Lightweight Binocular Facial Performance Capture under Uncontrolled Lighting

Levi Valgaerts[1] * Chenglei Wu[1,2] † Andrés Bruhn[3] ‡ Hans-Peter Seidel[1] § Christian Theobalt[1] ¶
[1]MPI for Informatics [2]Intel Visual Computing Institute [3]University of Stuttgart

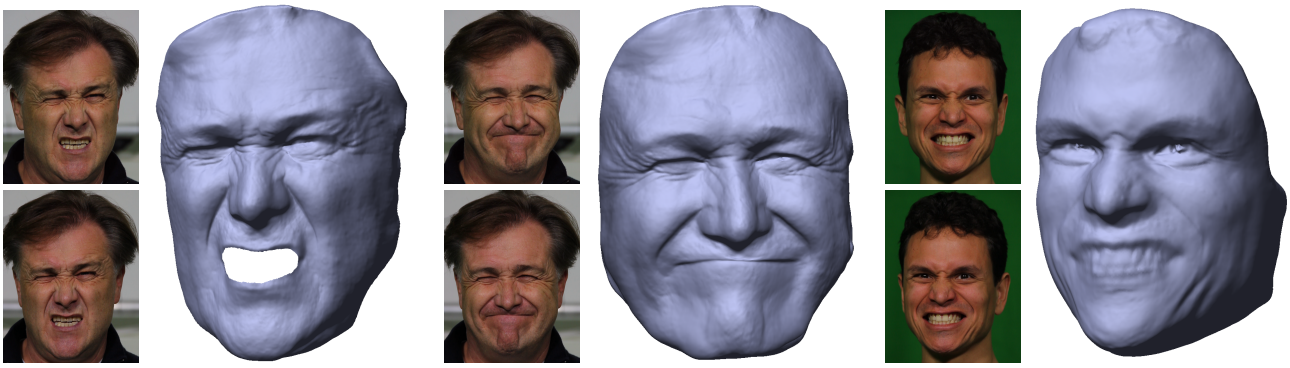

Figure 1: *Three results of our facial performance capture method for two indoor sequences with fast and expressive motion.*

Abstract

Recent progress in passive facial performance capture has shown impressively detailed results on highly articulated motion. However, most methods rely on complex multi-camera set-ups, controlled lighting or fiducial markers. This prevents them from being used in general environments, outdoor scenes, during live action on a film set, or by freelance animators and everyday users who want to capture their digital selves. In this paper, we therefore propose a lightweight passive facial performance capture approach that is able to reconstruct high-quality dynamic facial geometry from only a single pair of stereo cameras. Our method succeeds under uncontrolled and time-varying lighting, and also in outdoor scenes. Our approach builds upon and extends recent image-based scene flow computation, lighting estimation and shading-based refinement algorithms. It integrates them into a pipeline that is specifically tailored towards facial performance reconstruction from challenging binocular footage under uncontrolled lighting. In an experimental evaluation, the strong capabilities of our method become explicit: We achieve detailed and spatio-temporally coherent results for expressive facial motion in both indoor and outdoor scenes – even from low quality input images recorded with a hand-held consumer stereo camera. We believe that our approach is the first to capture facial performances of such high quality from a single stereo rig and we demonstrate that it brings facial performance capture out of the studio, into the wild, and within the reach of everybody.

CR Categories: I.3.7 [COMPUTER GRAPHICS]: Three-Dimensional Graphics and Realism; I.4.1 [IMAGE PROCESSING]: Digitization and Image Capture—Scanning; I.4.8 [IMAGE PROCESSING]: Scene Analysis;

Keywords: Facial Performance Capture, Scene Flow, Shading-based Refinement, Uncontrolled Lighting

*e-mail: valgaerts@mpi-inf.mpg.de
†e-mail: chenglei@mpi-inf.mpg.de
‡e-mail: andres.bruhn@vis.uni-stuttgart.de
§e-mail: hpseidel@mpi-inf.mpg.de
¶e-mail: theobalt@mpi-inf.mpg.de

Links: DL PDF WEB VIDEO DATA

1 Introduction

Two essential features of a realistic virtual actor are a convincingly rendered face and a convincingly animated facial performance. If virtual facial detail is not believably modeled and lit, and if facial motion and expression does not exhibit authentic high spatial and temporal detail, it will not be perceived as realistic. To meet these high quality demands, the research community has developed a variety of facial performance capture techniques that aim to reconstruct very detailed dynamic facial geometry, motion and possibly appearance from sensor measurements of real subjects.

On the one hand, there are active optical systems that use markers, active illumination or invisible paint to capture facial performance [Bickel et al. 2007; Zhang et al. 2004; Furukawa and Ponce 2009]. However, such reconstructions often lack detail and appearance capture is difficult or impossible. On the other hand, passive approaches use multiple cameras and vision-based reconstruction techniques to capture facial performance, e.g. [Bradley et al. 2010]. Reconstructions are of high quality, but pore-level detail is often missing. Moreover, accumulating drift makes it hard to capture very expressive motion. Active lighting methods can bring out pore-level shape detail, but the price to be paid is a complex controlled light and camera set-up [Vogiatzis and Hernández 2011; Wilson et al. 2010]. In other words, to capture facial performance with high-quality spatial and temporal detail, current state-of-the-art techniques require a large number of cameras in a controlled indoor environment, possibly actively controlled illumination, and in many cases some form of active interference with the scene.

ACM Reference Format
Valgaerts, L., Wu, C., Bruhn, A., Seidel, H., Theobalt, C. 2012. Lightweight Binocular Facial Performance Capture under Uncontrolled Lighting. *ACM Trans. Graph. 31* 6, Article 187 (November 2012), 11 pages. DOI = 10.1145/2366145.2366206 http://doi.acm.org/10.1145/2366145.2366206.

These strong requirements are the reason why facial performance capture has so far mostly been a privilege of high-budget animation productions. In addition, facial performance capture has hardly been used where it would actually be most effective: in arbitrary uncontrolled settings, such as indoor and outdoor movie sets where the actors perform in their natural environment. Also, in the light of an ever growing amount of existing stereo movie footage, a method that makes use of a small number of cameras could serve as a cornerstone for new movie-production applications, such as facial performance capture from the principal stereo camera feed. In a much broader scope, simple stereo camera systems will soon be available on many hand-held consumer devices. It is clear that these devices will not be equipped with complex multi-camera systems, nor will they be used within the controlled confines of a studio. All this shows that there is a clear use case for a performance capture method that succeeds with a minimum number of cameras and under general uncontrolled lighting, thereby opening up the path towards facial performance capture "in the wild".

In this paper, we take a step towards this goal by proposing a new image-based facial performance capture approach that uses an extremely simple acquisition set-up: *a single stereo pair of video cameras*. It captures performance under uncontrolled and changing lighting conditions, either indoors or outdoors. From the stereo data, our approach reconstructs facial performance data whose quality comes close to much more complex studio-based approaches – they exhibit both high spatial detail, i.e. creases and folds in the face, as well as high temporal detail, i.e. accurate facial motion. The main contributions of our work are the effective combination and adaptation of a variety of image-based reconstruction and tracking approaches: 1) a robust variational stereo reconstruction and scene flow method for coarse reconstruction and correspondence finding specifically tailored to face capture, 2) a robust template tracking approach with an active scene flow-based motion refinement mechanism and 3) an adapted spatio-temporal probabilistic framework that estimates incident illumination and face albedo and refines coarse face geometry using shading information.

Our algorithm succeeds under uncontrolled and time-varying illumination, allows both the performer and the stereo pair of cameras to move independently, and yields detailed 3D facial performance geometry that is fully spatio-temporally coherent, even when performers make very expressive faces. We will show highly detailed 3D facial performance results, optionally with texture, for two different camera systems: high-quality results from a pair of SLR cameras capturing indoors, and results of previously unseen detail captured outdoors with a low quality consumer stereo camera.

It is important to understand that we do not claim to achieve higher reconstruction quality than state-of-the-art multi-camera approaches under controlled studio conditions. *We want to solve a different problem for which these methods may not be suitable:* To make detailed facial performance capture with cheap devices in uncontrolled environments feasible, even for inexperienced users. We show that a carefully designed reconstruction method enables this, and that purely passive capture of dynamic face geometry from a *single stereo rig* is possible at an unprecedented level of detail that comes close to state-of-the-art studio-based results.

2 Related Work

For many years, researchers in graphics and vision have investigated facial performance capture approaches that differ in the employed sensors and reconstruction techniques. Some methods solely rely on dynamic 3D shape scanner data, i.e. time-varying point clouds, and no additional input images. Anuar and Guskov [2004] track an initial template mesh from point cloud data using a purely geometric 3D scene flow method. Reconstruction of high frequency detail is difficult with their approach and the purely geometric 3D scene flow method more frequently suffers from drift. Wand et al. [2009] simultaneously build up and track a template of a face from point cloud data, but reconstructions lack some high-frequency shape detail. Popa et al. [2010] propose a similar framework that can capture more high-frequency detail by means of a change prior. Weise et al. [2011] use point clouds from a Kinect and a template with an attached blend shape model to track facial performances. However, their goal is animation transfer, not authentic reconstruction of fine-scale shape detail.

Image-based approaches help to overcome the resolution limits and the limits in tracking accuracy that purely geometric methods still have. Following the marker-based motion capture paradigm widely accepted in industry, researchers attempted to reconstruct facial performances by tracking attached or painted markers on a face with several cameras, or by tracking the distortion of an invisible paint applied to the skin [Williams 1990; Guenter et al. 1998; Furukawa and Ponce 2009; Bickel et al. 2007]. Active fiducials greatly enhance tracking accuracy and enable robust reconstruction of even extreme facial expressions. However, the resolution of the captured geometry is limited, the mark-up phase can be cumbersome, and due to the active intrusion into the scene, the simultaneous reconstruction of geometry and appearance is not feasible. Huang et al. [2011] try to overcome some of these limitations in a data-driven way by transferring geometric detail from a sparse set of 3D scans to dynamic face geometry recorded with a marker-based motion capture system.

Instead of markers, active illumination, e.g. patterns emitted from projectors, can be used to facilitate image-based face geometry reconstruction from multiple cameras [Zhang et al. 2004; Wang et al. 2004; Weise et al. 2007]. For these approaches, texture acquisition requires interleaving of pattern and texture frames, and temporal reconstruction artifacts may occur since several subsequent images are required for a single reconstruction. Also, establishing geometric correspondence between subsequent reconstructions is still a challenge. Template-based methods fit a deformable shape model to images of a face [DeCarlo and Metaxas 1996; Pighin et al. 1999; Blanz et al. 2003]. While these methods yield spatio-temporally coherent reconstructions, the captured face geometry is often coarse and lacks fine-scale detail.

High-quality facial performances were reconstructed with purely passive stereo-based approaches in combination with mesh tracking [Bradley et al. 2010]. Borshukov et al. [2003] developed the *Universal Capture* system for the movie *The Matrix*, which deforms a laser-scanned 3D facial model by using optical flow fields computed from a multi-camera system. These approaches usually require dense multi-camera set-ups and a controlled studio environment. Also, reconstructing pore-level detail is difficult from pure stereo, and temporal drift in the reconstructions often prevents capturing expressive facial motions. Beeler et al. [2011] try to overcome the drift problem in dense multi-view face reconstruction by stabilizing mesh tracking with a set of key facial poses. The commercial system by DepthAnalysis[1] also reportedly uses stereo reconstruction from a dense multi-camera system under controlled studio lighting. Stereo reconstruction is also used in the MOVA Contour system, which employs an even denser array of tens of cameras and invisible make-up to aid reconstruction[2].

Establishing 3D correspondences between subsequent image-based face reconstructions is still a challenge and drift may easily occur. Most approaches resort to a form of template mesh tracking based

[1] www.depthanalysis.com
[2] www.mova.com

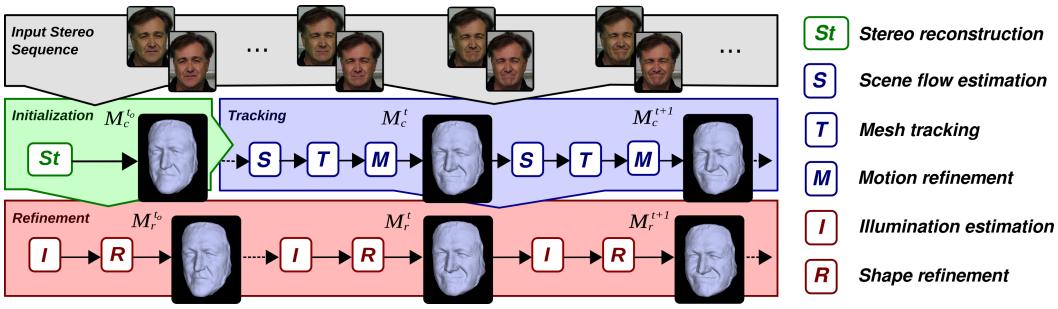

Figure 2: *Overview of our facial performance capture method with its tracking and refinement pipelines.*

on features or non-rigid registration [Anuar and Guskov 2004; Bradley et al. 2010; Beeler et al. 2011]. Recently, image-based scene flow methods have taken great strides forward. In particular, those approaches that treat the coupled problems of stereo and motion estimation jointly in a variational framework turned out to be successful [Basha et al. 2010; Valgaerts et al. 2010]. In this context, there have even been efforts to consider the estimation of temporally coherent scene flow correspondences, e.g. by constraining the temporal evolution using basis functions [Birkbeck et al. 2011]. However, most of the image-based scene flow methods focus on application scenarios that involve piecewise rigid body motion such as driver assistance systems [Wedel et al. 2008]. Reconstruction of human faces that require the estimation of heavily non-rigid body motion, in contrast, can hardly be found in the literature. Moreover, if facial reconstructions are presented, they are typically limited to details of medium scale [Valgaerts et al. 2010].

Passive acquisition of true fine scale surface detail with image-based methods is still difficult. Several approaches have recently shown that shading and reflectance effects under controlled lighting can boost reconstruction resolution dramatically. Vogiatzis and Hernández [2011] use controlled tri-colored studio illumination and a combination of multi-view stereo and photometric stereo to capture facial geometry. Combining active structured light scanning and marker-based facial performance capture with a complex light stage illumination set-up also enables high-quality capture of geometry and appearance in a studio [Alexander et al. 2009]. Light stage illumination requires recording of multi-view images under several light conditions to obtain a single reconstruction. To cope with the resulting spatio-temporal alignment problem in the data, Wilson et al. [2010] developed an approach to establish correspondences between images taken under starkly varying spherical gradient illuminations. This enables a combination of stereo and photometric normal reconstruction in a spatio-temporal way.

On-set facial performance capture would require an algorithm to handle general uncontrolled, possibly time-varying illumination. Beeler et al. [2010] have shown that shading cues under uncontrolled illumination can be used to synthesize artificial pore-level detail in multi-view stereo reconstructions of a static face. Basri et al. [2007] introduced a photometric stereo method for general unknown illumination. Jin et al. [2008] combine multi-view stereo and shape-from-shading under an unknown point light on static scenes. Using results obtained by a full-body performance capture approach from multi-view video, Wu et al. [2011a] estimate time-varying low-frequency illumination using spherical harmonics. The lighting estimates are used to perform shading-based refinement of the coarse performance capture geometry. We follow a similar approach to process video from only two cameras.

Only a few methods focus on additionally estimating the surface reflectance. Carceroni et al. [2002] capture coarse surfel-based geometry of a moving face and reflectance estimates from multi-view video footage. Georghiades [2003] reconstructs a static face model and a coarse BRDF from multiple images under point light illumination with varying but unknown positions.

Our method takes inspiration from recent progress in the above mentioned individual domains in order to overcome several limitations of previous approaches. To the best of our knowledge our approach is the first purely passive technique that enables highly detailed and spatio-temporally coherent facial performance capture using only two cameras, while being applicable in uncontrolled or even changing lighting scenarios at the same time.

3 Our Facial Performance Capture Method

As input, our approach expects a stereo video sequence of a face captured in an uncontrolled environment. Our method is composed of two main computational pipelines (Fig. 2):

I In a first pass, we track a coarse-detail face template throughout a binocular stereo sequence. This *template tracking step* (Sec. 5) produces a sequence of coarse face meshes that are in full correspondence and exhibit minimal drift. To enable this, our approach makes use of a new highly accurate image-based scene flow method and relies on a Laplacian deformation model to regularize the moving geometry.

II In a second pass, we add fine time-varying detail, e.g. wrinkles and folds, to the tracked meshes. This *shape refinement step* (Sec. 6) exploits shading information to produce accurate surface detail under uncontrolled and changing lighting. We build upon a framework for incident lighting and albedo estimation, and contribute with a new albedo clustering approach and an improved, faster shape refinement optimization.

Thus, we capture facial performance in a coarse-to-fine manner: While the first pipeline is responsible for the recovery of coarse-scale head motion and facial deformation, the second pipeline refines the results to include fine-scale details at skin level.

In the next sections we will discuss both pipelines in detail. Henceforth, we will indicate by f_0^t the left frame of a binocular stereo sequence at time t, and by f_1^t the corresponding right frame. For any time t, we can assume that f_0^t and f_0^{t+1} (f_1^t and f_1^{t+1}) are two consecutive frames in the left (right) image sequence. We further denote by t_0 the time at which we start capturing, i.e. $(f_0^{t_0}, f_1^{t_0})$ is the first stereo pair in our tracking and refinement algorithm. A reconstructed triangular mesh at time t will be denoted by M^t and is characterized by its set of n vertices and their connecting edges. The Euclidean coordinates of a vertex at time t will be denoted by the vector $\boldsymbol{X}^t$. Our two processing pipelines reconstruct a coarse mesh M_c^t and a refined mesh M_r^t at each time step, both of which are based on the same vertex set and connectivity.

4 Initialization

We assume that the stereo camera pair is calibrated off-line (MATLAB toolbox). Our method starts from a smooth 3D reconstruction of the face that will serve as a *template mesh* for the tracking step. During mesh tracking, this template will be moved and deformed according to the detected motion in the stereo sequence.

Template Reconstruction It is assumed that the face at time t_0 is in rest. To obtain an initial 3D reconstruction from the first stereo pair $(f_0^{t_0}, f_1^{t_0})$, we apply a variant of the variational stereo method of [Valgaerts et al. 2011] for calibrated images. This method recovers the dense 2D displacement field between $f_0^{t_0}$ and $f_1^{t_0}$ by minimizing an energy of the form:

$$E = \int_\Omega \left(E_\mathrm{D} + \alpha\, E_\mathrm{G} + \beta\, E_\mathrm{S}\right) \mathrm{d}\boldsymbol{x}\ , \tag{1}$$

where E_D imposes constancy assumptions on certain image features, E_G includes knowledge about the known stereo geometry and E_S assumes the displacement field to be piecewise smooth. The exact form of these terms is given by the equations (5), (7) and (9), respectively, and for the minimization of the total energy we refer to the next section on the related problem of scene flow estimation.

Once the 2D displacement field has been recovered, the corresponding pixels can be triangulated to obtain a 3D point cloud [Hartley and Zisserman 2000]. In practice, we perform a 3D reconstruction for both pairs $(f_0^{t_0}, f_1^{t_0})$ and $(f_1^{t_0}, f_0^{t_0})$. This ensures a sufficient amount of 3D points in regions that are badly visible in just one image, such as the sides of the nose.

Postprocessing In a post processing step, the background is removed manually and the point cloud is converted to a triangular mesh [Kazhdan et al. 2006]. We set the number of vertices roughly equal to that of the pixels in the face region such that each vertex corresponds to a pixel in the input views. Finally, the mesh is smoothed [Sorkine 2005] and each vertex is assigned a fixed color using projective texturing and blending from both input views. If desired, holes can be cut in the the mesh for the mouth or the eyes.

The above steps are illustrated in Fig. 3, where we show the starting frames $f_0^{t_0}$ and $f_1^{t_0}$ of a stereo sequence, together with the obtained 3D reconstruction and the final template mesh $M_c^{t_0}$.

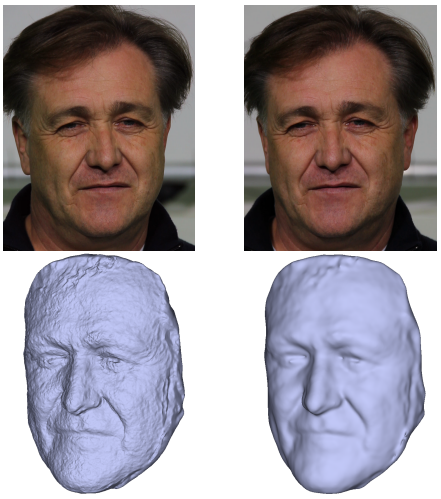

Figure 3: *Initialization. Top row: starting frames $f_0^{t_0}$ and $f_1^{t_0}$. Bottom row: stereo reconstruction and template mesh $M_c^{t_0}$.*

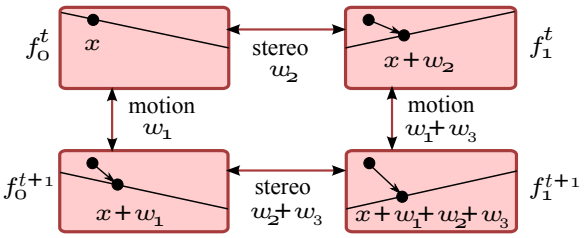

Figure 4: *The four-frame set-up for the scene flow computation.*

5 Template Tracking

The tracking step is responsible for propagating the template mesh throughout the stereo sequence. To accurately recover the motion of the face, we base our tracking on a state-of-the-art method for *scene flow* computation, which we extend with a new structure-aware regularization strategy. This method establishes a dense 3D displacement field, which is used to update the position of all the vertices in the tracked mesh from one time instance to the next. A smooth deformation of the face is obtained by regularizing the geometry of the surface via the *Laplacian operator*. Scene flow estimation is then used a second time to *refine the motion* and to minimize any reprojection error that might have been induced by the tracking.

5.1 Scene Flow Computation

To compute the scene flow between the time instances t and $t+1$, we build upon a recent variational 3D scene flow method [Valgaerts et al. 2010]. We propose an extended version of this method that assumes the stereo system to be calibrated, but does not require the recorded images to be preprocessed, e.g. rectified. In addition, we propose a novel anisotropic structure-aware regularization technique that is inspired by recent optical flow methods. This adaptive regularization strategy provides dense results of high quality and is specifically well suited for capturing facial features, such as mouth, nose, eyebrows and laugh lines.

Our scene flow method estimates a 3D reconstruction and 3D displacement field by establishing correspondences in the image domain. It is based on the four frame case depicted in Fig. 4. As one can see, all possible constraints between two consecutive stereo pairs (f_0^t, f_1^t) and (f_0^{t+1}, f_1^{t+1}) can be expressed in terms of three unknown optical flow fields: the *motion flow* $\boldsymbol{w}_1$, the *stereo flow* $\boldsymbol{w}_2$ and the *difference flow* $\boldsymbol{w}_3$. We propose to compute these flows $\boldsymbol{w}_i = (u_i, v_i)^\top$, $i = 1, 2, 3$, by minimizing an energy of the form:

$$E = \int_\Omega \Bigg(\underbrace{\sum_{i=1}^{4} E_\mathrm{D}^i}_{\text{data}} + \underbrace{\sum_{i=1}^{2} \alpha_i E_\mathrm{G}^i}_{\text{geometry}} + \underbrace{\sum_{i=1}^{3} \beta_i E_\mathrm{S}^i}_{\text{smoothness}} \Bigg) \mathrm{d}\boldsymbol{x}\ . \tag{2}$$

The four *data terms* E_D^i encode constancy assumptions between all frames, the three *smoothness terms* E_S^i assume the desired flows to be piecewise smooth and the *geometry terms* E_G^i model the geometric relations between the two stereo pairs. All deviations from model assumptions are weighted by positive weights α_i and β_i and are integrated over the rectangular image domain Ω of the reference frame $f_0^t(\boldsymbol{x})$, $\boldsymbol{x} = (x, y)^\top$. We now discuss these terms in detail.

Data Terms For the data constraints that model the relations between the four input images, we first assume that the brightness of corresponding image points is the same in all frames. Using the

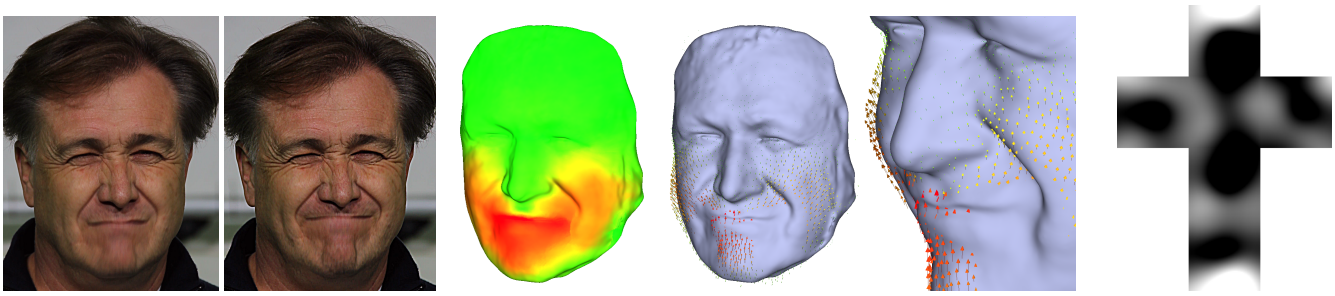

Figure 5: *Scene flow estimation and incident illumination estimation. First five images from left to right: two consecutive left frames f_0^t and f_0^{t+1}, the mesh M_c^t color coded by the scene flow magnitude (green to red for small to large motion), an overlay of the estimated scene flow vectors $\boldsymbol{W}^t$ on M_c^t, a detail of the same overlay. Last figure: an estimated lighting environment map for the same sequence.*

parameterization of [Valgaerts et al. 2010] with respect to the coordinates of the reference frame f_0^t, we obtain the four data terms

$$E_{\mathrm{D}}^1 = \Psi\left(|f_0^{t+1}(\boldsymbol{x}+\boldsymbol{w}_1) - f_0^t(\boldsymbol{x})|^2\right), \tag{3}$$

$$E_{\mathrm{D}}^2 = \Psi\left(|f_1^{t+1}(\boldsymbol{x}+\boldsymbol{w}_1+\boldsymbol{w}_2+\boldsymbol{w}_3) - f_1^t(\boldsymbol{x}+\boldsymbol{w}_2)|^2\right), \tag{4}$$

$$E_{\mathrm{D}}^3 = \Psi\left(|f_1^t(\boldsymbol{x}+\boldsymbol{w}_2) - f_0^t(\boldsymbol{x})|^2\right), \tag{5}$$

$$E_{\mathrm{D}}^4 = \Psi\left(|f_1^{t+1}(\boldsymbol{x}+\boldsymbol{w}_1+\boldsymbol{w}_2+\boldsymbol{w}_3) - f_0^{t+1}(\boldsymbol{x}+\boldsymbol{w}_1)|^2\right) \ . \tag{6}$$

While the first two terms result from motion constraints between two consecutive time instances, the last two terms arise from stereo constraints at the same time step. To handle outliers in all constraints independently, every data term is subject to a separate sub-quadratic penalization using the the regularized L_1 norm $\Psi(s^2) = \sqrt{s^2+\epsilon^2}$ as a cost function, with $\epsilon = 0.001$. To cope with varying illumination and to make use of color information, we additionally include the gradient constancy assumption in our final model and extend it to RGB color images [Valgaerts et al. 2010].

Geometry Terms The geometric relations between the left and the right image of the stereo pairs (f_0^t, f_1^t) and (f_0^{t+1}, f_1^{t+1}) are given by the associated *epipolar constraints*. These constraints relate corresponding points in a stereo pair via the fundamental matrix F – a projective entity that describes the geometry of the underlying stereo system [Hartley and Zisserman 2000]. The epipolar constraints between the two stereo pairs can be modeled as

$$E_{\mathrm{G}}^1 = \Psi\left(\left((\boldsymbol{x}+\boldsymbol{w}_2)_h^\top F\,(\boldsymbol{x})_h\right)^2\right), \tag{7}$$

$$E_{\mathrm{G}}^2 = \Psi\left(\left((\boldsymbol{x}+\boldsymbol{w}_1+\boldsymbol{w}_2+\boldsymbol{w}_3)_h^\top F\,(\boldsymbol{x}+\boldsymbol{w}_1)_h\right)^2\right) \ , \tag{8}$$

where the subscript h denotes the use of homogeneous coordinates, i.e. $(\boldsymbol{x})_h = (x, y, 1)^\top$. In contrast to [Valgaerts et al. 2010], we assume that our stereo system is calibrated with a known fundamental matrix F. Thus, in our case, only the flows $\boldsymbol{w}_i$ are unknown. Both terms E_{G}^1 and E_{G}^2 are soft constraints that penalize deviations of a point from its epipolar line. Together with a sub-quadratic penalizer function such as the regularized L_1 norm (see data terms), such soft constraints increase the robustness of the scene flow estimation with respect to small inaccuracies in the camera calibration.

Novel Structure-Aware Smoothness Terms Since the data terms and geometry terms alone may not guarantee a unique solution at every location, the problem needs to be regularized by imposing an additional *smoothness constraint*. In particular, this allows us to obtain dense scene structure and scene flow. In [Valgaerts et al. 2010], the isotropic total variation (TV, see L_1 norm above) regularizer is used. To recover typical facial features such as nose, eyebrows and laugh lines more realistically, however, we need a smoothness constraint that adapts better to the directional structure of the underlying reference image, while preserving sharp discontinuities in the reconstruction and the scene flow at the same time. To this end we make use of recent advances in the field of optical flow estimation [Sun et al. 2008; Zimmer et al. 2011] and propose the following anisotropic smoothness term

$$E_{\mathrm{S}}^i \ = \ \Psi_s\left(|\nabla \boldsymbol{w}_i^\top \boldsymbol{r}_1|^2\right) + \Psi_s\left(|\nabla \boldsymbol{w}_i^\top \boldsymbol{r}_2|^2\right) \ . \tag{9}$$

It splits the regularization locally into the direction *along* and *across* the image structures by projecting the Jacobian $\nabla \boldsymbol{w}_i$ onto $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$, respectively. Hereby, the directions $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$ are computed as eigenvectors of the structure tensor [Harris and Stephens 1988]

$$J_\rho = K_\rho * \nabla f_0^t \ \nabla f_0^{t\,\top} \ , \tag{10}$$

where $*$ denotes convolution with a Gaussian K_ρ. Since deviations from smoothness are penalized separately for each direction and typically a discontinuity-preserving cost function is used such as $\Psi_s(s^2) = 2\lambda_s^2\sqrt{1+s^2/\lambda_s^2}$, with $\lambda_s > 0$, discontinuities in the solution are preserved *independently* for both directions. This in turn allows us to handle structures of different intrinsic dimensionality such as corners, edges and homogeneous regions appropriately, thereby achieving the desired *structure-aware* anisotropic smoothing behavior. Extensions of these anisotropic smoothing ideas to color images are straightforward [Zimmer et al. 2011].

To introduce an additional feedback in the adaptivity of our smoothness term, we can apply the following modification: Instead of computing the directions $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$ only once – i.e. using the structure tensor from the reference image – we can recompute these directions during the scene flow estimation based on the structure of the evolving flows themselves. This leads to a highly adaptive *flow-aware* technique, where every flow is steered by a separate structure tensor computed by replacing f_0^t by u_i and v_i in Eq. (10).

In Fig. 6, we present a comparison of the TV regularizer used in [Valgaerts et al. 2010] (top row) with our structure-aware smoothness term (bottom row). It depicts two mesh geometries obtained by tracking the same coarse template over approximately 160 frames. From the zoom-in, it is clear that anisotropic regularized scene flow produces more realistic, tracked features, while TV regularized scene flow induces drift artifacts, such as double folding and degenerate triangles, in particular in the mouth and eyebrow region. The reason for the better tracking results is the selective smoothing of our structure-aware regularization, which allows an overall smoother scene structure and scene flow, while at the same time taking into account meaningful facial features. For a further visual comparison of scene flow results, see the supplementary material.

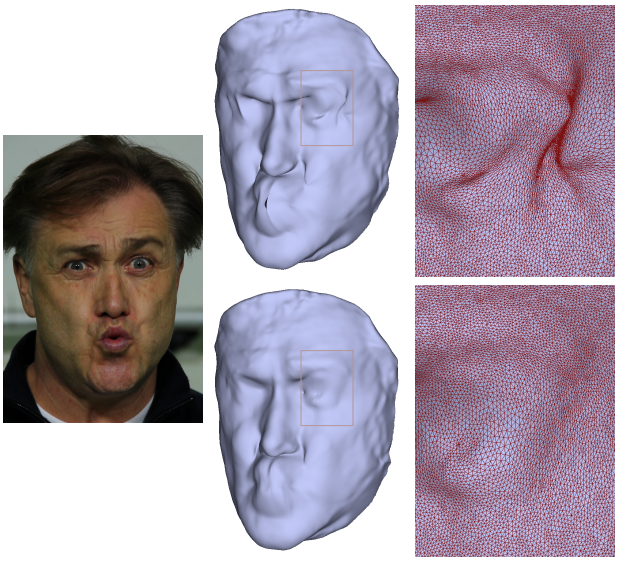

Figure 6: *Novel structure-aware smoothness terms. Top row: results obtained using [Valgaerts et al. 2010]. Bottom row: results obtained using our method. Left: left target frame. Middle column: tracked coarse mesh geometry. Right column: triangle-overlaid zoom-in into the highlighted region. Note the better tracking of expressive features such as mouth and eyebrows using our method.*

Minimization The final energy given in Eq. (2) has to be minimized with respect to the three unknown flows $\boldsymbol{w}_i$. To this end, we follow the the minimization scheme from [Valgaerts et al. 2010]: Large displacements are resolved by means of a coarse-to-fine multiresolution strategy, while the resulting nonlinear optimization problem at each resolution level is solved using a bidirectional multigrid method. In contrast to the original optimization scheme, we do not need to perform an alternating minimization between flows and fundamental matrix, since F is known from calibration. Due to its computational complexity, scene flow is computed for downsampled versions of the original images (half size in our experiments). As we will see in Sec. 5.2, the lower resolution of the scene flow will not compromise tracking accuracy, as motion vectors will only be assigned to a subset of the vertices.

All corresponding pixels can be triangulated to obtain a 3D reconstruction and a 3D displacement field. Note that we have only used 2D optical flow as an intermediate representation during the scene flow estimation. Our tracking algorithm effectively uses a 3D motion field, where each scene flow vector is characterized by a 3D starting position $\boldsymbol{S}^t$ and a 3D vector value $\boldsymbol{W}^t$. An example of the estimated scene flow field is depicted in Fig. 5. Note that we are able to cope with large motion and even noticeable motion blur.

5.2 Mesh Tracking

Once the scene flow $\boldsymbol{W}^t$ has been estimated for time instance t, it can be used to propagate the vertices of the current coarse mesh M_c^t to their new positions on $t+1$. However, moving each vertex by its corresponding scene flow vector is likely to induce local drift, which would quickly destroy the integrity of the template mesh. The reason for this is that the computed scene structure and motion contain errors, e.g. due to noise, which cannot be completely removed by our scene flow regularization. In addition, our scene flow lacks temporal coherence because it is estimated independently for all time instances. To ensure that the tracked geometry remains smooth over time, we have to regularize the moving geometry.

Positional Constraints To preserve the smoothness of the tracked mesh, we only assign a scene flow vector to a subset C^t of vertices. These will be denoted as *constrained vertices*, because their locations form the positional constraints in the regularization of the mesh geometry. We select the constrained vertices uniformly on the mesh (each tenth vertex in our experiments) to ensures a sufficient distribution of positional constraints. Additionally, we assure for each time instance t that all vertices in C^t are visible in both the left and the right image. This avoids erroneous tracking in regions that become occluded by head movement or expressive facial motion. For some outdoor sequences we experienced adverse interference of the estimated background motion at the side of the face. To avoid drift in more slanted regions for such cases, we restrict C^t to vertices for which the angle between the surface normal and the optical ray lies below a certain threshold ($\leq 70^\circ$).

Positional Update A first choice for updating the position of a constrained vertex $\boldsymbol{X}_i^t$, $i \in C^t$, is to move the vertex to the end point of the closest scene flow vector, thus updating its position as $\boldsymbol{X}_i^t \rightarrow \boldsymbol{S}_i^t + \boldsymbol{W}_i^t$. This strategy only produces good results if both the 3D reconstruction $\boldsymbol{S}_i^t$ and the 3D displacement field $\boldsymbol{W}_i^t$ are estimated with equally high accuracy. However, in practice, the structure $\boldsymbol{S}_i^t$ is more noisy than the motion $\boldsymbol{W}_i^t$ because the change in view point between the cameras induces a larger optical flow than the motion of the face. Especially for outdoor sequences captured with low quality cameras, this strategy lead to bumpy overfitting artifacts that could not be removed by Laplacian regularization.

For a smooth tracking result, instead, we determine the new position of a constrained vertex $\boldsymbol{X}_i^t$ by simply adding the closest scene flow vector $\boldsymbol{W}_i^t$ to the current vertex position. The updated constrained vertex position is then calculated as $\boldsymbol{X}_i^t \rightarrow \boldsymbol{X}_i^t + \boldsymbol{W}_i^t$. Possible small errors, introduced by adding the closest scene flow vector rather than moving to the scene flow end point, can be compensated by the optional motion refinement step of Sec. 5.3.

In our supplementary video we compare a noisy sequence of per-time-step 3D reconstructions with our final smooth tracking result.

Laplacian Regularization For a natural, shape preserving deformation of the face, we regularize the geometry of the target mesh M_c^{t+1} using the differential coordinates of the template mesh $M_c^{t_0}$ (similar in spirit to [Bradley et al. 2010]). The differential coordinates of $M_c^{t_0}$ encode the shape characteristics of the template surface and encapsulate information about the specific face that we are tracking. If we would use the differential coordinates of the current mesh M_c^t, the original shape of the face would not be preserved and the template structure would eventually be "forgotten". Using $M_c^{t_0}$ as a shape prior instead will avoid drift, while still allowing the capture of the low frequency component of strong facial deformations.

To deform M_c^t to M_c^{t+1} under the influence of the constrained vertices $\boldsymbol{X}_i^t$, $i \in C^t$, we minimize the energy

$$E = \left\| LX^{t+1} - LX^{t_0} \right\|^2 + \mu^2 \sum_{i \in C^t} \left\| \boldsymbol{X}_i^{t+1} - (\boldsymbol{X}_i^t + \boldsymbol{W}_i^t) \right\|^2, \quad (11)$$

where L is the Laplacian matrix with cotangent weights of $M_c^{t_0}$ [Sorkine 2005]. Further, X^{t+1} and X^{t_0} contain the vertex positions of the meshes M_c^{t+1} and $M_c^{t_0}$, and μ is a weighting factor.

5.3 Motion Refinement

Two possible sources of error remain in our tracking pipeline: First of all, the Laplacian regularization maintains mesh integrity but may prevent the vertices from moving to their true target positions. Secondly, we can expect a gradual accumulation of motion errors

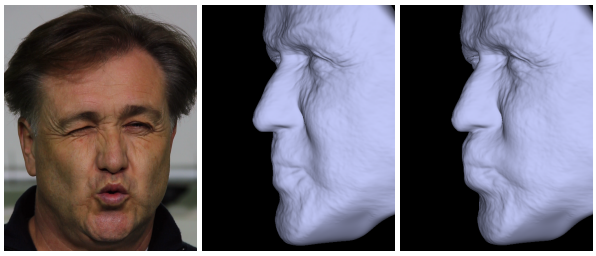

Figure 7: *Motion refinement. From left to right: input image, corresponding mesh obtained without motion refinement, corresponding mesh obtained with motion refinement. With motion refinement, the lips protrude as in the input image.*

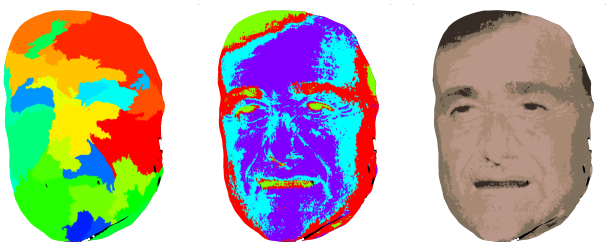

Figure 8: *Albedo clustering. Left to right: Original spatially coherent clustering from [Wu et al. 2011a], our new clustering result that corresponds better to typical facial feature distributions (e.g. around eyebrows, eyes etc.), average per material albedo coloring.*

over multiple frames. To compensate for such errors, we introduce a motion refinement step. The idea is to generate a synthetic image pair (f_0^r, f_1^r) by reprojecting the tracked mesh M_c^{t+1} onto the left and right image and to correct its position by minimizing the deviation between (f_0^r, f_1^r) and the ground truth (f_0^{t+1}, f_1^{t+1}). This effectively minimizes the reprojection error. We do this by computing the scene flow between (f_0^r, f_1^r) and (f_0^{t+1}, f_1^{t+1}) and by updating the position of M_c^{t+1} as explained in Sec. 5.2.

In Fig. 7, we illustrate the effect of motion refinement on a sequence of 30 frames. We see that the geometry obtained with motion refinement (right) is closer to what one would expect from the corresponding input image (left). This visual impression is confirmed quantitatively by a higher normalized cross correlation (NCC) between the reprojected image and the input image. For a quantitative confirmation in the form of a graph, see the supplementary material. For less expressive motion, such as speech, motion refinement can be considered optional as we found no large improvements in the estimated geometry. Motion refinement assumes that the texture of the face mesh does not change much over time and is thus less effective in case of changing illumination and cast shadows.

6 Shape Refinement

In this part, we explain how we employ shading cues to infer the high-frequency geometric detail and add it to the coarse tracked template. Our shading-based refinement algorithm consists of two steps: First, the lighting and albedo for each frame are estimated, after which both are utilized to optimize the geometry based on the shading information in the images. Our approach is inspired by the method of [Wu et al. 2011a], but it uses an albedo clustering that is better adapted to human faces as well as an improved refinement step which yields better results and faster convergence.

6.1 Albedo Clustering

In [Wu et al. 2011a], the surface albedo is assumed to be piece-wise uniform with larger coherent regions of similar reflectance. This could be efficiently segmented using a graph-based segmentation method [Felzenszwalb and Huttenlocher 2004]. However, when recording human faces from nearby camera positions, this assumption is less appropriate. While it is still fair to assume that there are a few albedo groups of vertices, their locations may not be spatially coherent, e.g. due to skin pigmentation, beard, shadows, etc. In contrast to [Wu et al. 2011a], we thus use a K-means clustering method to obtain k albedo groups, where vertices of the same group share the same albedo value. Particularly, given a set of initial per-vertex color albedo values $(a_1, a_2, \cdots, a_n)$, we aim at partitioning the n vertices of the mesh into k groups $S = \{S_1, S_2, \cdots, S_k\}$ to minimize the within-cluster sum of squares:

$$\arg\min_{S} \sum_{1}^{k} \sum_{a_j \in S_i} \|a_j - \mu_i\|^2 \quad , \tag{12}$$

where μ_i is the mean of the initial albedo of the vertices belonging to group i. The initial albedo value a_i is calculated from the shading equation with the geometry and lighting provided by the previous time frame. Once the albedo clusters are obtained, we utilize the same strategy as [Wu et al. 2011a] to estimate the incident illumination and the albedo value for each cluster.

An example of our improved albedo clustering strategy is shown in Fig. 8, where the different clusters are color coded. Fig. 5 shows an example of an estimated lighting environment map.

6.2 Surface Refinement

With the estimated illumination and albedos fixed, the coarse geometry of each frame is refined based on the shading cues in the images. The refined geometry is represented as the displacement of each vertex along its normal direction and is estimated by solving a spatio-temporal MAP inference problem.

A Novel Shading Energy Wu et al. [2011a] minimize a cost function that consists of a shading error term (data term) and a prior term (similarity term). Considering the fact that the reflectance of the face will not be purely Lambertian, such refinement will lead to noisy shape details when there are highlights on the skin. To account for this, we add to the energy a second prior term which requires the shape of the face to be spatially smooth (smoothness term). The cost function that we minimize then takes on the form:

$$E = \underbrace{E_{\mathrm{D}}}_{\text{data}} + \underbrace{\lambda_{\mathrm{M}} E_{\mathrm{M}}}_{\text{similarity}} + \underbrace{\lambda_{\mathrm{S}} E_{\mathrm{S}}}_{\text{smoothness}} \quad , \tag{13}$$

where λ_{M} and λ_{S} are weighting factors. The data term E_{D} is the shading error that measures the similarity of the shading gradients in the input images f_0^t and f_1^t to the predicted shading gradients:

$$E_{\mathrm{D}} = \sum_{i} \sum_{j \in N(i)} \sum_{c \in Q(i,j)} \left(r_c(i,j) - s(i,j)\right)^2 \quad , \tag{14}$$

where i and j are triangle indices, $N(i)$ is the set of neighboring triangles of triangle i, c is the camera index, $Q(i,j)$ is the set of cameras which see triangles i and j, and $r(i,j)$ and $s(i,j)$ are the measured image gradient and predicted shading gradient. The similarity term E_{M} is a prior term based on the previous frame geometry, that requires the current refined geometry M_r^t to be similar to the

refined geometry of the previous time step M_r^{t-1}, transplanted on the coarse mesh M_c^t. It constrains the reconstructed high-frequency shape detail in the face, such as fine folds and laugh lines, to change in a spatio-temporally coherent way. It takes on the form:

$$E_{\mathrm{M}} = \sum_i^n \sum_{u,v} \left(\hat{\boldsymbol{n}}_i^t \cdot (\boldsymbol{X}_u^t - \boldsymbol{X}_v^t)\right)^2 \quad , \tag{15}$$

where vertices $\boldsymbol{X}_u^t$, $\boldsymbol{X}_v^t$ and $\boldsymbol{X}_i^t$ belong to the same mesh triangle and $\hat{\boldsymbol{n}}_i^t$ is the propagated surface normal based on the already reconstructed high-frequency normal field of the previous time $t-1$. For more details on the propagation of the surface normals, we refer to the base-line method of [Wu et al. 2011a]. However, in contrast to the base-line method, we define the surface normals in the data and similarity term on triangles, since this allows a better approximation and easier calculation than for normals defined on a per-vertex basis. The third, newly-added term in our energy (13) is the smoothness term E_{S}, which has the following form:

$$E_{\mathrm{S}} = \sum_i^n \left\| \sum_{j \in N(i)} w_{ij} \left(\boldsymbol{X}_i^t - \boldsymbol{X}_j^t\right) \right\|_2^2 \quad . \tag{16}$$

Here $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ are the positions of the vertices i and j in the mesh M_r^t, $N(i)$ is the 1-ring neighborhood of vertex i, and w_{ij} are the common cotangent weights [Sorkine 2005].

Novel Fast Iterative Minimization The shading energy (13) is usually non-linear and not trivial to minimize. Wu et al. [2011a] employ a patch-based non-linear optimization strategy to refine the geometry within separate vertex patches. However, this strategy poses a trade-off between run time and refinement quality: While a small patch size may not constrain the neighboring vertices enough to achieve high quality large-displacement shape refinement, a large patch size will take much longer to compute.

We reduce this trade-off by replacing the non-linear optimization of the energy by an iterative linear one. To this end, we replace the only non-linear part in the energy – the argument of the shading error term (14) – with its first-order Taylor approximation. This way, all terms in the energy become squared linear with respect to the unknown vertex displacement. Since each vertex in the resulting energy is only coupled to its direct neighbors, the displacements can be easily found by solving a sparse linear system. A first-order Taylor approximation is only valid for small displacements, so in practice we update the vertex positions using the obtained solution scaled by an adjustable step size. We repeat this procedure, such that the sequence of newly-defined energies approximates the original one better. In our experiments, we use a step size of 0.7 and iterate 4 times to obtain the final refinement.

Fig. 9 shows that our novel shading energy and iterative minimization strategy lead to superior results compared to [Wu et al. 2011a]: Our estimated face surface suffers less from noisy artifacts, while exhibiting a higher level of fine-scale detail. The supplementary material shows with a graph that the proposed optimization strategy, which solves for all vertices simultaneously, converges to a lower energy, and thus a better optimal shape. It also provides evidence for the computational speed-up over [Wu et al. 2011a].

6.3 Temporal Postprocessing

After the final shape refinement, there might remain a slight temporal flicker in the visualization of the results due to small differences in the direction of the surface normals. To reduce this effect, we update the normals in the whole sequence by averaging them over a temporal window of size 5 and then adapting the geometry to the updated normals using the method of [Nehab et al. 2005].

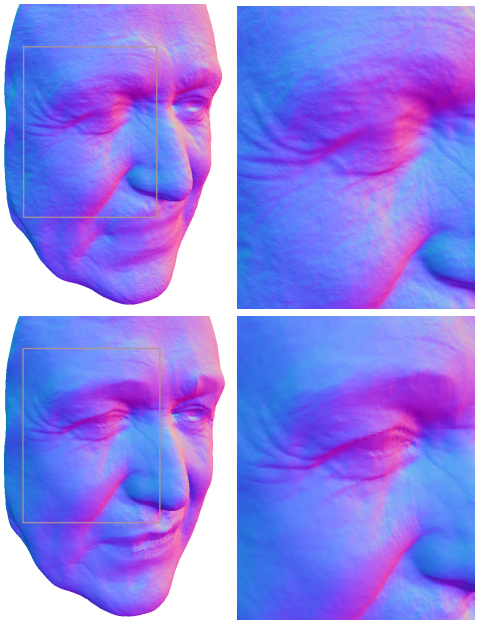

Figure 9: *Novel Shape Refinement. Top row: results obtained using [Wu et al. 2011a]. Bottom row: results obtained using our method. Both meshes are colored by normal orientation. The zoom-in shows that we obtain a smoother result with an even higher level of detail.*

Figure 10: *The set-ups used in our experiments. From left to right: A pair of Canon EOS 550D cameras, the GoPro 3D Hero system.*

7 Results

We evaluate the performance of our approach on real world data of three different test subjects captured with two different set-ups: 1) a pair of Canon SLR cameras in an indoor environment and 2) a pair of GoPro helmet cameras, used indoor and outdoor. Five sequences with a length of 300 to 560 frames (12s to 22s) will be presented.

Canon Set-up Our first set-up consists of two Canon EOS 550D cameras in an indoor environment (Fig. 10). These cameras record HD video with a resolution of 1920×1088 at 25 fps. They are not hardware synchronized, and synchronization is just verified by event-based temporal alignment. The green screen in the figure is not required and is just a standard feature of the room we used.

Fig. 11 shows the results for a subject captured with this set-up. All meshes consist of the same set of vertices and are produced by tracking a single template throughout a sequence of around 300 frames. The number of vertices is 100000. These results illustrate that we are able to capture very expressive facial motion at a level of detail that rivals more complex methods using more cameras and controlled lighting. Reconstructions are space-time coherent with no perceivable drift, as illustrated by the checkerboard result. With such high-quality reconstructions, realistic looking textured faces can be created via projective texturing with no perceivable ghosting.

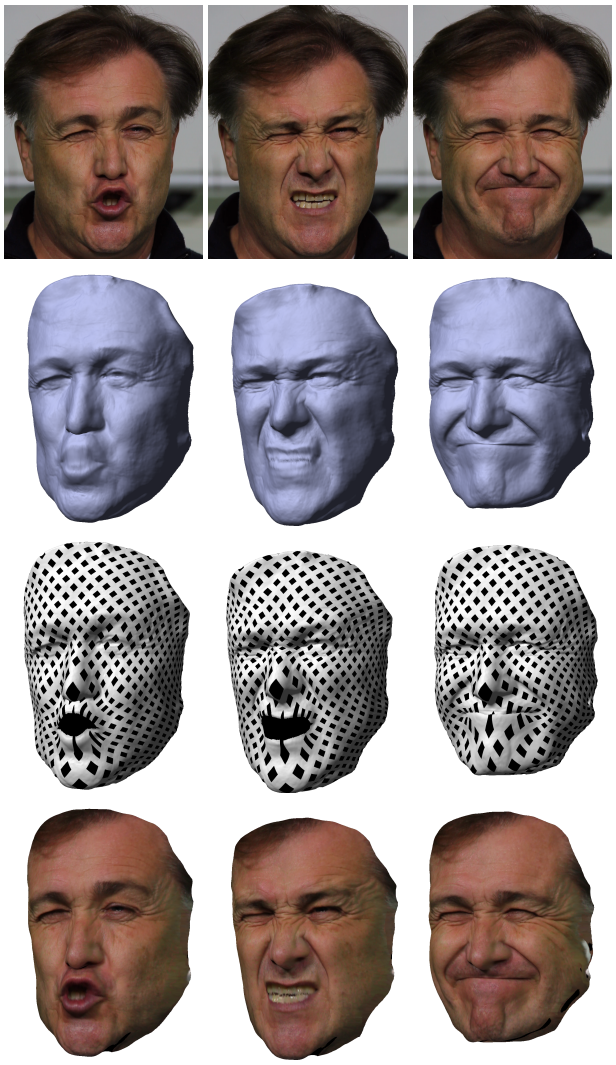

Figure 11: *Results for a pair of Canon cameras. From top to bottom: the left input image, the corresponding reconstructed mesh, the mesh overlaid with a checkerboard pattern to demonstrate geometric coherence, the mesh colored using projective texturing.*

In Fig. 12 we provide a long captured sequence for a different actor performing both extreme facial gestures and normal conversation. Both types of motion are captured by our method with high quality. Motion blur, due to the fast movement, makes the sequences especially challenging. However, as shown in Fig. 5, our approach is robust to this and captures fast motion reliably. Even after 560 frames, our method has hardly introduced any temporal drift.

The parameters used for both experiments are: $\alpha_1 = 5$, $\alpha_2 = 5$, $\beta_1 = 200$, $\beta_2 = 150$, $\beta_3 = 200$, $\rho = 3$, $\lambda_s = 0.1$, $k = 4$, $\lambda_{\mathrm{M}} = 2500$ and $\lambda_{\mathrm{S}} = 10000$. The weight μ was chosen as 1 and 0.9 respectively. The run time for our sequential, non-optimised code is around 9m per frame on an Intel Xeon@3.1GHz.

GoPro Set-up Our second setup uses a pair of GoPro HD Hero cameras that are hardware synchronized and combined in a single housing (Fig. 10). The pair records at 1920×1080 and 30 fps each. The camera is designed to be used outdoors on bike helmets and is at best comparable with an upscale web cam. Data are challenging due to the smaller baseline, cheap plastic wide angle lenses, the generally higher noise level, the rolling shutter, and potential automatic white balancing which can not be controlled.

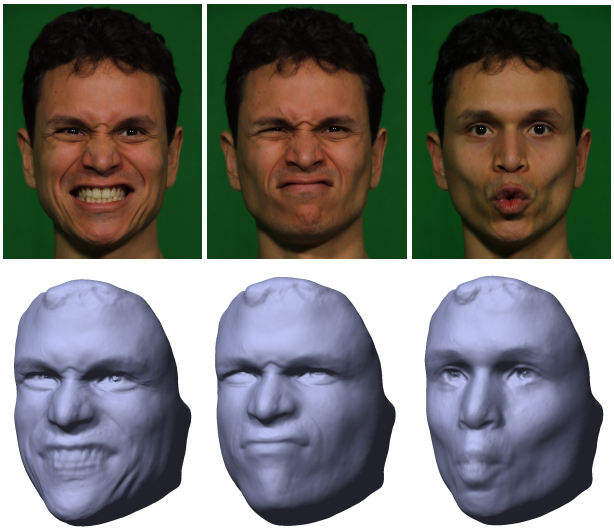

Figure 12: *Results for a pair of Canon cameras. From top to bottom: the left input image, the corresponding reconstructed mesh.*

Recordings are done with the hand-held GoPro HD Hero stereo system in both indoor and outdoor environments. One such set-up is depicted in the first two images in Fig. 13, where a speaking actor is recorded indoors. The general uncontrolled lighting make this scenario extremely difficult for any facial motion capture algorithm. Moreover, compared to the Canon set-up, the face of the actor only makes up a small portion of the HD images. Despite these challenges, we obtain reconstructions which exhibit a fair amount of detail, Fig. 13. We are able to capture the face, including the head motion, over extended periods of time with only little drift.

A second scenario is shown in the middle of Fig. 13, where an actor records himself outdoors in bright direct sunlight. For this sequence, the face was captured over 400 frames, and although the quality is not as high as for the indoor recordings, we are able to recover a large amount of detail. A strong shadow from the nose floats freely over the mouth region and motion refinement in the tracking step treats this incorrectly as physical motion. It was disabled for this sequence, but we are still able to achieve very expressive facial motion. The same strong shadow also leads to artifacts at the boundary caused by the shading-based refinement step. These high-frequency effects can be partly alleviated by the use of higher order spherical harmonics (see [Wu et al. 2011b]) to better approximate the visibility and shadow boundaries, but this would increase the run time substantially. Another option to handle this case, shown in the supplementary video, is the explicit detection of strong shadows to disable shape refinement there. We detect shadows by high shading errors and estimate them iteratively with the lighting. The results that we obtain do not have shadow boundary artifacts, but do exhibit less detail. This trade-off between fine detail and shadow artifacts should be chosen with respect to the application in mind.

The right two images of Fig. 13 show a third scenario of an actor recording himself walking outdoors under trees. This is a very challenging set-up, not in the least due to additional background motion and changing illumination on the face. Nevertheless, we are able to capture highly detailed and realistic facial motion, which shows that both our tracking and shape refinement pipeline are robust with respect to the aforementioned difficulties. These and additional outdoor self-capture results are included in the supplementary video.

The parameters are the same as for the Canon sequences, except $\beta_1 = 300$, $\beta_2 = 200$, $\beta_3 = 300$ and $\lambda_S = 40000$. The weight μ is 0.9 for the indoor sequence and 0.4 for the outdoor sequences.

Validation As shown in the video, all steps in our pipeline are essential to obtain detailed results on our challenging input. Scene flow estimation and geometry regularization need to be combined robustly, otherwise drift will deteriorate the tracked coarse mesh geometry quickly. Motion refinement helps to improve reconstruction quality, as shown visually in Fig. 7 and quantitatively in the supplementary material. Based on this coarse geometry, lighting and albedo can be estimated and shading-based refinement applied, as shown in the paper and the video. Overall, our shape refinement pipeline achieves a speedup over and provides better results than the baseline method of [Wu et al. 2011a], as shown in Fig. 9 and the graphs in the supplementary material. In the video, we also show that our results are superior to a state-of-the-art binocular stereo method [Valgaerts et al. 2011]. The supplementary material further compares our reconstruction of a static face against a laser scan of a similar pose (no hole-filling). While this does not provide a quantitative evaluation, it still shows that our reconstruction quality comes close to that of state-of-the-art static scanning technologies.

Recall that we are solving a different problem than high-quality multi-camera approaches that are designed for perfect studio conditions. Our data sets are stereo data with a small baseline, while those from a state-of-the-art multi-camera approach such as [Beeler et al. 2011] are multi-view data captured under studio lighting with a camera positioning naturally tailored to their approach and a comparably large baseline. Hence, applying our method to a subset of their cameras is difficult, as the positioning of these cameras suits the multi-view case but not our two-view case (for similar reasons, two distinct sets of benchmark data, one for multiple views[3] and one for two views[4], are available). Given the algorithm specific camera arrangements, we believe that direct comparison to a multi-view method would therefore not be meaningful. A reasonable comparison would be to record the same face under two parallel synchronized camera set-ups, one multi-view set-up and a separate two-view set-up, such that both categories of methods can be compared on the same scene. We will record such a data set in the future and make it available together with our other data sets.

Discussion We believe that our approach is the first to capture highly detailed facial performances from a single stereo rig under uncontrolled illumination and that it shows that on-set capture with consumer grade hardware is feasible. Nevertheless, our approach is subject to limitations. One of these are strong shadows. In our experiments, we demonstrated that we can reduce artifacts at shadow boundaries in the shape refinement step at the expense of less detail. Strong moving shadows are also a challenge for the tracking step, which interprets them as surface motion (see the supplementary material for an example). To counter these effects, we plan to investigate better ways of detecting shadows and the use of photometric invariants for scene flow [Zimmer et al. 2011]. Another limiting scenario for our method is low light conditions (e.g. dim rooms or overcast weather). Besides this, temporal drift is reduced by our method, but can not be completely prevented over extended periods of time. In this context, combining our template tracking approach with a key-frame-based regularization [Beeler et al. 2011] may be promising. Also, scene flow estimation could be further improved by using the refined geometry model as an explicit regularizer rather than relying merely on image constraints. We also plan to extract more advanced reflectance and lighting models from the data and investigate if these can further improve the results.

[3] vision.middlebury.edu/mview/

[4] vision.middlebury.edu/stereo/

8 Conclusion

We have presented an algorithm for capturing high-quality facial performances from a single stereo pair of video streams that were captured under uncontrolled illumination, even outdoors. This becomes possible through the use of a robust binocular variational scene flow method that was adapted to the face capture scenario, as well as through the combination of a mesh tracking and a shading-based refinement approach that captures space-time coherent and highly detailed geometry. With our approach we are able to produce results of a high quality that could not be achieved before using just a single stereo rig. We believe that our method can make hand-held facial performance capture feasible for everyone. It also opens the door for new applications in on-set performance capture, movie postprocessing, social media and teleconferencing.

Acknowledgements

We gratefully acknowledge our actors and thank Nils Hasler and Carsten Stoll for helping with the code.

References

ALEXANDER, O., ROGERS, M., LAMBETH, W., CHIANG, M., AND DEBEVEC, P. 2009. The digital emily project: photoreal facial modeling and animation. In *ACM SIGGRAPH 2009 Courses*, ACM, 12:1–12:15.

ANUAR, N., AND GUSKOV, I. 2004. Extracting animated meshes with adaptive motion estimation. In *Proc. VMV*, 63–71.

BASHA, T., MOSES, Y., AND KIRYATI, N. 2010. Multi-view scene flow estimation: a view centered variational approach. In *Proc. CVPR*, 1506–1513.

BASRI, R., JACOBS, D., AND KEMELMACHER, I. 2007. Photometric stereo with general, unknown lighting. *IJCV 72*, 3, 239–257.

BEELER, T., BICKEL, B., BEARDSLEY, P., SUMNER, B., AND GROSS, M. 2010. High-quality single-shot capture of facial geometry. *ACM TOG 29*, 40:1–40:9.

BEELER, T., HAHN, F., BRADLEY, D., BICKEL, B., BEARDSLEY, P., GOTSMAN, C., SUMNER, R. W., AND GROSS, M. 2011. High-quality passive facial performance capture using anchor frames. *ACM TOG 30*, 75:1–75:10.

BICKEL, B., BOTSCH, M., ANGST, R., MATUSIK, W., OTADUY, M., PFISTER, H., AND GROSS, M. 2007. Multi-scale capture of facial geometry and motion. *ACM TOG 26*, 33:1–33:10.

BIRKBECK, N., COBZAS, D., AND JÄGERSAND, M. 2011. Basis constrained 3D scene flow on a dynamic proxy. In *Proc. ICCV*.

BLANZ, V., BASSO, C., VETTER, T., AND POGGIO, T. 2003. Reanimating faces in images and video. *CGF (Proc. EUROGRAPHICS) 22*, 641–650.

BORSHUKOV, G., PIPONI, D., LARSEN, O., LEWIS, J. P., AND TEMPELAAR-LIETZ, C. 2003. Universal capture: image-based facial animation for "the matrix reloaded". In *ACM SIGGRAPH 2003 Sketches*, ACM, 16:1–16:1.

BRADLEY, D., HEIDRICH, W., POPA, T., AND SHEFFER, A. 2010. High resolution passive facial performance capture. *ACM TOG (Proc. SIGGRAPH) 29*, 41:1–41:10.

Figure 13: *Results for a pair of GoPro HD helmet cameras for a person being recorded in an uncontrolled indoor environment (left two images), a person recording himself outdoors in bright sunlight (middle two images) and under changing illumination (right two images).*

CARCERONI, R. L., AND KUTULAKOS, K. N. 2002. Multi-view scene capture by surfel sampling: from video streams to non-rigid 3D notion, shape and reflectance. *IJCV 49*, 2-3, 175–214.

DECARLO, D., AND METAXAS, D. 1996. The integration of optical flow and deformable models with applications to human face ahape and motion estimation. In *Proc. CVPR*, 231–238.

FELZENSZWALB, P. F., AND HUTTENLOCHER, D. P. 2004. Efficient graph-based image segmentation. *IJCV 59*, 2, 167–181.

FURUKAWA, Y., AND PONCE, J. 2009. Dense 3D motion capture for human faces. In *Proc. CVPR*, 1674–1681.

GEORGHIADES, A. S. 2003. Recovering 3-D shape and reflectance from a small number of photographs. In *Proc. EGSR*, 230–240.

GUENTER, B., GRIMM, C., WOOD, D., MALVAR, H., AND PIGHIN, F. 1998. Making faces. In *Proc. SIGGRAPH*, ACM, 55–66.

HARRIS, C. G., AND STEPHENS, M. 1988. A combined corner and edge detector. In *Proc. Alvey Vision Conf.*, 147–152.

HARTLEY, R., AND ZISSERMAN, A. 2000. *Multiple View Geometry in Computer Vision*. Cambridge University Press.

HUANG, H., CHAI, J., TONG, X., AND WU, H.-T. 2011. Leveraging motion capture and 3D scanning for high-fidelity facial performance acquisition. *ACM TOG 30*, 74:1–74:10.

JIN, H., CREMERS, D., WANG, D., PRADOS, E., YEZZI, A., AND SOATTO, S. 2008. 3-D reconstruction of shaded objects from multiple images under unknown illumination. *IJCV 76*, 3, 245–256.

KAZHDAN, M., BOLITHO, M., AND HOPPE, H. 2006. Poisson surface reconstruction. In *Proc. SGP*, 61–70.

NEHAB, D., RUSINKIEWICZ, S., DAVIS, J., AND RAMAMOORTHI, R. 2005. Efficiently combining positions and normals for precise 3D geometry. *ACM TOG 24*, 3, 536–543.

PIGHIN, F., SZELISKI, R., AND SALESIN, D. 1999. Resynthesizing facial animation through 3d model-based tracking. In *Proc. CVPR*, vol. 1, 143–150.

POPA, T., SOUTH-DICKINSON, I., BRADLEY, D., SHEFFER, A., AND HEIDRICH, W. 2010. Globally consistent space-time reconstruction. *CGF (Proc. SGP) 29*, 1633–1642.

SORKINE, O. 2005. Laplacian mesh processing. In *STAR Proceedings of Eurographics 2005*, Eurographics Association, 53–70.

SUN, D., ROTH, S., LEWIS, J. P., AND BLACK, M. J. 2008. Learning optical flow. In *Proc. ECCV*, vol. 5304, 83–97.

VALGAERTS, L., BRUHN, A., ZIMMER, H., WEICKERT, J., STOLL, C., AND THEOBALT, C. 2010. Joint estimation of motion, structure and geometry from stereo sequences. In *Proc. ECCV*, Springer LNCS, vol. 6314, 568–581.

VALGAERTS, L., BRUHN, A., MAINBERGER, M., AND WEICKERT, J. 2011. Dense versus sparse approaches for estimating the fundamental matrix. *IJCV*. Springer Online First.

VOGIATZIS, G., AND HERNÁNDEZ, C. 2011. Self-calibrated, multi-spectral photometric stereo for 3D face capture. *IJCV*.

WAND, M., ADAMS, B., OVSJANIKOV, M., BERNER, A., BOKELOH, M., JENKE, P., GUIBAS, L., SEIDEL, H.-P., AND SCHILLING, A. 2009. Efficient reconstruction of nonrigid shape and motion from real-time 3D scanner data. *ACM TOG 28*, 15:1–15:15.

WANG, Y., HUANG, X., SU LEE, C., 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. *CGF 23*, 677–686.

WEDEL, A., RABE, C., VAUDREY, T., BROX, T., FRANKE, U., AND CREMERS, D. 2008. Efficient dense scene flow from sparse or dense stereo data. In *Proc. ECCV*, vol. 5302, 739–751.

WEISE, T., LEIBE, B., AND GOOL, L. J. V. 2007. Fast 3D scanning with automatic motion compensation. In *Proc. CVPR*.

WEISE, T., BOUAZIZ, S., LI, H., AND PAULY, M. 2011. Realtime performance-based facial animation. *ACM TOG 30*, 77:1–77:10.

WILLIAMS, L. 1990. Performance-driven facial animation. In *Proc. SIGGRAPH*, ACM, 235–242.

WILSON, C. A., GHOSH, A., PEERS, P., CHIANG, J.-Y., BUSCH, J., AND DEBEVEC, P. 2010. Temporal upsampling of performance geometry using photometric alignment. *ACM TOG 29*, 17:1–17:11.

WU, C., VARANASI, K., LIU, Y., SEIDEL, H.-P., AND THEOBALT, C. 2011. Shading-based dynamic shape refinement from multi-view video under general illumination. In *Proc. ICCV*.

WU, C., WILBURN, B., MATSUSHITA, Y., AND THEOBALT, C. 2011. High-quality shape from multi-view stereo and shading under general illumination. In *Proc. IEEE CVPR*, 969–976.

ZHANG, L., NOAH, CURLESS, B., AND SEITZ, S. M. 2004. Spacetime faces: high resolution capture for modeling and animation. *ACM TOG 23*, 548–558.

ZIMMER, H., BRUHN, A., AND WEICKERT, J. 2011. Optic flow in harmony. *IJCV 93*, 3, 368–388.