PATCH-BASED RECONSTRUCTION AND RENDERING OF HUMAN HEADS

David C. Schneider[1], Anna Hilsmann[1], Peter Eisert[1,2]

[1]Fraunhofer Heinrich Hertz Institute
Computer Vision and Graphics Group
Einsteinufer 37, 10587 Berlin, Germany

[2]Department of Computer Science
Humboldt Universität zu Berlin
Unter den Linden 6, 10099 Berlin, Germany

ABSTRACT

Reconstructing the 3D shape of human faces is an intensively researched topic. Most approaches aim at generating a closed surface representation of geometry, i.e. a mesh, which is texture-mapped for rendering. However, if free viewpoint rendering is the primary purpose of the reconstruction, representations other than meshes are possible. In this paper a coarse patch-based approach to both reconstruction and rendering is explored and applied not only to the face but the whole human head. The approach has advantages on parts of the scene that are traditionally difficult to reconstruct and render, which is the case for hair when it comes to human heads.

In the paper, reconstruction of a patch is posed as a parameter estimation problem which is solved in a generic image-based optimization framework using the Levenberg-Marquard algorithm. In order to improve robustness, the Huber error metric is used and a geometric regularization strategy is introduced. Initial values for the optimization, which are crucial for the method's success, are obtained by triangulation of SIFT feature points and a recursive expansion scheme.

Index Terms— 3D reconstruction, optimization, face processing, image based rendering

1. INTRODUCTION

Reconstructing the 3D shape of human faces is an intensively researched topic with applications ranging from biometry to movie special effects. Most approaches aim at generating a geometrically precise closed surface representation of the face's geometry, i.e. a mesh, which can be texture-mapped for rendering. However, if rendering from a new viewpoint is the primary purpose of the reconstruction, representations other than meshes are possible. In this paper, we explore a coarse patch-based approach to reconstruction and rendering, not only of the face but the whole human head. By a *patch* we mean a convex quadrilateral image area that is assumed to be the projection of a planar part of the 3D scene; we also assume a patch to be visible in multiple views. Patches are rendered as a cloud of textured quadrilaterals without connectivity.

This approach has significant advantages when it comes to parts of a scene that are traditionally difficult for exact reconstruction methods due to their intricate geometry, their complex interaction with light (which interferes, for example, with projected patterns), and their self-similarity in the images (which fools keypoint detectors like SIFT). For human heads, this is the case for hair, which limits many reconstruction techniques to the (unbearded) face. Note that this paper presents first results of a work in progress. We believe that they prove the approach to be feasible and promising.

The paper is structured as follows. After a brief overview of related work, we outline our optimization framework in section 3. In section 4 we discuss in detail how this framework can be used to reconstruct image patches in a robust manner. In section 5 we address more practical but crucial issues of initialization and describe the overall procedure for reconstruction and rendering of a head.

authors' emails: david.schneider@hhi.fraunhofer.de
anna.hilsmannt@hhi.fraunhofer.de
peter.eisert@hhi.fraunhofer.de

2. RELATED WORK

The literature on 3D reconstruction, even if limited to faces, is too vast to be reviewed here. This work is primarily inspired by the patch-based approach to classical, mesh-oriented multiview stereo by Furukawa et al. [1, 2]. In contrast to our approach, they use extremely small patches and a large number of views of objects with high texture detail. Nonlinear optimization is employed in [1] while an approximation of the true warp function is used in [2]. Their initialization strategy resembles ours in the use of feature points (Harris corners and DoG in [1], Lowe's SIFT [3] in this work). However, on objects with high detail the distribution of features is more favorable than on faces and hair, which is highly self-similar.

Image-based optimization frameworks similar to ours have been used for a variety of parameter estimation problems such as image registration [4], tracking of deformable objects [5, 6], or monocular 3D tracking [7]. Often the methodology is described in a problem-oriented, bottom up manner. We chose a top-down description that clarifies the relation to standard nonlinear optimization. In most works, as in ours, the optimization is computed with the Gauss Newton algorithm. We use it with Levenberg Marquardt regularization [8] and the robust error metric of Huber [9, 10]. Our review of multiview geometry of planar patches follows Hartley and Zisserman [11].

3. OPTIMIZATION FRAMEWORK

We treat patch reconstruction as an optimization problem which is stated and solved in a robust least squares framework based on M-estimators. In the following, this framework is described in an abstract manner. The application to patches is addressed in section 4.

Be $\mathcal{I}, \mathcal{K} : \mathbb{R}^2 \rightarrow \mathbb{R}$ two images, regarded as mappings from coordinates to intensities. Optimization is based on the *brightness constancy assumption* which states that all intensity differences between $\mathcal{I}$ and $\mathcal{K}$ can be explained by pixels moving to another location without changing their intensity. Introducing a "warp" function $\mathcal{W} : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ that describes the pixel displacements, we can formalize this as

$$\mathcal{I}(\mathbf{x}) - \mathcal{K}(\mathcal{W}(\mathbf{x})) = 0 \tag{1}$$

for all pixel locations $\mathbf{x}$. If our warp is a parametric model $\mathcal{W} : \left(\mathbb{R}^2, \mathbb{R}^K\right) \to \mathbb{R}^2$ with K parameters, the parameter estimation problem can be stated as

$$\arg\min_{\mathbf{v}} \sum_{i=1}^{N} \psi\left\{\mathcal{I}\left(\mathbf{x}_i\right) - \mathcal{K}\left(\mathcal{W}\left(\mathbf{x}_i; \mathbf{v}\right)\right)\right\} \tag{2}$$

where the summation is over N pixel locations (e.g. of a patch) and ψ defines the error metric to use. With $\psi\left(x\right) = \frac{1}{2}x^2$ we get the least squares metric, with

$$\psi\left(x\right) = \begin{cases} {}^1\!/_2\, x^2 & |x| \leq \theta \\ \theta\left(|x| - {}^\theta\!/_2\right) & \text{otherwise} \end{cases} \tag{3}$$

we get the Huber estimator [9], which was used to produce the results shown in the paper. The threshold θ is determined with the median absolute deviation (MAD, [12]).

Note that $\mathbf{v}$ is the only unknown in equation (2), which is a nonlinear least-squares problem with residuals $r_i = \mathcal{I}\left(\mathbf{x}_i\right) - \mathcal{K}\left(\mathcal{W}\left(\mathbf{x}_i; \mathbf{v}\right)\right)$. The Jacobian of the residual vector $\left[r_1 \dots r_N\right]^T$ is

$$\mathbf{J} = -\begin{bmatrix} \nabla\mathcal{K}\left(\mathcal{W}\left(\mathbf{x}_1; \mathbf{v}\right)\right)^T \cdot \mathbf{J}_1 \\ \vdots \\ \nabla\mathcal{K}\left(\mathcal{W}\left(\mathbf{x}_N; \mathbf{v}\right)\right)^T \cdot \mathbf{J}_N \end{bmatrix} \tag{4}$$

where $\nabla\mathcal{K}$ is the gradient of image $\mathcal{K}$ evaluated at a "warped" location and

$$\mathbf{J}_i = \frac{\partial}{\partial \mathbf{v}} \mathcal{W}\left(\mathbf{x}_i; \mathbf{v}\right) \tag{5}$$

are the Jacobians of the warp *at each pixel*.

With a Jacobian defined, quasi-Newton type solvers can be used to estimate parameters. For M-estimators, the problem can be transformed to an iteratively reweighted least squares problem with a weight update in each optimization step. Currently, we use the Huber estimator [9] for robustness and the Levenberg-Marquard algorithm [8] for optimization.

4. RECONSTRUCTION OF PATCHES

4.1. Multiview geometry of patches

We regard a patch as a rectangular area in one view and a convex quadrilateral in all other views and in 3D space. In this section we briefly review the related geometry. Note that homogeneous coordinates are used throughout the paper. $\mathbb{P}^n$ denotes the projective space of dimension n, i.e. vectors of $\mathbb{P}^n$ have $n+1$ elements.

A camera is called *canonical* if its matrix is $[\mathbf{I}_3\, \mathbf{0}]$. Be $\mathbf{P} \in \mathbb{R}^{3\times4}$ a camera matrix and define a 4×4 matrix $\mathbf{C} := \left[\begin{smallmatrix}\mathbf{P} \\ \mathbf{c}\end{smallmatrix}\right]$ where $\mathbf{c}$ is chosen such that $\operatorname{rank}\left(\mathbf{C}\right) = 4$. Then $\mathbf{C}^{-1}$ transforms the camera $\mathbf{P}$ into a coordinate frame, where it is canonical, which we call the *canonical frame* in the following:

$$\mathbf{P}\mathbf{C}^{-1} = \left[\mathbf{I}_3\, \mathbf{0}\right] \tag{6}$$

Also, $\mathbf{C}$ transforms any world point $\mathbf{y} \in \mathbb{P}^3$ into the canonical frame:

$$\mathbf{P}\mathbf{y} = \mathbf{P}\left(\mathbf{C}^{-1}\mathbf{C}\right)\mathbf{y} = \left[\mathbf{I}_3\, \mathbf{0}\right]\left(\mathbf{C}\mathbf{y}\right) \tag{7}$$

Moreover, $\mathbf{C}^{-T}$ transforms planes into the canonical frame: If $\boldsymbol{\pi} \in \mathbb{P}^3$ is a world plane, then each point on $\boldsymbol{\pi}$ satisfies $\boldsymbol{\pi}^T\mathbf{y} = 0$. Transforming this into the canonical frame yields:

$$\boldsymbol{\pi}^T\mathbf{y} = \boldsymbol{\pi}^T\left(\mathbf{C}^{-1}\mathbf{C}\right)\mathbf{y} = \left(\mathbf{C}^{-T}\boldsymbol{\pi}\right)^T\left(\mathbf{C}\mathbf{y}\right) \tag{8}$$

Now be $\mathbf{P}_A$ and $\mathbf{P}_B$ two cameras. Chose $\mathbf{C}$ such that $\mathbf{P}_A\mathbf{C}^{-1} = [\mathbf{I}_3\, \mathbf{0}]$ and define $\mathbf{P}_B\mathbf{C}^{-1} =: [\mathbf{S}\ \mathbf{t}]$. Be $\boldsymbol{\pi} \in \mathbb{P}^3$ and assume without loss of generality that the plane transformed to the canonical frame satisfies

$$\mathbf{C}^{-T}\boldsymbol{\pi} =: \begin{bmatrix}\mathbf{v}^T\ 1\end{bmatrix}. \tag{9}$$

In the canonical frame, an image point $\mathbf{x} \in \mathbb{P}^2$ in camera *A* back-projects to a ray $\begin{bmatrix}\mathbf{x}^T\ \rho\end{bmatrix}^T$ parametrized by ρ since

$$\forall\rho: \quad \left[\mathbf{I}_3\ \mathbf{0}\right]\left[\begin{smallmatrix}\mathbf{x} \\ \rho\end{smallmatrix}\right] = \mathbf{x}. \tag{10}$$

Assume that $\mathbf{x}$ is a projection of a point $\mathbf{y} \in \mathbb{P}^3$ on the plane $\boldsymbol{\pi}$. Then $\mathbf{y}$ satisfies

$$\mathbf{y}^T\left[\begin{smallmatrix}\mathbf{v} \\ 1\end{smallmatrix}\right] = \begin{bmatrix}\mathbf{x}^T\ \rho\end{bmatrix}\left[\begin{smallmatrix}\mathbf{v} \\ 1\end{smallmatrix}\right] = 0 \tag{11}$$

and $\mathbf{y} = \left[\begin{smallmatrix}\mathbf{x} \\ -\mathbf{x}^T\mathbf{v}\end{smallmatrix}\right]$. The image of $\mathbf{y}$ in camera *B* is

$$\left[\mathbf{S}\ \mathbf{t}\right]\left[\begin{smallmatrix}\mathbf{x} \\ -\mathbf{x}^T\mathbf{v}\end{smallmatrix}\right] = \left(\mathbf{S} - \mathbf{t}\mathbf{v}^T\right)\mathbf{x}. \tag{12}$$

Note that $\left(\mathbf{S} - \mathbf{t}\mathbf{v}^T\right)$ is a homography that directly maps images of points on $\boldsymbol{\pi}$ in view *A* to their correspondences in view *B*.

4.2. Warp function

In order to reconstruct a patch, we require a warp and its Jacobian that is parametrized by the plane parameters $\mathbf{v}$. Then $\mathbf{C}^T\begin{bmatrix}\mathbf{v}^T\ 1\end{bmatrix}$ is the parameter vector of the patch's world plane in the non-canonical world frame. Note that for a point $\mathbf{x} \in \mathbb{P}^2$ in view *A*, the warp is not simply $\left(\mathbf{S} - \mathbf{t}^T\mathbf{v}\right)\mathbf{x}$ as we need to convert to non-homogeneous coordinates to use the warp in the optimization framework. Therefore, be

$$\mathbf{S} - \mathbf{t}\mathbf{v}^T =: \begin{bmatrix}\mathbf{s}_1^T \\ \mathbf{s}_2^T \\ \mathbf{s}_3^T\end{bmatrix} - \begin{bmatrix}t_1 \\ t_2 \\ t_3\end{bmatrix}\mathbf{v}^T. \tag{13}$$

Then the warp is given by

$$\mathcal{W}\left(\mathbf{x}_i; \mathbf{v}\right) = \begin{bmatrix} \frac{\mathbf{s}_1^T\mathbf{x}_i - t_1\mathbf{v}^T\mathbf{x}_i}{\mathbf{s}_3^T\mathbf{x}_i - t_3\mathbf{v}^T\mathbf{x}_i} \\ \frac{\mathbf{s}_2^T\mathbf{x}_i - t_2\mathbf{v}^T\mathbf{x}_i}{\mathbf{s}_3^T\mathbf{x}_i - t_3\mathbf{v}^T\mathbf{x}_i} \end{bmatrix}. \tag{14}$$

To derive the warp Jacobian with respect to $\mathbf{v}$, define

$$s_i := \mathbf{s}_i^T\mathbf{x}, \quad i = 1 \dots 3 \tag{15}$$

as this is a scalar unaffected by $\mathbf{v}$. Then the Jacobian of $\mathcal{W}\left(\mathbf{x}_i; \mathbf{v}\right)$ is

$$\mathbf{J}_i = \begin{bmatrix} \frac{(t_3 s_1 - t_1 s_3)}{\left(s_3 - t_3\mathbf{x}_i^T\mathbf{v}\right)^2}\mathbf{x}_i^T \\ \frac{(t_3 s_2 - t_2 s_3)}{\left(s_3 - t_3\mathbf{x}_i^T\mathbf{v}\right)^2}\mathbf{x}_i^T \end{bmatrix}. \tag{16}$$

4.3. More views

In order to improve quality and stability of the solutions, it is advisable to use more than two views in the optimization. This can be incorporated seamlessly in the optimization framework: In equation (14), the quantities $\mathbf{s}_i$ and t_i relate to the camera of a specific view (transformed into the canonical frame), while $\mathbf{v}$ relates to the patch seen in all views.

To optimize over V views, chose one reference image $\mathcal{I}$ and denote by $\mathcal{K}_j$, $j = 1 \dots V-1$ the other images. Be $\mathcal{W}_j$ the warp

of equation (14) with camera parameters of $\mathcal{K}_j$'s view. Then the optimization problem becomes

$$\arg\min_{\mathbf{v}} \sum_{j=1}^{V-1} \sum_{i=1}^{N} \psi \left\{ \mathcal{I}\left(\mathbf{x}_i\right) - \mathcal{K}_j\left(\mathcal{W}_j\left(\mathbf{x}_i; \mathbf{v}\right)\right) \right\}. \qquad (17)$$

Note that this is still a least squares problem in $\mathbf{v}$. From an optimization point of view, the only thing changed in comparison to equation (2) is the number of residuals.

4.4. Geometric regularization

Even with more than two views, patches with low magnitudes in the image gradient can go astray. In these cases, the optimization can be regularized by introducing additional cost terms to the error function. In the following we describe a cost term that can be used to incorporate geometric *a priori* knowledge.

Be $\mathbf{x} \in \mathbb{P}^2$ an image point in a patch that is not reconstructed and be $\mathbf{y} \in \mathbb{P}^3$ the (unknown) world point on the 3D patch that projects to $\mathbf{x}$. Assume that there is a world point $\mathbf{z} \in \mathbb{P}^3$—e.g. on a patch that has already been reconstructed—to which $\mathbf{y}$ should be close. According to equation (11), $\mathbf{y}^T = \begin{bmatrix} \mathbf{x}^T & -\mathbf{x}^T\mathbf{v} \end{bmatrix}$. Thereby a cost term favors patches $\mathbf{v}$ which bring $\mathbf{y}$ close to $\mathbf{z}$ is:

$$c\left(\mathbf{v}\right) = \lambda \left\| \mathbf{C}^{-1} \begin{bmatrix} \mathbf{x} \\ -\mathbf{x}^T\mathbf{v} \end{bmatrix} - \mathbf{z} \right\|^2 \qquad (18)$$

where λ weights the influence of the regularization. Note that $\mathbf{y}$ is transformed out of the canonical frame by $\mathbf{C}^{-1}$ before the error is computed as distances in the canonical frame are heavily distorted. Again, the cost term can be added to the overall optimization problem without changing its structure.

5. OVERALL RECONSTRUCTION AND INITIALIZATION

In the previous sections we outlined our optimization approach and its application to the estimation of patches. In this section we take a step back and describe the overall framework in which the reconstruction is applied to multi-view imagery of human heads with the goal of rendering them from new viewpoints.

We use a database of multi-view head images in XGA resolution that were shot simultaneously in a fully calibrated wide baseline rig. We use three views at a time with cameras equally spaced on a circle and roughly pointing towards the center of the head; see figure 1 for an example. The simultaneous use of more cameras or the integration of reconstructions from multiple view triplets have not yet been addressed.

A patch in our current approach is a quadratic area in the image of the central camera, which we call the *reference view* in the following. Image $\mathcal{I}$ in equation (2) is always the reference image. The patch centers lie on a uniform grid and patches overlap with their direct neigbors by 50%. For rendering, each reconstructed patch is displayed by two triangles and textured with its image region in the reference view.

Note that patch size has a contrary effect on reconstruction and rendering: Large patches are more likely to contain sufficient gradient information for stable reconstruction but they approximate the actual geometry worse than small patches.

Besides the amount of structure in a patch, the choice of the initial value for the optimization is crucial for the quality of the result as nonlinear optimization is prone to local minima of the respective error function. Issues of initialization are closely related to the order in which patches are reconstructed. Currently, the following three stage strategy is used.

Fig. 1. Typical multiview input to the algorithm.

Stage 1 In the first stage, feature points are extracted in all views using the SIFT algorithm [3]. The features are filtered by their descriptors; only those with matching descriptors over all image pairs are retained. The remaining features are then tested for epipolar consistency and inconsistent features are discarded. Finally, for each set of corresponding features a 3D point is computed by minimizing the back-projection error in all views. For each patch that contains a feature, the initial value of $\mathbf{v}$ is chosen such that the triangulated 3D feature lies on the 3D patch and its normal points to the reference camera. If a patch contains multiple features, the one closest to the patch center in the reference view is chosen. All patches that contain features are reconstruced independently with a coarse grid (30 pixel step) and a large patch size (60 pixels side length). In the following, be $\mathcal{F}$ the set of those patches.

Stage 2 For face images in XGA resolution, the distribution of SIFT features is highly uneven. Increasing sensitivity and tolerance of the feature extractor can increase the amount of feature points but does not improve their distribution. Therefore, most patches do not contain a feature and need to be initialized differently. Denote by $\mathcal{S}$ the set of unreconstructed, featureless patches and by $\mathcal{R}$ the set of reconstructed featureless patches which is empty in the beginning. To reconstruct the patches in $\mathcal{S}$, a patch $P \in \mathcal{S}$ with minimal distance to its nearest neighbor in $\mathcal{F}$ is chosen. All neighbors of P in $\mathcal{F} \cup \mathcal{R}$ are identified, and the initialization for P is found by averaging the neighbor's normal orientations and distances to the origin in the non-canonical frame. Patch P is then reconstructed using the geometric regularization described in section 4.4: A cost term is added for each corner of P that is shared with a reconstructed neighbor in $\mathcal{F} \cup \mathcal{R}$. In terms of equation (18), $\mathbf{x}$ is the corner coordinate of the candidate patch in the reference view and the target point $\mathbf{z}$ is the 3D corner of the reconstructed neighbor. The reconstructed patch is moved from $\mathcal{S}$ to $\mathcal{R}$ and the procedure is iterated. Stage 2 uses the same coarse grid and patch size as stage 1.

Stage 3 For small viewpoint changes the results of stage 2 produce good renderings. For more radical changes, the patch size is too large. Therefore, in stage 3 the grid is subdivided to a step length of 10 pixels and patch size of 20 pixels. Each patch is initialized with the parameters of its subdivision parent and reconstructed independently. Regularization is not used in stage 3.

6. RESULTS AND CONCLUSION

Figure 2 shows rendering results for a variety of faces reconstructed with the method described in the previous section. At its current

Fig. 2. Algorithm results: Patch-based renderings from different perspectives.

state, the approach allows for significant yet not radical changes of viewpoint. The results show that the proposed technique is indeed capable of reconstructing hair and rendering it convincingly from new viewpoints.

The key issues that have to be resolved in order to improve the quality of the results and increase the choice of possible viewpoints are the following:

Firstly, patches with little gradient information may converge to rather arbitrary 3D positions and small patches are more likely to be affected by this problem than large ones. This is currently addressed by geometric regularization from successfully reconstructed neighbors. However, if the neighbors are erroneous themselves, the errors propagate. This could be improved by developing a measure for the quality of a reconstruction. Note that the remaining MSE is a bad predictor as it is generally low on the problematic patches.

Secondly, a bad choice of initial values can spoil the optimization regardless of the quality of the patch. There is great potential for improvement in finding better strategies for propagating initialization information.

Finally, the rendering side is another interesting area for further research. Introducing transparency in patches or non-quadrilateral (yet still planar) patch shapes should greatly improve the visual quality of the results. Special treatment is required for patches that touch the border between head and background. Here, robust methods are especially promising. Generally we believe that looking at reconstruction and rendering simultaneously is a fruitful approach to vision and graphics of human heads.

7. REFERENCES

[1] Yatsuka Furukawa and Jean Ponce, "Accurate, Dense, and Robust Multi-View Stereopsis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 1, pp. 1–14, 2008.

[2] Yasutaka Furukawa and Jean Ponce, "High-Fidelity Image Based Modeling," Tech. Rep. Technical Report 2006-02, INRIA Rhone-Alpes, 2006.

[3] David G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, 2, pp. 91–110, 2004.

[4] Adrien Bartoli and Andrew Zisserman, "Direct Estimation of Non-Rigid Registrations," in *Proc. Of The Fifteenth British Machine Vision Conference*, 2004.

[5] A. Hilsmann and P. Eisert, "Joint Estimation of Deformable Motion and Photometric Parameters in Single View Video," in *ICCV Workshop on Non-Rigid Shape Analysis and Deformable Image Alignment*, Kyoto, Japan, September 2009.

[6] V. Gay-Bellile, A. Bartoli, and P. Sayd, "Direct Estimation of Non-Rigid Registrations with Image-Based Self-Occlusion Reasoning," *IEEE Trans on Pattern Analysis and Machine Intelligence*, vol. 32, no. 1, pp. 87–104, January 2010.

[7] Peter Eisert and Juergen Rurainsky, "Image-based Rendering and Tracking of Faces," in *Proc. International Conference on Image Processing*, 2005.

[8] Donald Marquardt, "An algorithm for least-squares estimation of nonlinear parameters," *SIAM Journal of Applied Mathematics*, vol. 11, pp. 431–441, 1963.

[9] Peter J. Huber, *Robust Statistics*, John Wiley & Sons, 1981.

[10] Zhengyou Zhang, "Parameter Estimation Techniques: A Tutorial with Application to Conic Fitting," *Image and Vision Computing Journal*, vol. 15, Nr. 1, pp. 59–76, 1997.

[11] Richard Hartley and Andrew Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, 2nd edition, 2003.

[12] David C. Hoaglin, Frederick Mosteller, and John W. Tukey, *Understanding Robust and Exploratory Data Analysis*, John Wiley & Sons, 1983.