

# 3D Facial Geometry Recovery via Group-wise Optical Flow

Hui Fang<sup>1</sup>, Nicholas Costen<sup>1</sup>, David Cristinacce<sup>2</sup>, John Darby<sup>1</sup>

<sup>1</sup>Department of Computing and Mathematics,  
Manchester Metropolitan University, U.K.

<sup>2</sup>Imaging Science and Biomedical Engineering,  
University of Manchester, U.K.

h.fang, n.costen, j.darby@mmu.ac.uk, david.cristinacce@manchester.ac.uk

## Abstract

*We describe an algorithm for automatically finding correspondences from face video sequences. This method is useful to many applications such as face tracking, face modeling and 3D face recovery. Given a sequence of images, the face feature points are tracked by a model-constraint optical flow algorithm. By employing a Minimum Description Length (MDL) point-refinement framework, the drift-off error caused by the optical flow algorithm can be reduced and the correspondences can be matched robustly by optimizing the statistical model. As a result, the face is able to be tracked precisely. Furthermore, it offers a new method of building an appearance model automatically. The objective root mean square error (RMSE) is used to prove the efficiency of the algorithm. At the same time, the performance is evaluated subjectively by generating 3D face models based upon it.*

## 1. Introduction

3D face modeling [4] has recently been recognized as a promising technique for improving face modeling and recognition. Thereby, the structure from motion (SfM) algorithm [15, 16] has become a popular method to recover 3D face shape from monocular video sequences, due to the economical and widespread nature of the acquisition devices. However, although there are a range of SfM algorithm variants, for all options the correspondence matches between facial locations is the most important and vital step. In this paper, a new technique is proposed for automatically matching correspondences across a monocular video sequence, allowing the reconstruction of 3D face models.

One common algorithm for image registration is optical flow [1]; this makes a non-rigid alignment between a pair of images. Although optical flow has been used by many researchers to find correspondences for building models and

reconstructing 3D faces [6, 14], it is not particularly robust to noise variation. In particular, the drift-off problem means that optical flow will accumulate errors as it tracks a set of salient points in a sequences of images. If the smoothness term is too strong, it makes the tracking points flow away as well. Therefore, it is necessary to include a group-wise registration step to refine the locations across the sequence and so reduce the error.

Statistical model-based algorithms [7, 8] provide good methods of aligning the faces and finding the correspondences. However, these usually require a labeled training set to allow initial construction of the models; building such a set is time-consuming and error-prone. This combination creates a chicken-and-egg problem between building models and matching correspondences. Another difficulty with model-based algorithms is the optimization procedure. Most model-based algorithms use the gradient descent line search algorithm, which needs a rough alignment as a searching start point to avoid being trapped in a local minimal value. This further enhances the chicken-and-egg nature of the problem.

This paper develops a method of integrating the strengths of the optical flow and model-based algorithms. Optical flow gives good initial correspondences for the model-based method and the statistical model provides an extra constraint to ensure the optical flow works robustly. Inspired by the Minimum Description Length (MDL) framework [10, 11], a model-based registration approach is combined with an optical flow algorithm [5], tracking the correspondences and simultaneously building the statistical model.

This paper uses a sparse optical flow method which has similarities with feature tracking algorithms [17]. Feature based tracking algorithms pick up a number of salient locations which are strong corner points or have sufficiently distinctive texture, before matching the correspondences. However, some edge points, for example those on the contour of faces, or at weakly textured points, for example the nose tip, are very important for effective recovery of the 3D

topology. Therefore a consistent set of topologically salient points (based on those used elsewhere [7]) is defined before tracking them.

We emphasize the advantages of the proposed algorithm as: (1) the effective initialization of the optical flow algorithm is used to deal with the manual labeling and local minimization trapping problems of the model-based algorithm; (2) the model-based algorithm can tackle the drift-off error problem of optical flow, and makes it possible to change the group-wise registration algorithm to a potentially real-time application. The paper is organized as follows: Section 2 introduces some related work and background. Section 3 describes details of our proposed algorithm. Section 4 gives some performance evaluations based on experiments. Conclusions are drawn in Section 5.

## 2. Background

The central aim of the correspondence matching algorithm is to minimize the energy objective function between the aligned images. Here the transformation  $T_i : x \mapsto x'$  maps the target image to the reference image.

A large number of studies have considered methods of matching the correspondences automatically. Typically, non-rigid alignment techniques, including grid-based diffeomorphisms [9], radial basis function (RBF) warper [2] and optical flow are designed to find improved transformation functions. Cootes *et. al.* [9] represent the warp field by deformation of a grid of control points. As the control points increase in number, the warp field has more degrees of freedom to generate flexible deformations. Bartoli *et al.* [2] describe a method of combining a linear affine transform and a nonlinear RBF part which includes a sum of weighted terms. The coefficients of the basis function are applied to the distance between the warped points and the control centers. Optical flow uses the Taylor Expansion to linearize the energy function for solving the displacement at the position of each pixel. Therefore, it can generate a dense non-rigid mapping between a pair of images. Compared with other correspondences matching techniques, optical flow exploits the characteristics of temporal consistency between adjacent image-frames to obtain an approximately linearized energy function which reduces the computation burden of finding the correspondences.

In the last decade, a large number of new concepts have been proposed to deal with the problems of optical flow. A multi-resolution strategy [3] was developed to track large displacements. Subspace constraints [13] have been used to deal with discontinuities and outliers of flows. Robust energy functions have been investigated [5], involving the combination of a brightness constancy assumption, a gradient constancy assumption and a discontinuity regularizer. This paper proposes a model-constrained algorithm to reduce the influence of the drift-off problem.

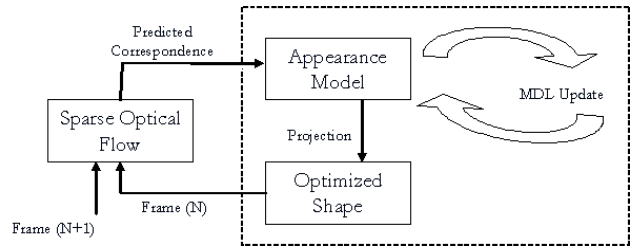


Figure 1. Illustration of the Groupwise Optical Flow algorithm

The differences from these similar studies should be emphasized. Multi-frame optic flow estimation [13] presents subspace constraints to enhance the robustness of optical flow. It assumes an affine displacement model to impose constraints on the flow domain. Although the algorithm removes the discontinuities and outliers from the optical flow, it still suffers from drift-off errors. In this algorithm, an appearance model is built in the image domain which adds high-level semantic information to the model constraints. Alternatively, group-wise construction of the appearance model [10] builds a leave-one-out model and iteratively updates the correspondences. The model is built based on the initialization of a pairwise step. This current work differs from the others in several aspects, including the use of the temporal relationship between the frames and updating of the model when the correspondences are matched by the MDL algorithm.

## 3. Proposed Algorithm

We seek to reduce the drift-off of the optical flow by fitting the correspondences into an appearance model. This model is robust, being updated from frame to frame. The description length of the parameters encoded by the iteratively upgraded model is then optimized to refine the salient points and find the best match. Thus the proposed algorithm can be regarded as an integration of optical flow estimation and statistical appearance model method. The outline of this algorithm can be briefly summarized in figure 1.

### 3.1. Optical Flow

The principle of optical flow estimation is based on an assumption that the intensity of a pixel is not changed by a relatively small displacement in the video sequence,

$$I(x, y, t) = I(x + u, y + v, t + \delta t) \quad (1)$$

where  $I(x, y, t)$  is the intensity value  $I$  at image position  $(x, y)$  in the  $t^{th}$  frame. The values of  $(u, v)$  represent the displacement between an image at the  $t^{th}$  frame and  $(t + \delta t)^{th}$  frame. By using first order Taylor Expansion to ignore the higher order terms, the equation can be linearized as

$$I_x u + I_y v + I_t = 0 \quad (2)$$

where subscripts are partial derivatives and  $u$  and  $v$  represent displacements in the horizontal and vertical directions. This linearized procedure makes optical flow more efficient than many correspondence matching algorithms.

Brox *et al.* [5] improved the optical flow algorithm by integrating a large number of techniques including a coarse-to-fine strategy, and redesign of the constancy function and Taylor Expansion, based on the hierarchical structure of the optic flow. All these changes create a more accurate and robust optical flow estimation compared with many other optic flow algorithms. Because of these advantages, it is used for the initial pairwise registration step in this paper.

In their method, the assumptions of gray value constancy, gradient constancy and smoothness are combined to make the constancy function. The cost function is shown as

$$\begin{aligned}
 E(u, v) = & \int_{\Omega} \Psi(\|I(\mathbf{x} + \mathbf{d}) - I(\mathbf{x})\| \\
 & + \gamma \|\nabla I(\mathbf{x} + \mathbf{d}) - \nabla I(\mathbf{x})\|) d\mathbf{x} \\
 & + \alpha \int_{\Omega} \Psi(\|\nabla u\| + \|\nabla v\|) d\mathbf{x}
 \end{aligned} \quad (3)$$

where  $\mathbf{d}$ , which is a vector of  $(u, v)$ , represents the displacements of the pixels;  $I(\mathbf{x})$  represents the intensity of the pixels at  $\mathbf{x}$ ;  $\nabla I(\mathbf{x})$  is the gradients of the pixels at  $\mathbf{x}$ ;  $\nabla u$  and  $\nabla v$  represent the gradient of the flow and  $\alpha$  and  $\gamma$  are two weight parameters to control the contributions of the flow gradient and the image gradient, and have standard values [5]. The function  $\Psi$  is set at  $\Psi(a^2) = \sqrt{a^2 + \epsilon^2}$  to optimize the minimization. In order to solve the non-linear cost functions and get the solution of displacements, Euler-Lagrange equations and Taylor Expansions are used to linearize the cost functions. Finally, the displacements can be solved with numerical methods such as Successive Over Relaxation (SOR) Method iterations.

This optical flow algorithm, just like most other optic algorithms, still has a major drawback in the drift-off problem. This problem is crucial when the SfM algorithm recovers the 3D shape using the optical flow correspondences. Therefore MDL model constraints are designed to prevent the salient points drifting away.

### 3.2. Updated appearance model

The statistical relations are encoded using an Appearance Model [7]. The variability within an object class can be interpreted using this model. An Appearance Model is normally computed from a training set consisting of images and manually marked salient points. When the shape is defined by a Delauney triangulated mesh, all the images are warped by a piece-wise affine transform to a base mesh shape. The model is then obtained by applying Principal Component Analysis (PCA). Thus the shape and appearance of an instance can be expressed as a linear combination

of a set of eigenvectors determined by parameters  $c$  as

$$\begin{aligned}
 \mathbf{s} &= \bar{\mathbf{s}} + \mathbf{Q}_s \mathbf{c}_s \\
 \mathbf{t} &= \bar{\mathbf{t}} + \mathbf{Q}_t \mathbf{c}_t
 \end{aligned} \quad (4)$$

where  $\bar{\mathbf{s}}$  is the mean shape,  $\bar{\mathbf{t}}$  is the mean texture and  $\mathbf{Q}_s$  and  $\mathbf{Q}_t$  represent the eigen-shape and eigen-texture respectively.

A fundamental problem in the application of this model is that it requires a manual identification of correspondences and relatively good starting parameters for the searching scheme. In this algorithm, this model is generated automatically and updated iteratively from frame to frame. Instead of labeling a training set for building the model, optical flow offers a solution to define the salient points in a video sequence and allow a good initialization if the drift-off error can be kept in a reasonable range.

Optical flow algorithms exploit the local texture and temporal relation to find the correspondences but do not consider global information. This is the reason why the drift-off problem occurs. The algorithm proposed here uses a global constraint on the optical flow. When the statistical model is kept in the buffer, it corrects the error when the correspondences move far away from the model.

### 3.3. MDL Point Optimization

Appearance models can be made active by using a gradient downhill line search strategy to find the best model parameters to minimize the texture error between the synthetic face and the region in the original image thought to contain a face [7]. This is not possible here, as there is only a bootstrapped model which is not fully trained. This implies  $\bar{\mathbf{s}}$ ,  $\bar{\mathbf{t}}$ ,  $\mathbf{Q}_s$  and  $\mathbf{Q}_t$  in Equation 4 are not constant, as in [7], but alter and grow with the sequence. It is necessary to both update all of them during the face tracking and simultaneously use them to constrain the optical flow, which here takes on the active part of the algorithm.

Suppose that whatever model is currently available is used to encode the face in the next consecutive frame. If the correspondences between the actual points and their ideal locations were perfectly matched, the coding cost of this face would be minimal, based on the MDL framework. This decomposes into four facets; the code lengths of the shape parameters, the texture parameters, the residues of shape and the texture errors of this instance. Together these can be used to calculate the description length and so the compactness and accuracy of the model. The length of each facet is related to the log of probability based on the entropy prin-

ciple,

$$E = \lambda_1 \sum_{s_m=1}^{n_1} \log P(s) + \lambda_2 \sum_{t_m=1}^{n_2} \log P(t) \quad (5)$$

$$+ \lambda_3 \sum_{d_m=1}^{n_3} \log P(d) + \lambda_4 \sum_{e_m=1}^{n_4} \log P(e)$$

where  $\lambda_i$  are weights to balance the complexity of the shape and texture, and  $n_i$  represents the number of shape or texture parameters, control points or pixels in the normalized shape respectively.  $P(x)$  is the probability of  $x$  where  $x$  is any of the parameters. It is assumed that  $P(x)$  obeys an exponential distribution, which is robust to outliers [10]. This group-wise step can be used to minimize this energy function. In the proposed algorithm, the search step directly optimizes all the feature points in the frame under consideration, using MDL as an objective function. Steepest gradient descent search is used, starting from the feature point locations provided by the optical flow. This maximizes the overlap between the model and new frame, while still allowing for the presence of un-encodable aspects of the face.

## 4. Experimental Results

The algorithm was used to track two faces in two down-sampled sequences (every 10<sup>th</sup> frame, 60 and 50 frames for individuals respectively), which are known to yield several singular columns if the SfM algorithm is used to recover the 3D shape via rank constraints. One sequence was from FG-NET [10] and the other, shorter sequence is from a new data-set [12]. These sequences had previously been manually annotated with a consistent set of 68 points. Movements are relatively small but plastic ensuring that all points were always visible although there are some occlusions existing in FG-NET sequence. To remove confounding effects, in all cases, the point-locations on the first frame were found manually. This could be replaced with a generic facial interpretation algorithm, such as an Active Appearance Model [7].

### 4.1. Objective Evaluation

To evaluate the proposed algorithm, the criteria used was the distance of the points between the correspondences found by the algorithms and manually labeled ground truth. The distance metric was root of mean square error (RMSE)

$$D_m = \sqrt{\frac{\sum_{i=1}^n d_i^2}{n}} \quad (6)$$

where  $d_i^2$  is the Euclidean point to point error, and  $n$  is the number of salient points.

The performance was compared with the pair-wise optical flow algorithm [5] and group-wise piece-wise affine reg-

istration [10]. Results of applying these methods are shown in Figure 2 to track the second of the test sequences.

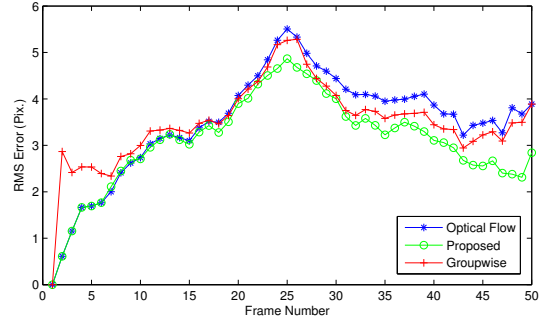


Figure 2. The objective comparison between different automatic registration algorithms. The peak error at frame 25 corresponds with the point of maximum facial rotation.

The optical flow algorithm accumulates errors due to the drift-off problem. Thus the error increases and fluctuates at a relative high level. It will further increase when the face is tracked in subsequent frames. The group-wise algorithm uses the MDL framework to update the correspondences in each frame, based on a leave-one-out model. However, incorrect information is encoded in the model because of errors in the pairwise step. While it improves the performance of some frames, the error in the other frames increases. The proposed algorithm gives the consistently best result. As the model expands, it offers an increasingly strong constraint on the correspondences and the error drops abruptly. The texture model modes (shown in figure 3) grow fast in approximately the first 25 frames and then become more stable (this corresponds with the point of maximum rotation of the face). Thereafter, the model constraints start working more robustly. Simultaneously, the shape model modes increase, which suggests that new shape instances are added into the shape model, and so more shape variation is obtained. The pattern of results for the FG-NET sequence shows a similar relationship with the facial behaviour.

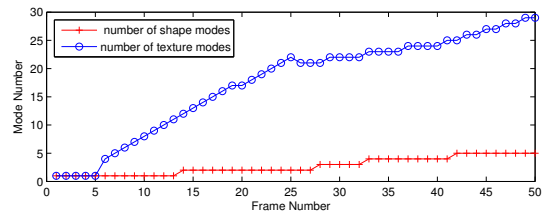


Figure 3. The number of modes in the shape and texture models.

The robustness of the algorithm is demonstrated by the average errors in table 1 tracking through the FG-NET sequence. It can be seen that the mean error of the proposed algorithm is the lowest, suggesting its performance is

more stable than the other two algorithms. Applying within-subject T-tests to the three sets of data shows that the difference in means between proposed algorithm and optic flow alone is significant ( $t = 9.66, d.f. = 59, P \ll 0.005$ ), as is that between the proposed algorithm and group-wise MDL ( $t = 11.35, d.f. = 59, P \ll 0.005$ ). However, the difference between optic flow and group-wise MDL is not significant ( $t = 0.78, d.f. = 59, P > 0.1$ ).

Algorithm	Error (pix)	S.D. (pix)
Optical Flow	8.3244	7.3253
Group-wise MDL	8.3022	6.0693
Proposed Algorithm	6.7769	2.9202

Table 1. Comparison of the mean error for the different algorithms.

As an application of the proposed algorithm, after the method was used to find the correspondences in the second sequence, the SfM algorithm [15] was used to recover the 3D shape of the face, expressed as a 3D triangulation. This algorithm assumes an orthographic projection model for 3D scene reconstruction. The RMS error in 3D domain was used to compare the 3D shape reconstructed by the automatic matching correspondence algorithms. The error is calculated between the 3D shape recovered by these methods and that derived from the manually labeled points. As shown in table 2, the proposed algorithm can recover a reasonable 3D face, unlike optic flow tracking or group-wise registration. The RMS error of 6.81 pixels is much lower than the error of the other automatic correspondence matching algorithms. Since each value reflects two global measures drawn from all 50 frames, it is not reasonable to calculate standard deviations. It should be noted that the manual points are not entirely error-free, and so it is not possible to state that the automatic configuration is in fact wrong, just that this method yields a solution which is more consistent than other techniques.

Algorithm	3D RMS Error (pix)
Optical Flow	18.77
Groupwise MDL	11.74
Proposed Algorithm	6.81

Table 2. Comparison of the RMS error of the reconstructed 3D shape.

## 4.2. Subjective Evaluation

In figure 4, the triangulation meshes built by the correspondences found by the proposed algorithm (left picture) and the optical flow algorithm (right picture) are shown. Some salient points such as the points at the right face contour and some points around the nose have been lost (they have drifted away) by the optical flow algorithm. Any

model built using the optical flow algorithm points will encode more wrong texture, e.g. the hair and some texture in a triangle which actually belongs to another triangle, and wrong shape variations. Under the global constraints of our algorithm, the salient points are still be tracked robustly. As a result, the model built by the proposed algorithm has greater power to represent the instances.

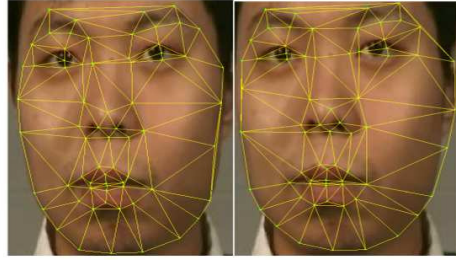


Figure 4. Comparison of the points recovered by the proposed algorithm (left) and optic flow algorithm (right) for the 50th frame.

## 4.3. Application in SfM algorithm

The rotation constraints alone are not enough to approach the correct solution in most automatic tracking cases. Shape constraints can be added via a generic shape model [12] to regularize the 3D recovered face. The RMS error between the recovered 3D shape and ground truth drops from 6.81 to 5.2 pixels. The effectively recovered 3D topology is shown in figure 5. The depth of each part of the face is subjectively correct. Finally, the mean texture of the

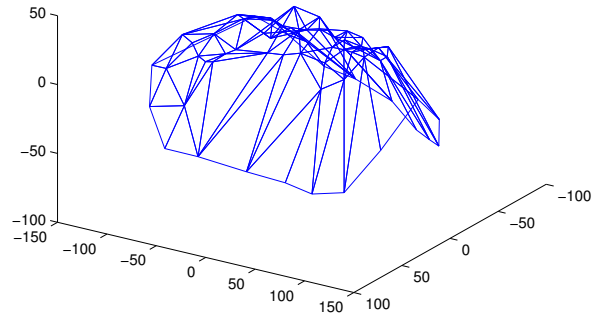


Figure 5. 3D shape recovered by the proposed algorithm.

face drawn from the normalized faces using the Delaunay triangulation built from the tracked sequence was mapped onto the 3D shape. The textured face is shown in figure 6.



Figure 6. The 3D textured face recovered by SfM using the correspondences of the proposed algorithm.

## 5. Discussion and Conclusions

In this paper we have investigated a method to integrate optical flow and model-based frameworks to find the correspondences in video sequences for reconstructing 3D faces. The disadvantages of the two frameworks are addressed by combining them into a complimentary scheme. We use high quality initialization from optical flow to deal with the manual labeling and local minimization trapping problems of the model-based algorithm. Also, the model-based algorithm tackles the drift-off error of optical flow. In addition, it is possible to use the algorithm as the basis of a real-time application, as it does not depend upon an iterative step to build the optimal model as does the group-wise registration algorithm. These correspondences then can be used to recover a 3D textured face representation.

In the future we will further test the efficiency of this work in longer sequences. Furthermore, the 3D shape will be refined by an optimization algorithm to reduce the projection error, in particular a bundle adjustment method.

## 6. Acknowledgments

We would like to thank Prof. Tim Cootes and Prof. Chris Taylor of Manchester University for providing comments and helpful suggestions. This work was supported by EPSRC grants EP/D056942 and EP/D054818.

## References

- [1] J. Barron, D. Fleete, and S. Beauchemin. Performance of optical flow techniques. *International Journal of Computer Vision*, 12(1):43–77, 1994.
- [2] A. Bartoli and A. Zisserman. Direct estimation of non-rigid registrations. In *Proc. BMVC*, pages 899–908, 2004.
- [3] J. R. Bergen, P. Anandan, K. J. Hanna, and R. Hingorani. Hierarchical model-based motion estimation. In *Proc. ECCV*, pages 237–252, 1992.
- [4] V. Blanz and T. Vetter. Face recognition based on fitting a 3D morphable model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1–12, 2003.
- [5] T. Brox, A. Bruhn, N. Papenbergh, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In *Proc. ECCV*, volume 4, pages 25–36, 2004.
- [6] A. K. R. Chowdhury and R. Chellappa. Face reconstruction from monocular video using uncertainty analysis and a generic model. *Computer Vision and Image Understanding*, 91:188–213, 2003.
- [7] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001.
- [8] T. F. Cootes and C. J. Taylor. Active shape models. In *Proc. BMVC*, pages 266–275, 1992.
- [9] T. F. Cootes, C. J. Twining, K. O. Babalola, and C. J. Taylor. Diffeomorphic statistical shape models. *Image and Vision Computing*, 26:326–332, 2008.
- [10] T. F. Cootes, C. J. Twining, V. Petrovic, R. Schestowitz, and C. J. Taylor. Groupwise construction of appearance model using piece-wise affine deformations. In *Proc. BMVC*, pages 879–888, 2005.
- [11] R. H. Davies, C. J. Twining, P. D. Allen, T. F. Cootes, and C. J. Taylor. Building optimal 2D statistical shape models. *Image and Vision Computing*, 21:1171–1182, 2003.
- [12] H. Fang and N. P. Costen. 3D face reconstruction under imperfect tracking circumstances using shape model constraints. In *Proc. AVC*, pages 519–528, 2007.
- [13] M. Irani. Multi-frame optical flow estimation using subspace constraints. In *Proc. ICCV*, volume 1, pages 626–633, 1999.
- [14] M. J. Jones and T. Poggio. Multidimensional morphable models: A framework for representing and matching object classes. *International Journal of Computer Vision*, 29(2):1–28, 1998.
- [15] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: a factorization method. *International Journal of Computer Vision*, 9(2):137–154, 1992.
- [16] L. Torresani, A. Hertzmann, and C. Bregler. Learning non-rigid 3D shape from 2D motion. In *Proc. NIPS*, pages 1555–1562, 2003.
- [17] Z. Zhang, J. Liu, D. Adler, M. F. Cohen, E. Hanson, and Y. Shan. Robust and rapid generation of animated faces from video images: A model-based modeling approach. *International Journal of Computer Vision*, 58(2):93–119, 2004.