# Method for Calculating View-Invariant 3D Optical Strain

Matthew Shreve, Sergiy Fefilatyev, Nestor Bonilla, Gerardo Hernandez,
Dmitry Goldgof, Sudeep Sarkar

Department of Computer Science and Engineering
mshreve@mail.usf.edu
University of South Florida
Tampa, Florida

**Abstract.** Optical strain maps calculated during a facial expression describe a bio-mechanical property of facial skin tissue, and are derived from the non-rigid motion incurred during facial expressions. In this paper, we propose a method for accurately estimating and modeling the three-dimensional strain impacted onto the face using standard optical flow techniques and a dynamic depth map. We demonstrate the robustness of this approach at different depth resolutions and views. Experimental results are given for a publically available dataset that contains high depth resolutions of facial expressions, as well as a new dataset collected using the Microsoft Kinect synchronized with two HD webcams.

**Keywords:** 3D, Optical Strain, Optical Flow

## 1 Introduction

Two-dimensional optical strain has been shown to be an effective feature for several applications such as expression spotting [1], biometrics [2], as well as medical analysis [3]. Due to the recent releases of several feasible solutions for real-time three-dimensional imaging, successfully recovering the three-dimensional elastic properties of the face is now practical and could potentially lead to improvements in each of these areas. In this work, we demonstrate a method for calculating the 3-D strain incurred on a subject's face based on the non-rigid facial motion observed during a facial expression.

Some key advantages of this method over traditional two-dimensional strain methods are the following: (i) horizontal motions that occur along the sides of the face are often projected as smaller displacements due to parallax projection. Our method of calculating three-dimensional strain has the advantage of reconstructing these vectors in order to represent a more accurate displacement; (ii) motion perpendicular to the camera axis is not factored in to two-dimensional strain maps, however this is captured with three-dimensional strain as an additional normal strain component.

The method is based on the observation that the captured depth images of the surface of the face exhibit 2-manifold qualities, i.e., local regions of the surface can be accurately estimated using two dimensional planar equations. Hence, we take ad-

Workshop on Depth Image Analysis, 2012.

vantage of this by estimating the three-dimensional correspondences using established two-dimensional optical flow methods.

Methods for calculating non-rigid three-dimensional disparity has found extensive application in the entertainment industry, where they are used to animate faces of humanoid avatars in movies and games. Most current systems require special makeup or markers [4-6] due to low texture variability of the skin. Such markers, however, provide only a limited number of anchor points and are not sufficient to capture fine expression details at all points on the face. Other approaches [7] use direct intensity of the skin to track the displacements between the frames by imposing constraints on non-rigid motion. Approaches for 3D reconstruction of the scene itself can be broadly categorized into three groups. The first group, *motion stereo*, requires a multi-view camera setup of a single scene, where 3D information is obtained through triangulation of 3D points from multiple views [7-10]. Most of the prior work for estimation of non-rigid 3-D disparity falls into this category. The second category, called *monocular sequence,* uses a single-camera setups, but exploits *a priori* knowledge about the reconstructed scene by developing a model that constrains and develops the observed 2D motion [11]. The last category, referred to as a *dynamic depth map* approach [12,13], assumes a monocular setup for both depth and color information using inexpensive sensors such as Microsoft Kinect. Our approach for 3D optical strain estimation is based on the monocular setup and direct intensity skin tracking with a dynamic depth map.

## 2 3-D Optical Strain

### 2.1 Optical Flow

Optical Flow is a well-known motion estimation technique that has two constraints: (i) the smoothness constraint, i.e., points within a small region move at some level of uniformity, and (ii) the brightness constraint, i.e. the intensity of a point in the image does not change over time. Optical flow is typically represented by the following equation.

$$(\nabla \mathrm{I})^T \mathbf{p} + I_{\mathrm{t}} = 0, \tag{1}$$

where $I(x,y,t)$ is the image intensity as a spatial and temporal function, $x$ and $y$ are the image coordinates and $t$ is time. $\nabla I$ and $I_{\mathrm{t}}$ are the spatial and temporal gradients of the intensity function. $\mathbf{p} = \left[p = \frac{dx}{dt}, q = \frac{dy}{dt}\right]^{\boldsymbol{T}}$ denotes horizontal and vertical motion.

After experimenting with several versions of optical flow, we decided to use an implementation of the Horn-Schunck [14] method found in the computer vision toolbox for Matlab 2012. The Horn-Schunk method consists of re-writing equation (1) as a global energy function that is constrained by a smoothness parameter $\alpha \in (0,1]$, and is optimized over k iterations. In general a lower alpha allows for less smooth flow fields (good for small motion), while a larger alpha restricts neighboring motions to be more uniform (good for large motions). For our experiments we used a low value for the smoothness constraint ($\alpha = .05$) to allow for the small, non-rigid motion inherent with facial expressions, and chose an iteration count of k=200.

## 2.2 Optical Strain

Considering a three dimensional surface of a deformable object, its motion can be described by a three-dimensional displacement vector $\boldsymbol{u} = [u, v, w]^T$. Next, if we assume both a small region and small motion for a point P, we can define the strain tensor:

$$\varepsilon = \frac{1}{2}[\nabla\boldsymbol{u} + (\nabla\boldsymbol{u})^{\boldsymbol{T}}]\,, \tag{2}$$

or in an expanded form:

$$\varepsilon = \begin{bmatrix} \varepsilon_{xx} = \frac{\partial u}{\partial x} & \varepsilon_{yx} = \frac{1}{2}\left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\right) & \varepsilon_{zx} = \frac{1}{2}\left(\frac{\partial u}{\partial z} + \frac{\partial w}{\partial x}\right) \\ \varepsilon_{xy} = \frac{1}{2}\left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\right) & \varepsilon_{yy} = \frac{\partial v}{\partial y} & \varepsilon_{zy} = \frac{1}{2}\left(\frac{\partial w}{\partial y} + \frac{\partial v}{\partial z}\right) \\ \varepsilon_{xz} = \frac{1}{2}\left(\frac{\partial w}{\partial x} + \frac{\partial u}{\partial z}\right) & \varepsilon_{yz} = \frac{1}{2}\left(\frac{\partial w}{\partial y} + \frac{\partial v}{\partial z}\right) & \varepsilon_{zz} = \frac{\partial w}{\partial z} \end{bmatrix} \tag{3}$$

Where$(\varepsilon_{xx}, \varepsilon_{yy,}, \varepsilon_{zz})$ are normal strain components, $(\varepsilon_{xy}, \varepsilon_{zy}, \varepsilon_{zx})$ are shear strain components, and *u*,*v*,*w* are the displacements in the *x*,*y*,*z* directions.

Since strain is defined with respect to the displacement vector (*u*,*v*,*w*) in continuous space, we make the following 2-D approximation from the optical flow data (*p*,*q*):

$$p = \frac{dx}{dt} \doteq \frac{\Delta x}{\Delta t} = \frac{u}{\Delta t}, u = p\Delta t\,, \tag{4}$$

$$q = \frac{dy}{dt} \doteq \frac{\Delta y}{\Delta t} = \frac{v}{\Delta t}, v = q\Delta t\,, \tag{5}$$

$$r = \frac{dz}{dt} \doteq \frac{\Delta z}{\Delta t} = \frac{w}{\Delta t}, w = r\Delta t. \tag{6}$$

where $\Delta t$ is the elapsed time between two image frames.

If we compute the optical flow and strain using a fixed frame interval throughout a particular video sequence, we can treat $\Delta t$ as a constant and estimate the partial derivatives as follows:

$$\frac{\partial u}{\partial x} = \frac{\partial p}{\partial x}\Delta t, \quad \frac{\partial u}{\partial y} = \frac{\partial p}{\partial y}\Delta t, \quad \frac{\partial u}{\partial z} = \frac{\partial p}{\partial z}\Delta t, \tag{7}$$

$$\frac{\partial v}{\partial x} = \frac{\partial q}{\partial x}\Delta t, \quad \frac{\partial v}{\partial y} = \frac{\partial q}{\partial y}\Delta t, \quad \frac{\partial v}{\partial z} = \frac{\partial q}{\partial z}\Delta t, \tag{8}$$

$$\frac{\partial w}{\partial x} = \frac{\partial r}{\partial x}\Delta t, \quad \frac{\partial w}{\partial y} = \frac{\partial r}{\partial y}\Delta t, \quad \frac{\partial w}{\partial z} = \frac{\partial r}{\partial z}\Delta t, \tag{9}$$

After the initial flow estimates are updated, we then take the spatial derivative in each direction. We chose the central difference method due to its accuracy and efficiency [2]. Hence,

$$\frac{\partial u}{\partial x} = \frac{u(x+\Delta x)-u(x-\Delta x)}{2\Delta x} \doteq \frac{p(x+\Delta x)-p(x-\Delta x)}{2\Delta x} \tag{10}$$

$$\frac{\partial v}{\partial y} = \frac{v(y+\Delta y)-v(y-\Delta y)}{2\Delta y} \doteq \frac{q(y+\Delta y)-q(y-\Delta y)}{2\Delta y} \tag{11}$$

$$\frac{\partial w}{\partial z} = \frac{w(z+\Delta z)-w(z-\Delta z)}{2\Delta z} \doteq \frac{r(z+\Delta z)-r(z-\Delta z)}{2\Delta z} \tag{12}$$

where $(\Delta x, \Delta y, \Delta y)$ are preset distances of 2-3 pixels.

Assuming a uniform stress across the face, elastorgrams based on the absolute strain value or relative strain ratio can be used to reveal underlying elastic property changes and the strain magnitude can be calculated as follows:

$$\varepsilon_m = \sqrt{\varepsilon_{xx}^2 + \ \varepsilon_{yy}^2 + \varepsilon_{zz} + \varepsilon_{xy}^2 + \varepsilon_{yx}^2 + \varepsilon_{zx}^2 + \varepsilon_{yz}^2}\ . \tag{13}$$

Lastly, for visualization, we normalize the strain magnitudes to [0,1] and illustrate the strain map with a color bar (see Fig. 1).

# 3 Results

## 3.1 Feasibility at Multiple depth resolutions

In order to test the feasibility of using multiple depth resolutions, we developed an experiment that subsamples high resolution depth data at different rates. This was done on a publically available 3D dataset released from Binghamton University [15]. Fig. 1 contains an example subject from this dataset.

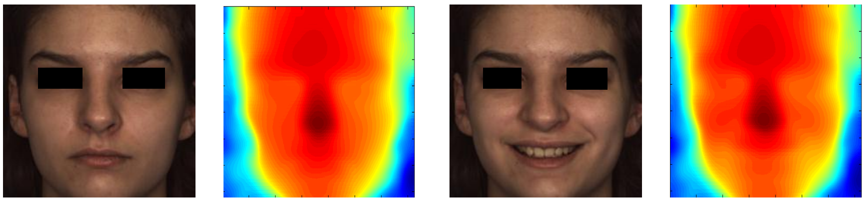

**Fig. 1.** Example data from BU dataset showing face image and corresponding depth map (red=closest, blue = farthest). Eyes masked for privacy concerns.

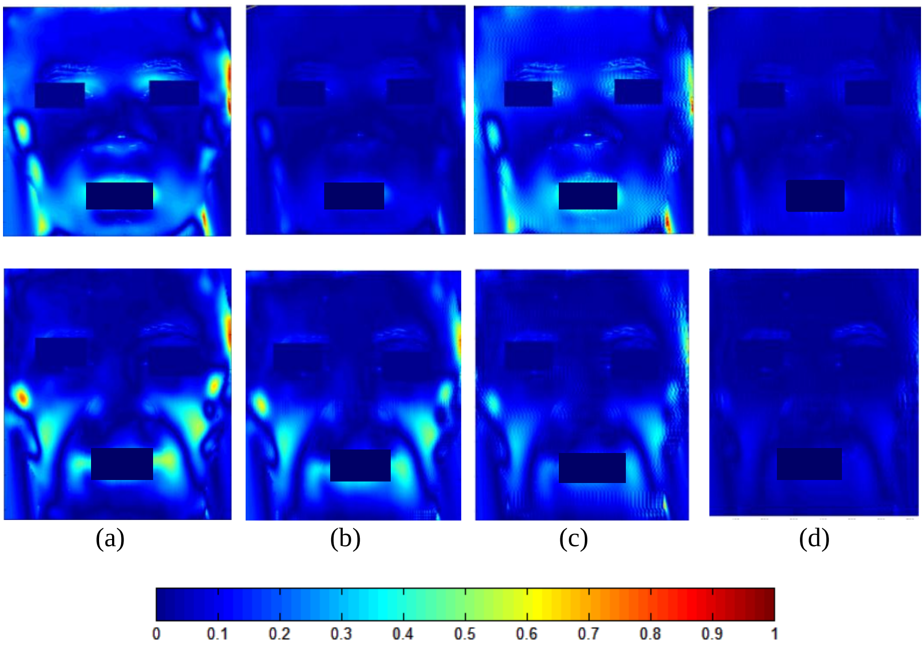

**Fig. 2.** Example normalized 3-D strain maps calculated for two subjects corresponding to the surprise and smile expressions (each row). Depths were sub-sampled at ratios of (a) 1:1, (b) 1:2, (c) 1:3, and (d) 1:4 (each row) resulting in depth resolutions of approximately 200x200, 100x100, 66x66 and 50x50. Regions around eyes / mouth are masked due to noise, as in [2].

We selected 20 subjects performing two expressions (smile, surprise) for a total of 40 sequences. The cropped face resolutions are approximately 700 x 700 pixels in dimension, with approximately every 3x3 pixel window containing a single depth value. We sampled the depth values at a 1:1, 1:2, 1:3, and 1:4 ratio and then used bilinear interpolation to scale the values back up to 700x700. Fig. 2 Contains some example strain maps calculated at each scale.

In order to measure the similarity between strain maps calculated at several different depth resolutions, the correlation coefficient was used (Table 1). Several observations can be made. First, at least an 80% similarity is maintained for both expressions even when using approximately 70x70 of 200x200 (a third) of available depth points.

**Table 1.** Correlation coefficients for 40 expression (20 smile, 20 surprise) after subsampling at the given ratios and compared with a 1:1 sampling ratio.

| **Exp. / Ratio** | **1:2** | **1:3** | **1:4** |
|---|---|---|---|
| **Smile** | .90±.11 | .80±.12 | .69±.18 |
| **Surprise** | .95±.02 | .89±.05 | .79±.15 |
| **Both** | .93±.08 | .85±.10 | .74±.17 |

### 3.2 View Invariance

To demonstrate the view invariance of the method and give further evidence of the methods stability at low resolutions, we collected several subjects using the Microsoft Kinect sensor that was synchronized with an additional two additional HD webcams at approximately 30 degree angles to the face. We then registered each webcam to the automatically calibrated image provided by the Kinect using manually labeled eye coordinates. The Kinect sensor provides depth imaging at an image resolution of 640x480 and a depth resolution of 320x280. However, due to mechanical restrictions on the minimum distance allowed to the sensor (roughly 3 feet), the face image is typically 175x175 in image resolution with a depth resolution of roughly 90 x 90. Some examples images and depth maps for this dataset can be found in Figure 3.

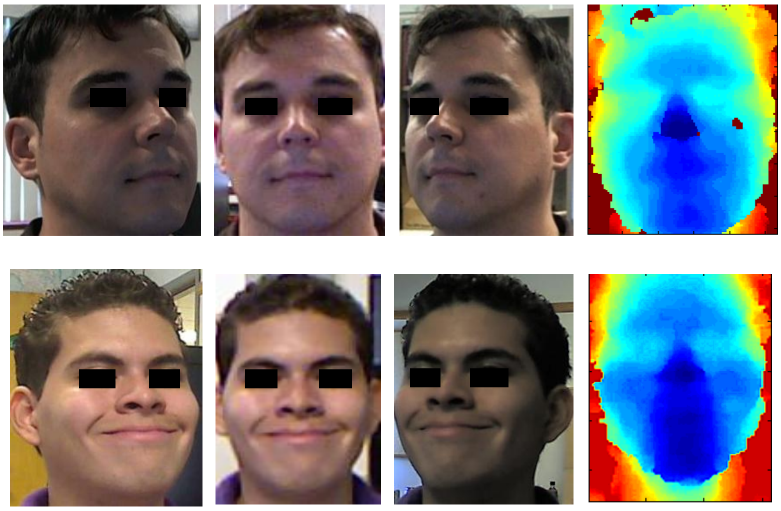

**Fig. 3.** Example images captured from approx. 30 degree angles with depth map (blue=closest, red = farthest) captured from Kinect synchronized with two webcams. Eyes masked for privacy concerns.

Due to the amount of noise in the depth values provided by the Kinect, smoothing was required as pre-processing step. It is worth noting that we tested several kernel sizes and standard deviations, and a 3x3 Gaussian kernel with a standard deviation of .1 led to the best trade-off between keeping the three-dimensional structure of the face intact and minimizing erroneous noisy depth gradients. Some example 3-D strain maps can be found in Figure 4 for two different views that demonstrate the view-invariance of the method.

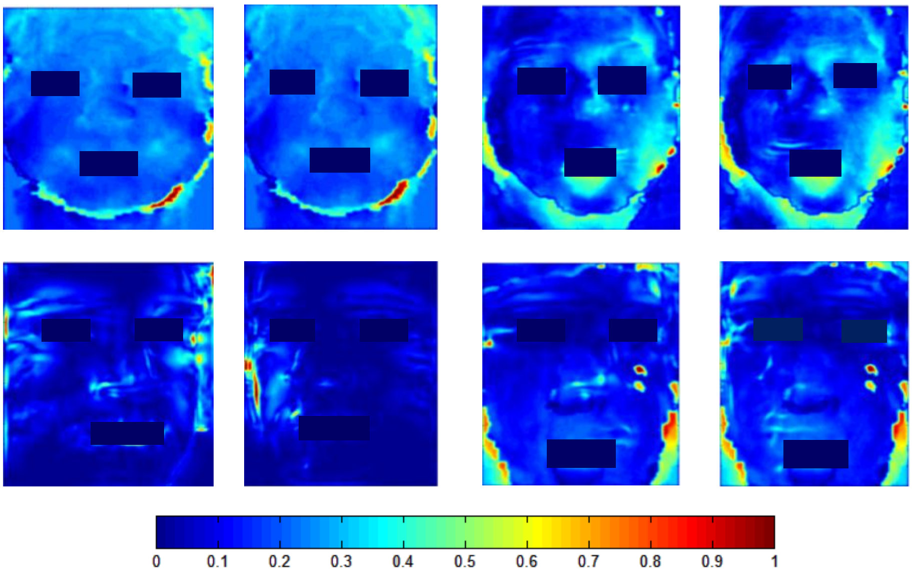

**Fig. 4.** Example strain maps calculated at two views roughly 45 degrees apart, for two subjects (each row). The first two pairs of columns are for the smile expression, the second pair of columns are for the surprise expression. Regions around eyes / mouth are masked due to noise, as in [2].

## 4 Conclusions

Optical strain maps calculated during facial expression are a bio-mechanical feature that has been shown to have broad significance in facial motion analysis. In this paper, we propose a method for calculating view-invariant three-dimensional strain maps based on 2-D optical flow correspondences over a dynamic depth map. We have demonstrated that the method is robust at several different depth resolutions and views by first giving results on the high resolution BU dataset sampled at different resolutions, as well as a low-resolution Kinect that was synchronized with two HD webcams at different angles.

# 5 References

1. M. Shreve, S. Godavarthy, D. Goldgof, “Macro- and micro-expression spotting in long videos using spatio-temporal strain”, Proceedings of Int. Conference on Automatic Face and Gesture Recognition, pp. 51-56, 2011.
2. M. Shreve, V. Manohar, D. Goldgof, S. Sarkar, “Face recognition under camouflage and adverse illumination”, Proceedings of International Conference on Biometrics: Theory Applications and Systems, pp. 1-6, 2010
3. M. Shreve, N. Jain, D. Goldgof, S. Sarkar, W. Kropatsch, C. Tzou, M. Frey, “Evaluation of facial reconstructive surgery on patients with facial palsy using optical strain”, Computer Analysis of Images and Patterns, pp. 512-519, 2011.
4. B. Bickel, M. Botsch, R. Angst, W. Matusik, M. Otaduy, H. Pfister, M. Gross, “Multi-scale capture of facial geometry and motion”, ACM Transactions on Graphics, v. 29, nu. 3, pp. 33, 2007.
5. V. Blanz, C. Basso, T. Poggio, T. Vetter, “Reanimating faces in images and video”, Computer Graphics Forum, 22(3):641-650, 2003.
6. I. Lin, M. Ouhyoung, “Mirror MoCap: Automatic and efficient capture of dense 3D facial motion parameters”, Visual Computer, 21(6):355-372.
7. D. Bradley, W. Heidrich, T. Popa, A. Sheffer, “High resolution passive facial performance capture”, ACM Transactions on Graphics, 29(4):41, 2010.
8. Y. Furukawa, J. Ponce, “Dense 3D motion capture from synchronized video streams”, Image and Geometry Processing for 3-D Cinematography, pp. 193-211, 2010.
9. Y. Furukawa, J. Ponce, “Dense 3D motion capture for human faces”, Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 1674-1681, 2009.
10. J. Pons, R. Keriven, O. Faugeras. Multi-view stereo reconstruction and scene flow estimation with a global image-based matching score”, International Journal of Computer Vision, 72(2):179–193, 2007.
11. M. Penna, “The Incremental Approximation of Nonrigid Motion”, Computer Vision, Graphics, and Image Processing, 60(2):141-156, 1994.
12. S. Hadfield, R. Bowden, “Kinecting the dots: particle based scene flow from depth sensors”, Proceedings of International Conference on Computer Vision, pp. 2290-2295, 2011.
13. T. Weise, S. Bouaziz, H. Li, M. Pauly, “Realtime Performance-Based Facial Animation”, ACM Transactions on Graphics, 30(4),pp. 77, 2011.
14. B. Horn, B. Schunck, “Determining optical flow”, Artificial Intelligence, 17:185-203, 1981.
15. J. Neumann , Y. Aloimonos, “Spatio-temporal stereo using multi-resolution subdivision surfaces”, International Journal of Computer Vision, 47(1-3):181–193, 2002.
16. L. Yin, X.Chen, Y. Sun, T. Worm, M. Reale, “A High-Resolution 3D Dynamic Facial Expression Database” International Conference on Automatic Face and Gesture Recognition,2008