Non-Rigid Registration of 3D Facial Surfaces with Robust Outlier Detection

Moritz Kaiser, Andre Störmer, Dejan Arsić and Gerhard Rigoll
Institute for Human-Machine Communication
Technische Univerität München
Arcisstr. 3, 80333 München, Germany
{moritz.kaiser,andre.stoermer,dejan.arsic,rigoll}@tum.de

Abstract

Non-rigid registration of 3D facial surfaces is a crucial step in a variety of applications. Outliers, i.e.*, features in a facial surface that are not present in the reference face, often perturb the registration process. In this paper, we present a novel method which registers facial surfaces reliably also in the presence of huge outlier regions. A cost function incorporating several channels (red, green, blue,* etc.*) is proposed. The weight of each point of the facial surface in the cost function is controlled by a weight map, which is learned iteratively. Ideally, outliers will get a zero weight so that their disturbing effect is decreased. Results show that with an intelligent initialization the weight map improves the registration results considerably.*

1. Introduction

There are many applications in computer vision and computer graphics that stand or fall on the exact non-rigid registration of 3D facial surfaces. Examples include transfer of texture between faces [18], transfer of expressions [11, 16], facial animation [19], or texture mapping [20]. Another popular application is building up a 3D face model [3].

Non-rigid registration of facial surfaces is not a trivial problem. In some approaches the non-rigid registration of facial surfaces is based on a sparse set of feature points. In [12], the authors suggest to select those points manually. Wang *et al*. [18] map the 3D facial surface into the 2D plane using harmonic mapping and feature points are tracked at high video rate. In other works dense registration methods are proposed. Blanz and Vetter [3] suggested an automatic registration process similar to the Kanade-Lucas-Tomasi (KLT) algorithm [17]. Savran and Sankur [15] proposed a method where the 3D surface is mapped to the 2D plane with least squares conformal mapping. As smoothness constraint they employ a Green-Lagrange strain tensor. However, the authors mention that the algorithm is not designed to deal with outliers such as an open mouth. A similar approach was presented by Litke *et al*. in [10]. The authors propose a mapping function to map the 3D facial surface into the 2D plane which minimizes length and area distortions. A sophisticated cost function depending on feature demarcations, curvature and texture is minimized to match the two 2D images. Also in this approach outliers were not considered. Bronstein *et al*. [4] proposed generalized multidimensional scaling that allows to embed a facial surface directly into the reference facial surface without any 2D registration process in between. The approach works without texture information. Outliers, such as an open mouth, have to be treated manually.

Relatively little has been done on the treatment of outliers, *i.e.*, features that are not existent in the reference face, such as facial hair, open mouth, glasses, borders, *etc*. However, it is most likely that outliers exist for faces of two different individuals. If outliers are present they can seriously corrupt the registration. Hellier *et al*. [7] presented a study on dense deformation fields for the registration of brain magnetic resonance images (MRI) in the presence of outliers, where they apply robust error functions.

In this paper we particularly investigate registration of 3D facial surfaces with outliers (*i.e.* open mouth, glasses, beard, *etc.*) that corrupt the registration of the surface, if not treated specifically. The 3D facial surface is mapped into the 2D plane with least squares conformal mapping. Subsequently, a cost function that accounts for texture/depth constancy and spatial coherence is minimized. It is shown that if the disturbing effect of outliers is not decreased the registration fails. The main constribution of this work is the weight map that determines the significance of every pixel in the registration process. Outliers should have a small weight or ideally a zero weight so that their disturbing effect is suppressed. We present an algorithm to learn the weight map iteratively. With an intelligent initialization the method registers the face and detects outliers reliably.

The paper is organized as follows. In Sec. 2 it is explained how the 3D facial surfaces are mapped into the 2D

978-1-4244-5496-9/09/$25.00 ©2009 IEEE

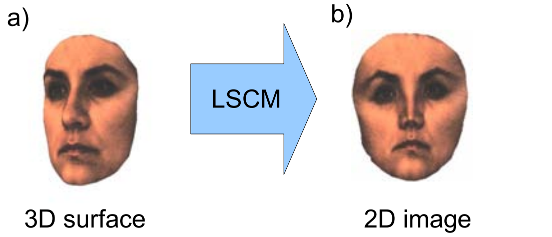

Figure 1. Mapping of a 3D facial surface (a) into the 2D plane (b) via least squares conformal mapping.

plane. A cost function to solve the 2D registration problem is described in Sec. 3. In Sec. 4 the weight map is introduced. The results of the non-rigid registration method are presented in Sec. 5. Section 6 gives a conclusion and outlines future work.

2. Mapping of 3D Surfaces into the 2D Plane

The 3D facial surface is mapped into the 2D plane in order to convert the registration of a 3D surface into a 2D registration problem. The simplest approach would be considering the depth component of the 3D points of the facial surface as fourth channel of a 2D RGB image. Blanz and Vetter [3] suggested to apply cylindrical projection. However, such projections may cause large distortions in regions where the angle between the surface normal and the projection direction is large. More sophisticated methods to map 3D surfaces into the 2D plane have been used in [4, 10, 15, 18]. Many of these methods were developed in the context of texture mapping. An overview of surface parameterization methods is given in [6].

We decided to apply LSCM presented by Lévy *et al.* in [9]. For our purpose it is important that the distortions caused by the mapping remain small. LSCM minimizes angle deformations and non-uniform scalings. Furthermore, by using conjugate gradient methods or quasi-Newton methods for minimization the mapping can be computed very fast. Consequently, 2D images with four channels (red, green, blue, and depth) are obtained. The 2D image of the reference facial surface is denoted by $\boldsymbol{I}_{\mathrm{ref}}$ and the image of the face that we want to register with respect to the reference facial surface is denoted by $\boldsymbol{I}_{\mathrm{reg}}$. Figure 1 illustrates a 3D facial surface (a) from the Bosphorus database [14] mapped into the 2D plane (b) with LSCM.

3. 2D Registration Method

A non-rigid 2D registration problem can be solved with an optical flow estimation. A cost function that has to be minimized can be formulated locally, as in [3], or globally, as in [10] or [15]. We decided to apply a global cost function, because thereby it is easier to integrate a robust error norm (introduced in Sec. 3.1) and a weight map (introduced in Sec. 4.1).

3.1. Cost Function

Following the ISO typesetting standards, matrices are denoted by capital bold letters ($\boldsymbol{I}$) and scalars/elements of a matrix by normal letters (I, u). The cost function is based on two assumptions:

Texture/Depth Constancy. The shift between a pixel in $\boldsymbol{I}_{\mathrm{ref}}$ and a corresponding pixel in $\boldsymbol{I}_{\mathrm{reg}}$ is the unknown that we want to compute. The shift of a pixel (x, y) in x-direction and y-direction is denoted by $u(x, y)$ and $v(x, y)$, respectively. For better readability, we will write u instead of $u(x, y)$, I instead of $I(x, y)$, *etc.*

Since the beginning of optical flow estimation, it has been assumed that the gray value of a pixel is not changed by displacement. This assumption leads to the famous optical flow constraint [8] for each pixel: $I_x \cdot u + I_y \cdot v + I_t = 0$. The partial derivatives of the gray value image in x- and y-direction are denoted by I_x, I_y, the temporal derivative by I_t.

We modify this assumption for our purpose. Instead of only one gray value channel, six channels were employed. Color information is available so the first three channels are the red, green, and blue channel. The fourth channel is a gray value derivative channel, which further improves the illumination independency, as shown in [5]. The fifth and the sixth channel are the partial derivatives of the depth in x- and y-direction, respectively. Thereby, the susceptibility to variable skull shapes of different individuals can be reduced compared to just taking the depth channel.

The optical flow constraint for all six channels is combined in the following cost function:

$$E_{\mathrm{t/d}}(u, v) = \Psi\Big(\sum_{c=1}^{C} w_c \cdot (I_{x,c} \cdot u + I_{y,c} \cdot v + \Delta I_c)^2\Big), \quad (1)$$

where the index c stands for the channels (r, g, b, *etc.*) and C is the total number of channels. ΔI_c is the subtraction of $I_{\mathrm{ref},c}$ from $I_{\mathrm{reg},c}$. Each channel is weighted with a weight w_c. Note that, instead of a quadratic error norm, we use $\Psi(r^2) = \sqrt{r^2 + \epsilon^2}$, where ϵ is set to a small fixed value (here $\epsilon = 0.001$), as suggested in [5]. This provides robustness against small portions of outliers. Other robust error norms have been suggested and intensively investigated in [1] and [2].

Spatial Coherence. The spatial coherence assumption postulates that the flow field of neighboring points should not differ too much. Several cost functions to integrate this assumption have been proposed (see [15, 8, 2]). Facial skin has properties similar to a rubber membrane. Thus, the membrane model [2] is suitable to model spatial coherence in a face:

$$E_{\mathrm{s}}(u,v) = \sum_{(u_n,v_n)\in\mathcal{N}} \Psi\Big((u-u_n)^2 + (v-v_n)^2\Big), \quad (2)$$

where $(u_n, v_n) \in \mathcal{N}$ are pixels in the 3×3 or 5×5 neighborhood of (u, v).

The total cost function is a weighted sum of the texture/depth constancy term and the spatial coherence term summed over all pixels in Ω_{ref}:

$$E(\boldsymbol{u},\boldsymbol{v}) = \sum_{\Omega_{\mathrm{ref}}} \lambda \cdot E_{\mathrm{t/d}}(u,v) + E_{\mathrm{s}}(u,v). \quad (3)$$

Ω_{ref} is the set of all pixels that display facial texture in the reference image. Background, *i.e.*, parts of the image that do not display the face, should not be considered.

In previous works (*e.g.* [2, 7, 5, 15]), λ is a fixed constant, that has the same value for all pixels. In Sec. 4.1, we propose to employ a $\lambda(x, y)$ which is different for each pixel. $\lambda(x, y)$ should be low for potential outliers. Hence, disturbing outliers can be suppressed.

3.2. Multiscale Approach

The total cost function is minimized via conjugate gradient minimization [13]. In order to cope with larger motions the well-known coarse-to-fine strategy is employed [17]. The optical flow field is first estimated at the coarsest level of a Laplace pyramid. The coarse-scale estimate is used to warp to the next pyramid level.

In Fig. 2, we illustrate some results of the registration process with the suggested cost function. The reference face is depicted in Fig. 2 (a). Several faces from the Bosphorus Database [14] are shown as original in the left column. The right column displays the texture of these faces mapped to the shape of the reference face. For the individual in Fig. 2 (b), which is quite similar to the reference face, the registration was successful. If faces contain features that are not existent in the reference face, as shown in Fig. 2 (c), (d), and (e), the registration is significantly corrupted. The glasses in Fig. 2 (c) are mapped to the eyes, because they have also dark texture. The texture of the hair in Fig. 2 (d) and the beard in Fig. 2 (e) is very different from the corresponding texture in the reference face, so those points are pushed outside the area that is considered by the cost function (Ω_{ref}). The labels *correct* and *incorrect* on the faces are based on the qualitative appearance of the warped faces. Quantitative results are given in Sec. 5.2 (Tab. 1). The evaluation method is explained in detail also in this section. As baseline system a registration incorporating the KLT algorithm [17] instead of the global cost function we suggest was used. The global cost function performs only slightly better than the baseline system.

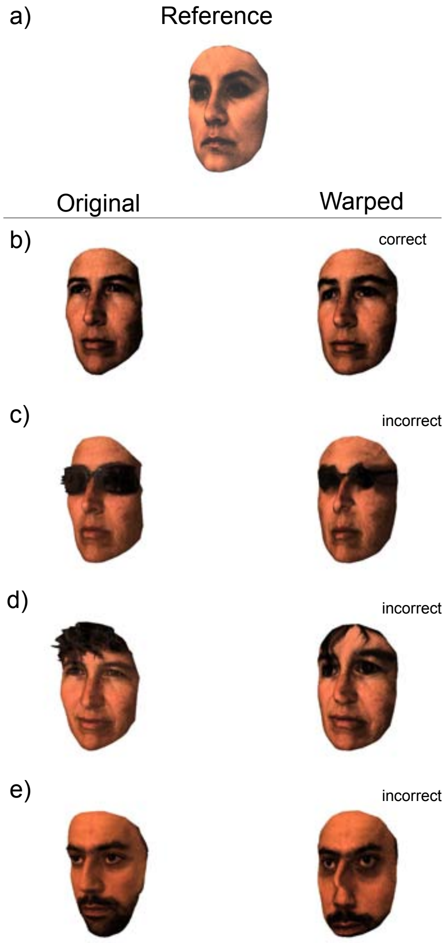

Figure 2. Left column: Several facial surfaces of the Bosphorus Database [14]. Right column: Texture of those facial surfaces mapped to the shape of the reference facial surface, that is shown in (a). Without special treatment of the outliers the registration works fine if the faces are similar (b). However, if there are features that are not existent in the reference face, such as glasses (c), hair (d), or beard (e), the registration is poor.

4. Outlier Detection

Outliers are points with depth or color values that are not existent or cannot clearly be assigned to a point in the reference face, such as facial hair, open mouth, glasses, border, *etc.* and thus can corrupt the registration process. Several approaches (*e.g.* [2, 7, 5]) propose to include a robust error norm that detects and reduces the influence of an outlier. However, for larger outlier regions the results with a robust error norm are unsatisfying. A robust error norm was included in Eq. 1 and 2 and Fig. 2 (c), (d), and (e) show that it fails. If there are regions with more outliers than non-outliers, outliers cannot be detected and, even worse, non-outliers are detected as outliers. The weight map that we will present in this work is able to handle these cases much better.

4.1. Weight Map

A novel approach to reduce the disturbing effect of outliers is proposed in this work. A different weight $\lambda(x, y)$ for each pixel in Eq. 3 is used instead of a constant λ for all pixels. We call the matrix that stores the weight for each pixel *weight map* $\boldsymbol{\lambda}$. Thus, Eq. 3 can be reformulated as

$$E(\boldsymbol{u}, \boldsymbol{v}) = \sum_{\Omega_{\mathrm{ref}}} \lambda(x, y) \cdot E_{\mathrm{t/d}}(u, v) + E_{\mathrm{s}}(u, v). \qquad (4)$$

Ideally, the weight of an outlier should be 0 and the weight of a non-outlier 1. Note that the weight for the texture/depth constancy term is controlled by the weight map $\boldsymbol{\lambda}$, while the factor for the spatial coherence term remains constant so that the flow field for outliers is guided by the nearest non-outliers. The weight map is learned iteratively. It is initialized with 1 for each pixel in Ω_{ref} and with 0 for pixels lying outside. After the registration process the weight for each pixel is updated:

$$\lambda_{i+1}(x, y) = \frac{1}{2}\Big(\frac{1}{\sum_{c=1}^{C} \Delta_c(x, y) + \epsilon} + \lambda_i(x, y)\Big), \qquad (5)$$

with $\epsilon = 0.001$ preventing a division by zero. The difference of the value of channel c (red, green, blue, *etc.*) of a pixel (x, y) in $\boldsymbol{I}_{\mathrm{ref}}$ and the corresponding pixel in $\boldsymbol{I}_{\mathrm{reg}}$ according to the current estimate of (u, v) is denoted by $\Delta_c(x, y)$. Subsequently, the weight map is normalized so that the sum of all weights remains constant. With the new weight map the registration process can be repeated and a finer registration is obtained since the disturbing effect of the outliers is reduced. The repetition of the registration process can be stopped after a fixed number of iterations or if the weight map is not changing considerably any more. In order to let the weight map converge correctly, a good initialization is needed. If initially too many non-outliers have a low weight, the algorithm converges to a local minimum which results in an improper registration.

Both quantitative results and a detailed explanation of the evaluation method are given in Sec. 5.2 (Tab. 1). Compared to the registration without a weight map $\boldsymbol{\lambda}$ there is some improvement, but it can be significantly enhanced with a proper initialization, which is described in the next section. The problem that still has to be solved is that for some individuals the iterative registration converges only to a local minimum.

4.2. Intelligent Initialization of the Weight Map

As could be seen in the last section, the registration process stops at a local minimum, if the initial estimate of the weight map is too poor. A registration only based on depth information, although not very exact, is always a good rough first estimate and it is not corrupted too much by outliers. Once the facial surface is roughly registered, a proper initial guess for the weight map is computed and the registration process can be repeated with all channels for finer registration. The algorithm can thus be divided into the following steps:

1. Weight for all pixels in Ω_{ref} is 1, for other pixels 0.
2. Repeat
   (a) Register facial surface.*
   (b) Compute weight map with the result.
   (c) Stop after fixed number of iterations or if weight map does not change considerably any more.

*For the first iteration only channels that depend on depth information are applied for a robust initialization. At later iterations all channels are considered for refinement.

Compared to the baseline system, the weight map with an intelligent initialization could decrease the average registration error by 16%. In Sec. 5 it is explained how the average registration error is computed. Quantitative results (Tab. 1) and qualitative results (Fig. 3) are commented in detail also in this section.

5. Results

5.1. Parameters

A 256 × 192 raster has been employed to map the 3D surface into the 2D plane. We used 5 resolution levels for the coarse-to-fine pyramid and 6 iterations for the registration process (one only with channels depending on the depth information and five depending on all channels, as proposed in Sec. 4.2). The weights of the channels in Eq. 1 were chosen so that the two channels depending on depth information together have the same weight as the four channels depending on color information together. The registration of the facial surfaces has been implemented in C++ and one iteration took approximately 20 seconds on a 3GHz Intel® Pentium® Duo-Core so 6 iterations took roughly 120 seconds.

Some registration results are depicted in Fig. 3. Again, the texture of the faces in the database is mapped to the shape of the reference face according to the correspondence, which has been computed. The reference face and the sample faces that have been used in the previous figures are shown in Fig. 3 (a) and (b), respectively. The outliers are illustrated in the right column. It can be seen that with a weight map $\boldsymbol{\lambda}$ and an intelligent initialization the faces could be registered and the outliers detected reliably. The glasses are not shrunk and facial hair is not pushed outside the reference face as before. Figure 3 (c) shows the results of the registration process for several other faces with outliers, such as beard, hair and open mouth. For all those faces, which partly deviate considerably from the reference face, the registration process was qualitatively successful.

5.2. Evaluation

The Bosphorus Database [14] was employed for the evaluation. The data is labeled with 22 landmarks per face. These landmarks were considered as ground truth. For those 22 sample points the accuracy of the registration process can be measured. The average distance $\bar{d} = \sum_i \sqrt{\Delta x_i^2 + \Delta y_i^2 + \Delta z_i^2}$ between the position estimated by our method and the true coordinates of this landmark was computed. The average distance was normalized to the height of the reference face. For example, a distance of 0.05 stands for 5% of the height of the reference face.

Table 1 shows the average distance for all 105 faces in the database with a neutral expression. For comparison the classical Kanade-Lucas-Tomasi (KLT) algorithm [17], as proposed in [3], was chosen as baseline system. It can be seen that the registration works to some degree with the KLT algorithm (average distance of 0.0537) or the global cost function that has been presented in Sec. 3 (average distance of 0.0521). With only a weight map (Sec. 4.1) an average distance of 0.0494 can be achieved. However, with both a weight map (Sec. 4.1) and an intelligent initialization (Sec. 4.2) an average distance of 0.0450 can be achieved. Compared to the baseline system with an average distance of 0.0537 our method could decrease the average distance by $(0.0537 - 0.0450)/0.0537 = 16\%$. It is observed that with the method that has been proposed the global minimum tends to be found reliably. The quantitative results coincide with the qualitative results. Visually it can be verified in Fig. 3 that the faces are registered accurately and also huge outlier regions could be detected reliably.

6. Conclusion and Future Work

In this paper, we propose a novel method to register 3D facial surfaces and to detect features that are not present in the reference face more robustly. A weight map, which is learned iteratively, controls the weight of each point of the facial surface in the registration cost function. The first estimate of the weight map is computed only with depth information, which is less exact but also less outlier prone. At later iterations also texture information is considered. Remarkable results could be achieved only after a few iterations even in the presence of large outlier regions, as shown in Fig. 3. Quantitative results are given using landmarks available with the database as ground truth. Our method could improve the registration accuracy by 16 % compared to the method [3] which was chosen as baseline system.

In our ongoing research, we will employ other strategies to map the 3D facial surface into the 2D plane and investi-

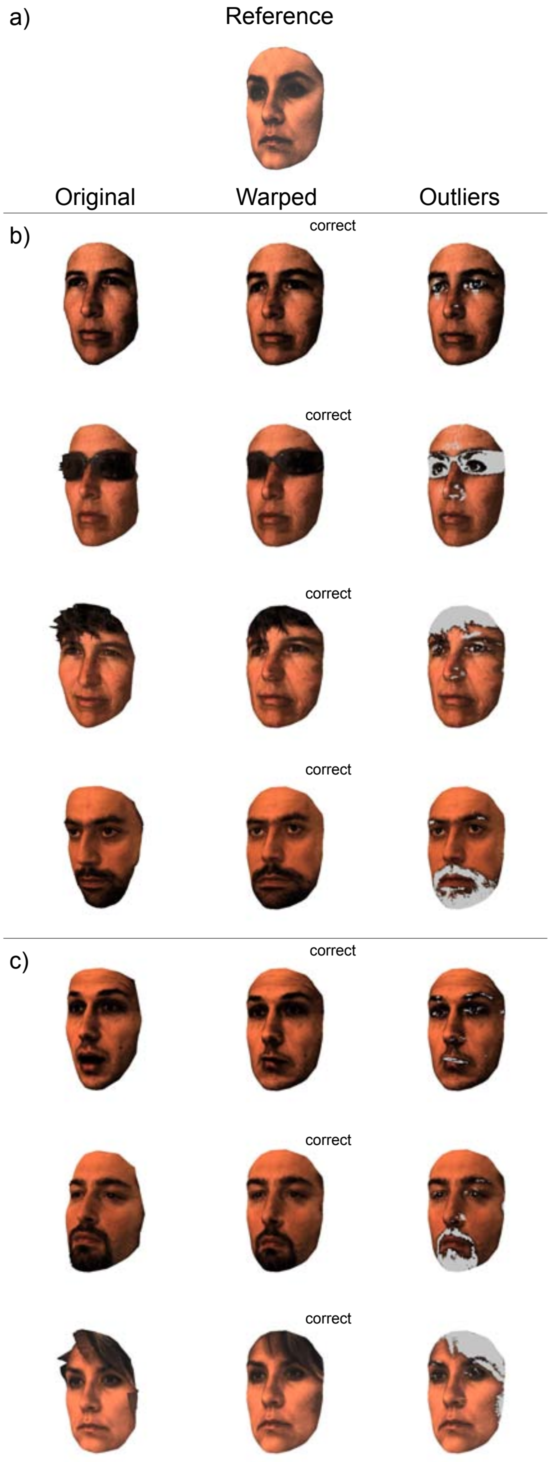

Figure 3. Left column: Several facial surfaces of the Bosphorus Database [14]. Middle column: Texture of those facial surfaces mapped to the shape of the reference facial surface, that is shown in (a). Right column: Outliers painted in silver gray. (b) shows that all faces from Fig. 2 could be registered properly. In (c) some more faces with outliers are depicted.

average distance	landmarks at eye brows	landmarks at eyes	landmarks at nose	landmarks at mouth	landmarks at chin	overall
KLT algorithm as in [3]	0.0522	0.0548	0.0517	0.0582	0.0422	0.0537
our method without weight map	0.0539	0.0585	0.0537	0.0457	0.0460	0.0521
our method with weight map	0.0507	0.0539	0.0501	0.0450	0.0460	0.0494
our method with weight map and intelligent initialization	**0.0474**	**0.0505**	**0.0457**	**0.0391**	**0.0399**	**0.0450**

Table 1. Average distance between the point on the facial surface to register that corresponds to a certain landmark in the reference face and the true coordinates of this landmark in the facial surface to register (number of landmarks: 22). The distances are normalized to the height of the reference face. A global cost function with a weight map in combination with an intelligent initialization could decrease the average distance by 16% compared to a registration with the Kanade-Lucas-Tomasi (KLT) algorithm.

gate their influence on the results. It is also planned to combine our method with an anterior rigid 3D registration step to be able to register faces that are recorded from a lateral point of view.

References

[1] M. J. Black. *Robust incremental optical flow*. PhD thesis, New Haven, CT, USA, 1992.

[2] M. J. Black and P. Anandan. The robust estimation of multiple motions: parametric and piecewise-smooth flow fields. *Computer Vision and Image Understanding*, 63(1):75–104, 1996.

[3] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. In *Proc. ACM SIGGRAPH*, pages 187–194, 1999.

[4] A. Bronstein, M. Bronstein, and R. Kimmel. Calculus of nonrigid surfaces for geometry and texture manipulation. *IEEE Trans. Visualization and Computer Graphics*, 13(5):902–913, 2007.

[5] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In T. Pajdla and J. Matas, editors, *Europ. Conf. Computer Vision*, volume 3024 of *LNCS*, pages 25–36. Springer, May 2004.

[6] M. S. Floater and K. Hormann. Surface parameterization: a tutorial and survey. In *Advances in Multiresolution for Geometric Modelling*, pages 157–186, 2005.

[7] P. Hellier, C. Barillot, E. Mémin, and P. Pérez. Hierarchical estimation of a dense deformation field for 3D robust registration. *IEEE Trans. Medical Imaging*, 20(5):388–402, May 2001.

[8] B. K. P. Horn and B. G. Schunk. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981.

[9] B. Lévy, S. Petitjean, N. Ray, and J. Maillo t. Least squares conformal maps for automatic texture atlas generation. In *Proc. ACM SIGGRAPH*, Jul 2002.

[10] N. Litke, M. Droske, M. Rumpf, and P. Schröder. An image processing approach to surface matching. In *Proc. Eurographics*. Eurographics Association, 2005.

[11] J. Y. Noh and U. Neumann. Expression cloning. In *Proc. ACM SIGGRAPH*, pages 277–288, New York, NY, USA, 2001.

[12] E. Praun, W. Sweldens, and P. Schröder. Consistent mesh parameterizations. In E. Fiume, editor, *Proc. ACM SIGGRAPH*, pages 179–184, 2001.

[13] W. H. Press, S. A. Teukolsky, W. T. Vettering, and B. P. Flannery. *Numerical Recipes in C++. The Art of Scientific Computing.: The Art of Scientific Computing*. Cambridge University Press, 2nd edition, 2002.

[14] A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun. Bosphorus database for 3d face analysis. pages 47–56, 2008.

[15] A. Savran and B. Sankur. Non-rigid registration of 3D surfaces by deformable 2D triangular meshes. *Computer Vision and Pattern Recognition Workshop*, 0:1–6, 2008.

[16] R. W. Sumner and J. Popović. Deformation transfer for triangle meshes. In *Proc. ACM SIGGRAPH*, pages 399–405, 2004.

[17] C. Tomasi and T. Kanade. Detection and tracking of point features. Technical report, Carnegie Mellon University, April 1991.

[18] Y. Wang, M. Gupta, S. Zhang, S. Wang, X. Gu, D. Samaras, and P. Huang. High resolution tracking of non-rigid 3D motion of densely sampled data using harmonic maps. *Int. Conf. Computer Vision*, 1:388–395, 2005.

[19] L. Zhang, N. Snavely, B. Curless, and S. M. Seitz. Spacetime faces: high resolution capture for modeling and animation. *ACM Trans. Graphics*, 23(3):548–558, August 2004.

[20] G. Zigelman, R. Kimmel, and N. Kiryati. Texture mapping using surface flattening via multidimensional scaling. *IEEE Trans. Visualization and Computer Graphics*, 8(2):198–207, 2002.