by a significant margin, with the exception of example 10, which is an eye-replacement example (last column in Figure 8). This suggests that sometimes the 2D method is sufficient for eye replacement when the two faces have roughly the same pose, since the upper-half of the face is more rigid and opening or closing the eyes may not involve any significant global change to the face. The expression flow is insignificant in this case.

Some examples used in the user study are shown in Figure 1, 5, 6 and 8. All other examples are included in the supplementary material, which is downloadable at: http://juew.org/projects/expressionflow.htm.

To further evaluate the effectiveness of the proposed expression flow, we conducted another user study where we only compare our results against those generated by disabling expression flow on the target image. Since 3D alignment is still applied, these results are more natural than the 2D results. We chose 6 examples on which our method were rated significantly better than 2D method, and conducted a second round side-by-side comparison on AMT. Each pair was evaluated by 100 users. The results are shown in Figure 9. This study clearly suggests that the users consistently favored results with expression flow being applied.

5.2 Comparison with General Face Modeller

There are existing general face modellers that can construct a 3D face model from an input image. One may wonder if they can be applied to build 3D models for computing the expression flow, instead of using the 3D fitting method proposed in Section 3.4. To

Figure 8: *Example 7, 13, 11, 10 used in the user study. For each example, top row: target image (left) and after being warped by the expression flow (right); second row: reference image (left) and after being warped by the alignment flow (right); third row: our result; last row: 2D result.*

test this idea we applied two single image 3D fitting methods, the popular FaceGen Modeller [Singular Inversions Inc. 2009] and Algorithm 1 proposed in this paper applied to each of the faces in our examples separately, as shown in Figure 10. Note that the difference flow computed using single image fitting significantly distorts the faces, and the final composites are much worse than our results shown in Figure 1 and 8. This is because single image fitting methods will vary all possible internal parameters to best fit a 3D model for a face image, thus the two 3D models contain not only expression difference, but also *identity difference*. Using this difference flow to warp the face will lead to significant artifacts.

In Figure 10d we also show comparison results by replacing the whole target face with the 3D-corrected reference face generated by our system, inspired by the face replacement system of [Bitouk et al. 2008]. Note the various artifacts around the hair region in the bottom example as whole face replacing cannot deal with occlusions properly, and the changed gaze direction in the top example. This suggests that whole face compositing is not always reliable nor is it always desirable. Local component transfer is preferable in many cases.

5.3 Apply Expression Flow Only

Instead of being applied for transferring facial components from one image to another, the expression flow can also be used directly for expression enhancement that does not involve large topology changes, e.g., opening a mouth. Figure 11 shows two such examples, where the target face has a neutral or slight smile expression and the reference face has a large smile. In this case the computed expression flow accurately captures the characteristics of the person's smile, thus can be applied on the target image for smile enhancement. Since no blending is applied, these results have very high fidelity.

Note that previous expression mapping techniques, such as the expression ratio image [Liu et al. 2001], cannot be directly applied in these cases due to the 3D pose difference between input face images.

5.4 Apply Alignment Flow Only

In some cases the user may only want to transfer a local component without modifying the other correlated ones. As the example shown in Figure 12, one may only want to copy the eyebrows and wrinkles

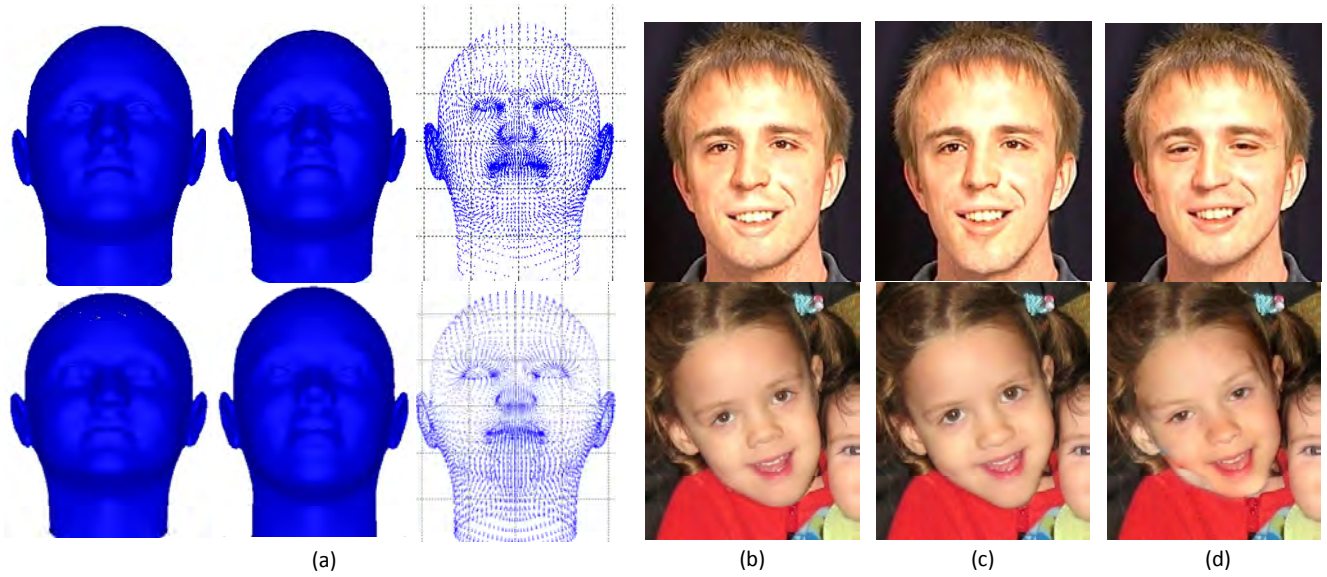

Figure 10: *Comparisons with other methods. (a). 3D Models and the difference flows generated by FaceGen Modeller. (b). Composites generated using FaceGen models. (c). Composites generated using the single image fitting algorithm (Algorithm 1). (d). Whole face replacement results. Compared with our results in Figure 1 and 8, these results contain more severe artifacts.*

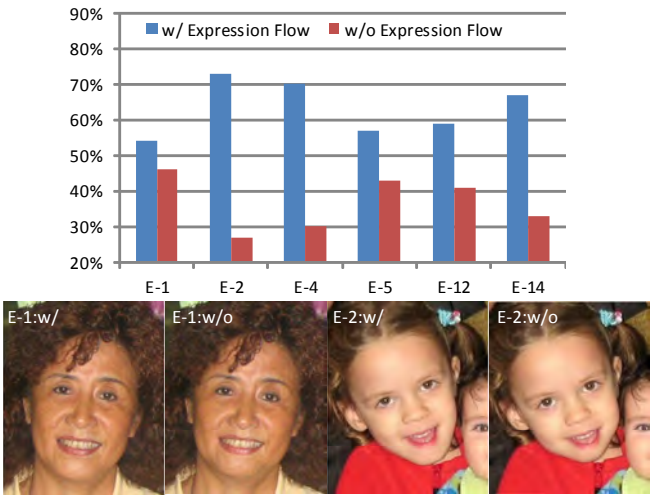

Figure 9: *Top: user study results on comparing results with and without applying expression flow. Vertical axis is the percentage of users favoring results with expression flow. Bottom: Examples with the most significant (E-2) and insignificant (E-1) differences.*

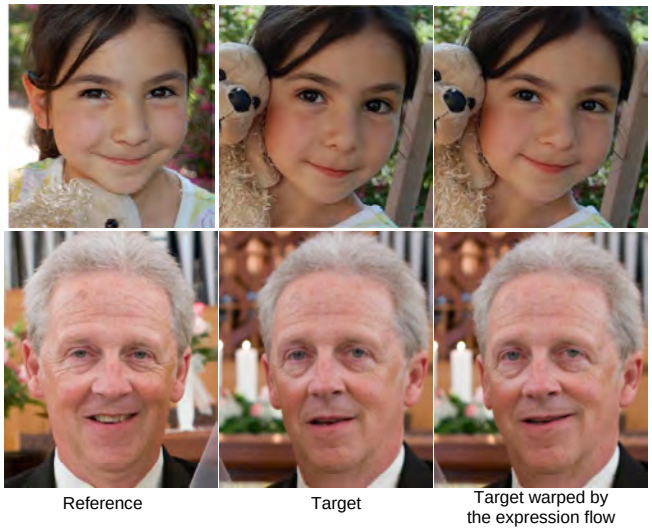

Figure 11: *Using expression flow for expression transfer only.*

from the reference face to the target face to make the expression more expressive. In this case we can disable the expression flow, and only apply the alignment flow computed from the 3D models to the reference face. Compared with the 2D result, our composite is more realistic since the correct 3D transformation has been applied to the eyebrows and wrinkles.

5.5 Other Expressions

Although most examples we have shown aim at changing a non-smiling face to a smiling one, our system can handle other less common expressions within the range of expressions in our training data set. Figure 13 shows two examples where we change a neutral face to a surprise expression. Note how the eyes of the person changes along with the expression in the top example, and the large 3D pose difference between two input images in the bottom example.

5.6 Failure Cases

Our system does not always generate realistic composites. In general, if the input faces have a very large pose difference, then the face region cropped from the reference image has to go through a large geometric transformation before being composed onto the target image. Artifacts are likely to be introduced in this process. Furthermore, in difficult cases our 3D fitting may contain errors, which will lead to inaccurate transformation. Figure 14 (top) shows one such example. Our system failed to compute an accurate 3D trans-

Figure 12: *An example of creating composite without applying expression flow. Eyebrows and wrinkles are transferred from the reference to the target image. Note the right eyebrow in the result images.*

Figure 13: *Changing a neutral expression to a surprise expression.*

formation for the mouth region, thus in the result the mouth region is clearly not compatible with the pose and shape of the target face, although our result is still significantly better than the 2D result. To avoid this problem one can find another reference image where the face pose is closer to that of the target face. This is not a problem if a large personal photo album of the subject is available.

Figure 14(bottom) shows another limitation of our system on handling asymmetric expressions. For this expression transfer example, our system cannot raise one eyebrow and at the same time lower the other one. This is because our training dataset contains only symmetric expressions. Using a richer training dataset will help in this case.

5.7 Computational Efficiency

We found that the iterative joint fitting algorithm (Algorithm 2) converges fast in practice. In our experiments we use only 10 iterations, each being a closed-form solution. In our Matlab implementation, it takes less than one second to run the algorithm. Our approach is also robust to local minima for two reasons. First, although we use an alternating minimization approach, in each stage we have a closed-form solution to efficiently minimize the energy in that stage. Second, we found that our initialization by aligning the 3D shapes to the images using internal facial landmarks, is accurate enough and leads to a quick convergence in all our examples.

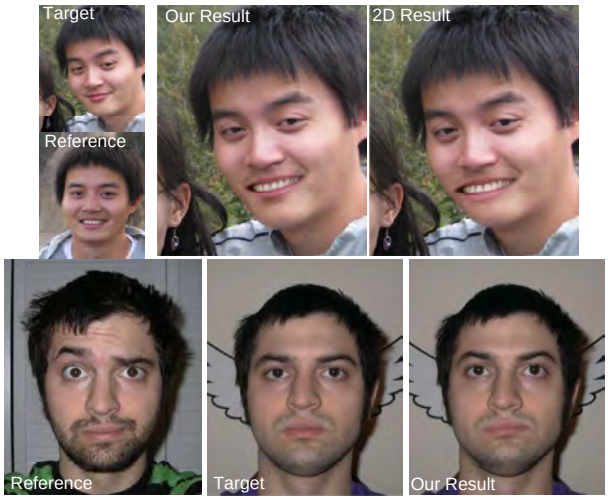

Figure 14: *Two failure examples. Top: an unsuccessful compositing. Bottom: an unsuccessful expression transfer.*

6 Conclusion and Future Work

In this paper we address the problem of transferring local facial components between face images with different expressions of the same person. To account for the expression difference, we propose a novel expression flow, a 2D flow field which can warp the target face in a natural way so that the warped face becomes compatible with the new component to be blended in. The expression flow is computed from a novel joint 3D fitting method, which jointly reconstructs 3D face shapes from the two input images, so that the identity difference between them is minimized, and only expression difference exists. A comprehensive user study was conducted to demonstrate the effectiveness of our system.

Currently our system relies on the user to provide the reference image for improving the target image. In the future we plan to develop a reference image search tool, which can automatically identify good reference images to use given the target image, in the personal photo album of the subject. This will greatly improve the efficiency of the personal face editing workflow.

Our system currently uses the face expression dataset collected by Vlasic et al. [2005]. While we demonstrate in this paper that our system can work reliably well on a wide variety of people of different races, ages and genders, we are also aware that the dataset is not rich enough to handle all possible expressions, especially asymmetric ones. As future work we plan to use existing 3D face capturing methods [Zhang et al. 2004; Wang et al. 2004] to capture more data to enrich our dataset, and explore whether it can boost the performance of the system.

As pointed out by some user study subjects, some of our results still contain minor but noticeable photometric artifacts. For instance, some subjects pointed out that the mouth region shown in Example 14, Figure 8 is grainy. While fixing these blending artifacts is not the main focus of this paper, we plan to incorporate more advanced harmonization methods [Sunkavalli et al. 2010] into our system to further improve the quality of the final results.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful feedback. Fei Yang and Dimitris Metaxas are partially supported

by the following grants to Dimitris Metaxas: NSF-0549115, NSF-0725607, NASA-NSBRI-NBTS01601, NASA-NSBRI-NBTS004, ONR-N000140910104 and Adobe systems.

References

AGARWALA, A., DONTCHEVA, M., AGRAWALA, M., DRUCKER, S., COLBURN, A., CURLESS, B., SALESIN, D., AND COHEN, M. 2004. Interactive digital photomontage. In *Proceedings of ACM SIGGRAPH*, vol. 23, 294–302.

BITOUK, D., KUMAR, N., DHILLON, S., BELHUMEUR, P., AND NAYAR, S. K. 2008. Face swapping: automatically replacing faces in photographs. In *Proceedings of ACM SIGGRAPH*, 39:1–39:8.

BLANZ, V., AND VETTER, T. 1999. A morphable model for the synthesis of 3d faces. In *Proceedings of ACM SIGGRAPH*, 187–194.

BLANZ, V., BASSO, C., POGGIO, T., AND VETTER. 2003. Reanimating faces in images and video. *Computer Graphics Forum 22*, 3.

BLANZ, V., SCHERBAUM, K., VETTER, T., AND SEIDEL, H.-P. 2004. Exchanging faces in images. *Computer Graphics Forum 23*, 3, 669–676.

BOYKOV, Y., VEKSLER, O., AND ZABIH, R. 2001. Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Analysis and Machine Intelligence 23*.

COOTES, T., TAYLOR, C., COOPER, D., AND GRAHAM, J. 1995. Active shape models: Their training and application. *Computer Vision and Image Understanding 61*, 1, 38–59.

COOTES, T. F., EDWARDS, G. J., AND TAYLOR, C. J. 2001. Active appearance models. *IEEE Trans. Pattern Analysis and Machine Intelligence 23*, 681–685.

DECARLO, D., AND METAXAS, D. 2000. Optical flow constraints on deformable models with applications to face tracking. *Int. J. Comput. Vision 38*, 99–127.

DOVGARD, R., AND BASRI, R. 2004. Statistical symmetric shape from shading for 3d structure recovery of faces. In *Proceedings of ECCV*, 99–113.

FAIGIN, G. 1991. *The Artist's Complete Guide to Facial Expression*. Watson-Guptill Publications Inc., New York.

FARBMAN, Z., HOFFER, G., LIPMAN, Y., COHEN-OR, D., AND LISCHINSKI, D. 2009. Coordinates for instant image cloning. In *Proceedings of ACM SIGGRAPH*, vol. 28, 67:1–67:9.

FATTAL, R. 2009. Edge-avoiding wavelets and their applications. In *Proceedings of ACM SIGGRAPH*, vol. 28.

FELZENSZWALB, P. F., AND HUTTENLOCHER, D. P. 2005. Pictorial structures for object recognition. *International Journal of Computer Vision 61*, 55–79.

JOSHI, N., MATUSIK, W., ADELSON, E. H., AND KRIEGMAN, D. J. 2010. Personal photo enhancement using example images. *ACM Trans. Graphics 29*, 12:1–12:15.

LEYVAND, T., COHEN-OR, D., DROR, G., AND LISCHINSKI, D. 2008. Data-driven enhancement of facial attractiveness. In *Proceedings of ACM SIGGRAPH*, 38:1–38:9.

LIANG, L., XIAO, R., WEN, F., AND SUN, J. 2008. Face alignment via component-based discriminative search. In *Proceedings of ECCV*.

LIU, Z., SHAN, Y., AND ZHANG, Z. 2001. Expressive expression mapping with ratio images. In *Proceedings of ACM SIGGRAPH*, 271–276.

LIU, C., SHUM, H.-Y., AND FREEMAN, W. T. 2007. Face hallucination: Theory and practice. *International Journal of Computer Vision 75*, 115–134.

LUCAS, B. D., AND KANADE, T. 1981. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 1981 DARPA Image Understanding Workshop*, 121–130.

MILBORROW, S., AND NICOLLS, F. 2008. Locating facial features with an extended active shape model. In *Proceedings of ECCV*, 504–513.

PÉREZ, P., GANGNET, M., AND BLAKE, A. 2003. Poisson image editing. In *Proceedings of ACM SIGGRAPH*, 313–318.

PIGHIN, F., HECKER, J., LISCHINSKI, D., SZELISKI, R., AND SALESIN, D. H. 1998. Synthesizing realistic facial expressions from photographs. In *Proceedings of ACM SIGGRAPH*, 75–84.

ROMDHANI, S., AND VETTER, T. 2003. Efficient, robust and accurate fitting of a 3d morphable model. In *Proceedings of ICCV*, 59–66.

SARAGIH, J., LUCEY, S., AND COHN, J. 2009. Face alignment through subspace constrained mean-shifts. In *Proceedings of the 12th ICCV*, 1034–1041.

SHLIZERMAN, I. K., AND BASRI, R. 2011. 3d face reconstruction from a single image using a single reference face shape. *IEEE Trans. Pattern Analysis and Machine Intelligence 33*, 394–405.

SHLIZERMAN, I. K., SANKAR, A., SHECHTMAN, E., AND SEITZ, S. M. 2010. Being john malkovich. In *Proceedings of ECCV*, 341–353.

SINGULAR INVERSIONS INC. 2009. Facegen modeller manual. In *www.facegen.com*.

SUNKAVALLI, K., JOHNSON, M. K., MATUSIK, W., AND PFISTER, H. 2010. Multi-scale image harmonization. In *Proceedings of ACM SIGGRAPH*, vol. 29.

VLASIC, D., BRAND, M., PFISTER, H., AND POPOVIĆ, J. 2005. Face transfer with multilinear models. In *Proceedings of ACM SIGGRAPH*, vol. 24, 426–433.

WANG, Y., HUANG, X., LEE, C.-S., ZHANG, S., LI, Z., SAMARAS, D., METAXAS, D., ELGAMMAL, A., AND HUANG, P. 2004. High resolution acquisition, learning and transfer of dynamic 3-d facial expressions. In *Proceedings of EuroGraphics*, 677–686.

WILLIAMS, L. 1990. Performance-driven facial animation. In *Proceedings of ACM SIGGRAPH*, vol. 24, 235–242.

ZHANG, L., SNAVELY, N., CURLESS, B., AND SEITZ, S. M. 2004. Spacetime faces: High-resolution capture for modeling and animation. In *Proceedings of ACM SIGGRAPH*, 548–558.