Semantic 3D Reconstruction of Heads

Fabio Maninchedda[1(✉)], Christian Häne[2], Bastien Jacquet[3], Amaël Delaunoy[1], and Marc Pollefeys[1,4]

[1] ETH Zurich, Zurich, Switzerland
fabiom@inf.ethz.ch
[2] UC Berkeley, Berkeley, USA
[3] Kitware SAS, Villeurbanne, France
[4] Microsoft, Redmond, USA

Abstract. We present a novel approach that jointly reconstructs the geometry of a human head and semantically segments it into labels such as skin, hair and eyebrows. In order to get faithful reconstructions from data captured in uncontrolled environments, we propose to adapt a recently introduced implicit volumetric surface normal based shape prior formulation. Shape prior based approaches critically rely on an accurate alignment between the data and the prior to succeed. To this end, we propose an automatic alignment procedure for the used shape prior formulation. We evaluate our alignment procedure thoroughly and show head reconstruction results on challenging datasets.

Keywords: Face · Head · Semantic · Multi-label · Shape prior · Alignment

1 Introduction

Reconstruction of human faces and heads is an ongoing topic in computer vision and related areas. There is much interest due to the wide field of applications and the inherent difficulty of the problem. Use cases are for example content generation for movie production, computer games, virtual make over, physical manufacturing of figurines, i.e. 3D printing, and many more. Due to the wide range of applications many different capturing technologies are utilized in practice. When generating content for movies, high quality capturing setups that facilitate a very accurate geometry acquisition are the natural choice [2]. However, this is expensive and needs expert knowledge during the capturing process. In this paper we focus on less constrained scenarios, such as a person taking a 3D selfie [32] or a person capturing a 3D head model of another person by using a

C. Häne, B. Jacquet and A. Delaunoy—Work done while authors were at the Department of Computer Science, ETH Zurich.

Electronic supplementary material The online version of this chapter (doi:10.1007/978-3-319-46466-4_40) contains supplementary material, which is available to authorized users.

B. Leibe et al. (Eds.): ECCV 2016, Part VI, LNCS 9910, pp. 667–683, 2016.
DOI: 10.1007/978-3-319-46466-4_40

hand held camera. Therefore, there is little control over the conditions in which the images are taken. They can be badly exposed, blurry and are generally of lower quality than with a dedicated capturing setup. A common way to address these issues is to use shape priors [7,13,23].

For many applications also semantic labels are of interest. In video games the hair of characters can be physically simulated in real time. Being able to generate a semantically segmented 3D model would directly facilitate such a simulation on user generated content. Similarly, for 3D printing different semantic labels could be manufactured with different materials. For augmented reality the head could be augmented with a hat which would interact with the hair, but not affect the shape of the head. For such applications not only the visible surfaces, such as skin or hair, need to be modeled but also the hidden, invisible surfaces, for example the surface between skin and hair needs to be estimated convincingly. In this paper, we show that this can be achieved by posing the reconstruction of human heads as a volumetric multi-label segmentation problem [15] together with a multi-label shape prior [14]. When using shape priors one has to establish the correspondence between the input data and the shape prior and eventually recover a good alignment between them. To this end, we propose a novel alignment procedure that allows us to align the implicit volumetric shape prior of [14] fully automatically to the input data. In previous work the alignment for this type of shape prior was done manually.

1.1 Related Work

Using synchronized, high resolution multi-camera systems in controlled environments with good lighting, high quality face models can be acquired by stereo matching [2,8]. An extension [3], estimates facial hair as separate layer. A skin surface is always present underneath the hair, however it is only a pseudo surface which is not meant to be a plausible reconstruction of the unobserved surface.

In uncontrolled environments where data is captured with lower resolution, face reconstruction is often achieved by fitting a blend shape to the images. A classical way is to generate a statistical shape model (of faces) [4,23,26], which is fitted into the input data. First, facial landmarks [10,17,27,28] are extracted, which are then used to register the input images to the shape model. Using such a blend shape model with additional refinements, [12,16] focus on reconstructing dynamic face models. Even though realistic reconstructions are obtained using a low-dimensional statistical shape model, they generally do not capture instance specific shape variations, such as big moles. Also for 3D reconstruction of hair methods that exploit the specific structure of hair were proposed [22,34]. Most of the methods focus on reconstructing either the face or the hair. In this work, we reconstruct complete, printable, 3D models of human *heads* similar to [8]. While [8] uses a similar capturing setup as [2], we tackle the challenging problem of working with images captured using a hand-held camera, e.g. a mobile phone or a compact camera. We achieve this with volumetric multi-label formulation.

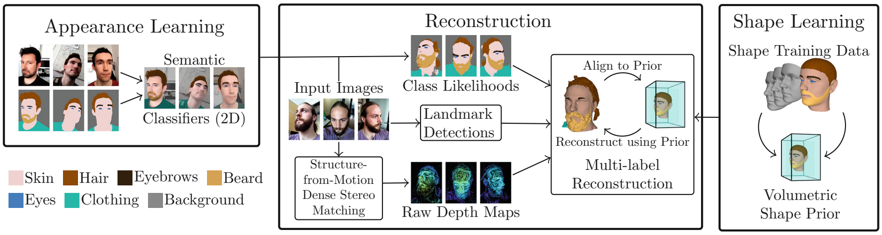


Fig. 1. Overview of our method.

Volumetric 3D reconstruction dates back to [6]. A voxel space is labeled into *free space* and *occupied space*. Regularizing the input data by penalizing the surface area was proposed in [20,36] for the discrete graph-based and the spatially continuous (variational) formulation, respectively. Continuous formulations for multi-label segmentation have been proposed in [5,37,38]. Instead of using a single *occupied space* label, [15] proposed to use multiple semantic classes to segment the occupied space. This continuously inspired method, penalizes transitions between different labels anisotropically and can therefore include priors on the direction of the surfaces. The idea of using anisotropic surface area penalization in the continuous setting [15], was extended in [14] to describe 3D object shape priors, learned from training data, in form of an implicit normal direction based shape prior. This leads to a very powerful object shape prior, however the alignment between the prior and the input data is assumed to be given as input.

1.2 Contributions

Our main contributions are the following:

- We present a system which reconstructs and semantically segments human heads from images captured with standard hand-held cameras in uncontrolled environments. In contrast to previous systems we do not only reconstruct the geometry of the head but also acquire a semantic segmentation into classes such as skin, hair, beard and eyebrows. This includes a plausible reconstruction of the unobserved surfaces (e.g. skin underneath hair). Moreover, our system is able to recover instance specific shape details which are typically lost when using a low-dimensional statistical shape model.
- We propose an automatic alignment procedure for the implicit shape prior formulation of [14], which was considered as an input in the original publication and hence done manually. Our key insight is that despite the volumetric nature of the shape prior we can formulate the alignment as an optimization over the surface. We propose an optimization scheme which alternates between optimizing for the geometry and the alignment. Despite the non-convexity of the optimization we can robustly infer the geometry and the alignment. This part is detailed in Sect. 4.1.

Moreover, we propose generalizations and modifications to the used formulations:

- In Sect. 3.2 we present a data term which allows for thin layers of semantic classes without the additional complexity of ray potentials [29,30]. The idea is to represent parts of the input data in the regularization term, instead of fully representing the data cost as unary terms as proposed in [15].
- The implicit normal direction based shape prior, discretizes the normals regularly over all directions. However, often the training data locally suggest just one single or very few predominant directions with little variation. We propose to detect and exploit this for a more efficient formulation, as explained in Sect. 3.3.

1.3 Overview

Figure 1 illustrates our reconstruction pipeline. In the training part of the method, we train an image based semantic classifier and the volumetric shape prior. From the input images camera poses and depth maps are computed through structure from motion and subsequent dense matching. For each of the images pixel-wise semantic likelihoods are obtained by running the trained semantic classifier. An approximate alignment of the input data to the shape prior is based on detecting landmarks around the eyes, nose and mouth in the input data. The core of our method is an optimization with respect to both the geometry and the alignment.

2 Optimization Problem

Our method is based on a volumetric multi-label problem, formulated as a convex optimization that does usually not include an alignment. We propose to include the alignment into the formulation leading to an energy which is convex with respect to the labeling and non-convex with respect to the alignment. The actual choices for the unary cost and the regularization term will be detailed in Sect. 3, they are based on pixel-wise semantic classifications and depth maps.

Mathematically, we have a voxel space Ω, understood as discretization of a subset of $\mathbb{R}^3$. Each voxel gets assigned a label $\ell \in \mathcal{L}$. Indicator variables $x_s^i \in [0,1]$, indicate if label i is assigned at voxel s. In addition to the original formulation, we propose to include a similarity transform $\mathcal{T}$ into the optimization problem. The transform $\mathcal{T} : \mathbb{R}^3 \to \mathbb{R}^3$, is defined as $y \mapsto \alpha R y + t$, with a positive scaling factor $\alpha > 0$, a rotation matrix R and a translation vector t.

$$E(\mathbf{x}, \mathcal{T}) = \sum_{s\in\Omega} \left(\sum_i \rho_s^i(\mathcal{T}) x_s^i + \frac{1}{\alpha^2} \sum_{i,j:i<j} \phi_s^{ij}(\mathcal{T}, x_s^{ij} - x_s^{ji}) \right) \tag{1}$$

$$\text{s. t. } x_s^i = \sum_j (x_s^{ij})_k, \quad x_s^i = \sum_j (x_{s-e_k}^{ji})_k\ , \quad \sum_i x_s^i = 1, \quad x_s^i \geq 0, \quad x_s^{ij} \geq 0.$$

Next, we intuitively explain the meaning of the formulation. A thorough derivation of the basic formulation without the alignment is given in [38]. The first line defines the objective of the minimization problem. It is split into two parts: the unary term and the regularization term. The values $\rho_s^i(\mathcal{T})$ define the cost for assigning a label i to a voxel s. The second part is a spatially varying anisotropic regularization term $\phi_s^{ij}(\cdot,\cdot) \rightarrow \mathbb{R}^+$, which is derived in the continuum and discretized afterwards [9]. It assigns a cost to a surface between labels i and j in voxel s with a surface normal pointing into the direction of $x_s^{ij} - x_s^{ji} \in [-1,1]^3$. The functions $\phi_s^{ij}(\cdot,\cdot)$ need to be convex and positively 1-homogeneous in their second argument. The variables $x_s^{ij} \in [0,1]^3$ describe how much the assignment of label i changes to label j in the direction in which they point. In order to allow for arbitrary convex non-metric smoothness terms the x_s^{ij} need to be non-negative, which limits the possible directions they can point to. This is resolved by using $x_s^{ij} - x_s^{ji}$, which allows for arbitrary directions, for details see [38]. The first two constraints, are called marginalization constraints. They connect the x_s^i and x_s^{ij} variables. k indexes the components of the vector and e_k denotes the k-th canonical basis vector, i.e. $e_1 = (1,0,0)^T$. Intuitively, these constraints describe that if label i is assigned to voxel s and label j in a neighboring voxel then the x_s^{ij} variables need to reflect such a transition. Next, the normalization constraint enforces that one label is assigned. Finally, all the x_s need to be non-negative.

As mentioned above we included the similarity transform $\mathcal{T}$ into the original convex multi-label formulation. $\mathcal{T}$ transforms the input data into the coordinate frame of the shape prior. The smoothness term is dependent on the transformation $\mathcal{T}$ because it includes parts of the data cost. The normalization of the smoothness term with respect to α^2 ensures that a change in scaling does not change the cost of the surface. This is crucial for the optimization of the alignment as we will see in Sect. 4.1.

3 Choices for ρ and ϕ

The key difficulty that needs to be tackled, when defining the unary cost and the regularization term, is thin layers of semantic classes such as eyebrows in front of the skin. It has already been pointed out in [29,30] that this is problematic when using the data term of [15] (c.f. Fig. 2). The solution given in [29,30] is a formulation which represents the dataterm as a potential over viewing rays. They propose a purely discrete graph-based scheme [30] and a continuous (variational) formulation [29]. Both versions introduce the additional complexity that also the assignment to additional per-voxel variables for each viewing ray that crosses a specific voxel needs to be determined during the optimization, which makes the optimization problem much more complex. This can be resolved using a coarse-to fine scheme in the discrete setting but remains a problem for the continuous setting. To this end, we propose an alternative representation in the continuous setting which does not add any additional variables. Our solution can be seen as an alternative to ray potentials in cases where the only feature that is needed is the representation of thin layers of semantic classes.

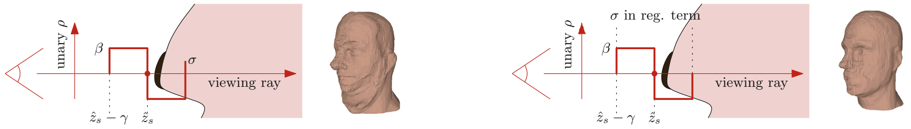


Fig. 2. Unary term for a ray going through the eyebrow next to the skin layer of an example reconstruction (Left) data term of [15]. (Right) our proposed data term. Both sides illustrate the weight added to the voxels along the ray for the class eyebrow by the unary term. (Left) The per-pixel semantic cost σ, is entered into the last voxel of the uncertainty region. In this case eyebrow is visible in the image but the weight ends up inside the skin layer due to the very little thickness of the semantic class eyebrow, which leads to artifacts in the reconstruction. (Right) in our proposed data term the weight σ is moved to the regularization term. The unary term only captures the geometric information about free and occupied space. This resolves the artifacts in the reconstruction.

3.1 Unary Term

We only include the information from the depth maps in the per voxel unary term and represent the likelihood of the semantic class in the surface regularization term. The rationale behind this is the following. The semantic classifier only gives a likelihood for which semantic label should be closest to the camera along the ray, but not where along the ray this transition from free space to occupied space happens. The depth measurement roughly tells us the region where we expect the transition. If we now decrease the smoothness cost of a transition from free space to the desired semantic label in that region, then our formulation prefers to place the observed semantic class as the transition from free to occupied space but does not affect a potential additional transition from one semantic label to another one just behind it (c.f. Fig. 2).

The unary cost $\rho_s^i(\mathcal{T})$ contains the information from the depth maps. There is one free space label $i = 0$ and several occupied space labels $i > 0$. Therefore, we have $\rho_s^i(\mathcal{T}) {:=} \rho_s(\mathcal{T})$, $\forall i > 0$ and $\rho_s^0(\mathcal{T}) {:=} 0$. We denote the non-zero unary cost that a single depth map contributes to voxel s by $\rho_s(\mathcal{T})'$, the complete unary cost is formed by summing over all the depth maps. Further, z_s is the depth of voxel s and $\hat{z}_s(\mathcal{T})$ is the depth at the depth map position to which the voxel s projects to with the alignment transformation $\mathcal{T}$. Using the assumption that in front of an observed depth we expect *free space* in a region γ and behind the observed depth *occupied space*, we set the unary cost to

$$\rho_s(\mathcal{T})' = \begin{cases} \beta & \text{if } z_s - \hat{z}_s(\mathcal{T}) \in [0, \gamma] \\ -\beta & \text{if } z_s - \hat{z}_s(\mathcal{T}) \in [-\gamma, 0). \end{cases} \qquad (2)$$

3.2 Data Dependent Regularization Term

The regularization term $\phi_s^{ij}(\mathcal{T}, n)$ describes the cost of a transition between label i and j with normal direction n. We derive our novel regularization term based on the underlying probabilities. $\leftrightarrow_s$ denotes that there is a surface at location s, $\leftrightarrow_s^{ij}$ denotes the existence of a surface between label i and j at location s and n_s^{ij} indicates that a surface with normal n between label i and j is present at location s. Finally, we denote the per pixel knowledge about the semantic labels as $\mathit{\Gamma}$ and also need a dependency on the alignment transformation $\mathcal{T}$. We start by stating the probability of a surface element as

$$P(n_s^{ij}|\mathcal{T},\mathit{\Gamma}) := P(n_s^{ij}|\leftrightarrow_s^{ij})P(\leftrightarrow_s^{ij}|\leftrightarrow_s,\mathcal{T},\mathit{\Gamma})P(\leftrightarrow_s). \tag{3}$$

The probability is modeled as a Bayesian network and factored into three parts. The rightmost term $P(\leftrightarrow_s)$ captures the probability of observing a surface at voxel s. $P(\leftrightarrow_s^{ij}|\leftrightarrow_s,\mathcal{T},\mathit{\Gamma})$ is the probability to have a surface between two specific labels i and j given there is a surface. This part includes the knowledge about the per pixel semantic labels $\mathit{\Gamma}$ in the input images and hence is dependent on the alignment $\mathcal{T}$. $P(n_s^{ij}|\leftrightarrow_s^{ij})$ takes into account the surface orientation and is essentially capturing the implicit normal direction based shape prior. In the following we will explain how we approximate the above model in our energy formulation. To simplify the notation for the rest of this section we will consider the alignment $\mathcal{T}$ to be fixed and drop it from the equations. The mathematical formulation [38] allows any convex positively 1-homogeneous function as function $\phi_s^{ij}(\cdot)$. To find a function which fulfills these properties and approximates the above model well, we rewrite it in its dual form in terms of a Wulff shape [9]. Every convex positively 1-homogeneous function can be written as

$$\phi_s^{ij}(x) = \max_{p\in\mathcal{W}_s^{ij}} \{p^T x\}. \tag{4}$$

$\mathcal{W}_s^{ij}$ is the Wulff shape. It defines the regularizer and can be any closed convex shape which contains the origin. Any convex shape can be written as intersection of half spaces. [14] proposes to use a discrete set of normal directions $n \in \mathcal{S} \subset \mathbb{S}^2$ to form a discretized Wulff shape $\mathcal{W}_{\mathcal{H}_s^{ij}}$ by intersecting the half spaces $h_s^{n,ij} \in \mathcal{H}_s^{ij}$. The distance of the half space boundary to the origin at voxel s with normal n for the boundary between i and j is denoted as $d_s^{n,ij}$. Looking at the probabilistic meaning of the energy formulation and assuming all the half spaces $\mathcal{H}_s^{ij}$ share a boundary with $\mathcal{W}_{\mathcal{H}_s^{ij}}$, it follows that

$$P(n_s^{ij}|\mathit{\Gamma}) = \exp\left(-\phi_s^{ij}(n_s^{ij})\right) = \exp\left(-\max_{p\in\mathcal{W}_{\mathcal{H}_s^{ij}}}(p^T n_s^{ij})\right) = \exp\left(-d_s^{n,ij}\right) \tag{5}$$

and hence using the model of Eq. 3 leads to

$$d_s^{n,ij} := -\log(P(n_s^{ij}|\leftrightarrow_s^{ij})) - \log(P(\leftrightarrow_s^{ij}|\leftrightarrow_s,\mathit{\Gamma})) - \log(P(\leftrightarrow_s)). \tag{6}$$

The resulting Wulff shape is a convex approximation to the original probability model. In cases where the assumption that all the half spaces $\mathcal{H}_s^{ij}$ share a boundary with $\mathcal{W}_{\mathcal{H}_s^{ij}}$ does not hold, the cost of unlikely transitions can be underestimated. However, for the most likely directions and hence most relevant directions the approximation will model the true likelihood exactly (c.f. [14]).

In order to use Eq. 6 we also need to approximate the probabilities. $P(n_s^{ij}|\leftrightarrow_s^{ij})$ is estimated from training data, given as a collection of surface meshes, by building a histogram over the training data's normals [14]. The term $P(\leftrightarrow_s^{ij}|\leftrightarrow_s, \mathcal{T}, \Gamma)$ is dependent on the input data and hence changes with the per image classifications Γ and the alignment $\mathcal{T}$. Computing the convex shape as the intersection of the half spaces is computationally demanding (computation of a 3D convex hull on the dual points using point plane duality [25]). Directly inserting the above term would require such a computation whenever the alignment changes. Hence, we want to only do this during the training of the shape prior. To achieve this, we follow the often used approach of weighting the regularization term by the input data.

We fix the structure of the Wulff shape at the training stage by dropping the dependence on the input data. To bring the lost information back to the model we scale the Wulff shape with a weight w_s^{ij}, giving an approximation of Eq. 6:

$$\tilde{d}_s^{n,ij} := w_s^{ij}(\mathcal{T},\Gamma)\big(-\log(P(n_s^{ij}|\leftrightarrow_s^{ij})) - \log(P(\leftrightarrow_s^{ij}|\leftrightarrow_s)) - \log(P(\leftrightarrow_s))\big). \quad (7)$$

This is in analogy to, image segmentation, where often the regularization term is weighted by the input image gradient magnitude.

3.3 Training Data Dependent Parametrization of the Wulff Shapes

A disadvantage of the discretized Wulff shape approach is that a complex Wulff shape composed of the intersection of many half spaces needs to be stored for all the voxels which contained training data (for the other voxels a strong isotropic cost is used). However, often most of the training data normals point in a very similar direction and therefore it is not necessary to store such a complex Wulff shape. To this end, we propose to cluster the input training data and whenever all the training normals lie in up to three clusters we replace them with a surrogate Wulff shape which serves as a faithful approximation (c.f. Fig. 3). For multiple clusters the intersection of multiple surrogate Wulff shapes is used. Using a soft clustering where 95 % of the normals closest to the cluster center with a maximal deviation of 10° are considered, we obtained 74.6 % of voxels with 1 cluster, 10.6 % with 2 clusters and 6.3 % with 3 clusters. Note, that in these cases we do not need to compute a Wulff shape based on half spaces and

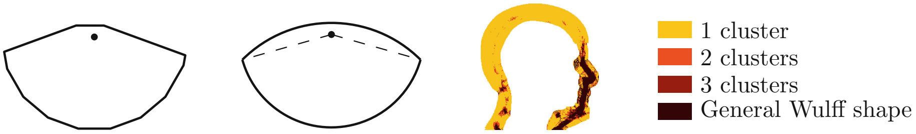


Fig. 3. (Left) 2D illustration of a discretized Wulff shape, where all the training data lies close to a single direction. (Middle) our approximation of the general shape with a surrogate parametric Wulff shape composed out of a spherical sector with an attached spherical cap. (Right) Slice through the volumetric shape prior that indicates the type of Wulff shape used at each place.

hence can directly fit the surrogate Wulff shape into the original training data, which circumvents the discretization of the directions (see supplementary material for more details). Furthermore, when using the prosed clustering approach the memory requirements are reduced by a factor of 3.75 in our implementation.

4 Optimization

The critical part in most algorithms exploiting shape priors is to establish the correspondence between the input data and the shape prior. One of our main contributions is to equip [14] with an automatic alignment procedure. Our optimization strategy alternates between optimizing for the geometry and optimizing for the alignment. The geometry is optimized first, therefore an initialization for the alignment needs to be determined beforehand. We follow the often used strategy of detecting landmark positions, such as points around the eyes and nose. Determining these positions in multiple images allows us to get an estimate of the head pose [7,10]. There is no direct correspondence between the triangulated landmark positions and the implicit volumetric shape prior, as the shape prior is based on many training shapes, and hence the landmark positions end up at slightly different positions in the volume. Our shape prior is trained from shapes that are sampled from a statistical shape model. Therefore, we register the triangulated landmark positions to the ones of the mean shape of the statistical model.

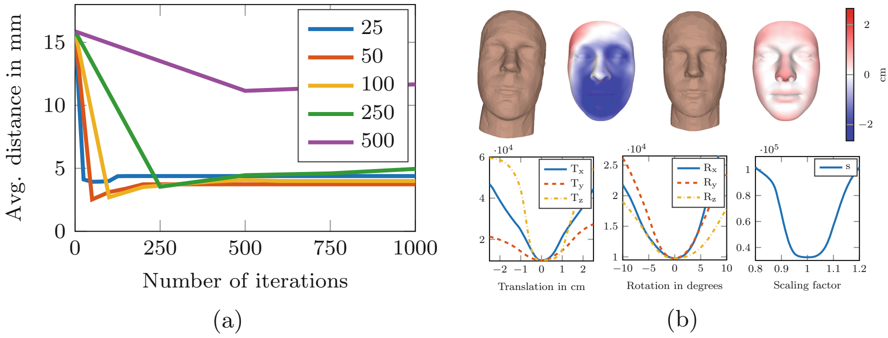
4.1 Optimization with Respect to the Alignment

The energy from Eq. 1 is convex in the variables $\mathbf{x}$, which describe the geometry and labeling, but it is non-convex in the alignment $\mathcal{T}$. It is important to note that for the alignment, only the observed geometry can be used. This means surfaces which are purely filled in by the prior should ideally not be taken into account for the alignment. This can be surfaces which are simply not observable in the input data such as a transition between hair and skin or areas which are filled in by the prior where data is missing. Taking all this into account is important to get a good alignment that can be robustly inferred.

Before we further discuss the optimization we detail the rationale behind the way the alignment transformation is introduced into the formulation. Generally, there are two different ways for defining the alignment, either the input data is at a fixed position and the shape prior gets transformed or the shape prior is at a fixed position and the input data gets transformed. The former one has the disadvantage that the shape prior would not be fixed and hence would need to be adapted for different alignments, by either recomputing or interpolating. Both of these choices add additional computational effort. Therefore, we keep the shape prior at a fixed position and align the input data into the volume of the prior. In this way only the unary cost of the energy and the scaling factors of the data dependent regularization need to be adjusted when the alignment changes. This can be done very efficiently on the GPU in a few seconds by re-evaluating the per voxel data costs using the new alignment transformation. For the alignment with respect to the scaling factor α we need to ensure that a rescaling does not change the energy proportionally to the surface area. Otherwise, the optimization would just try to shrink the object to a reconstruction with 0 surface area and hence no regularization cost. Therefore we normalize the smoothness term with respect to the scaling factor α. In the following derivation we will see that this factor cancels out from the optimization with respect to the alignment.

Given that the convex optimization algorithm which is commonly used to optimize the continuously inspired multi-label assignment problems, the first order primal-dual algorithm [24], essentially executes gradient descent and ascent steps with subsequent proximity operations, it would be tempting to include additional gradient steps in each iteration that account for the alignment. However, this comes with problems and disadvantages. The optimization of the alignment would be an additional update over the volume, we argue that the alignment can be optimized on a surface level and hence more efficiently. Besides the gradient steps that would need to be executed over the volume, a change in alignment also means that the data cost changes due to the dependence on the alignment transformation $\mathcal{T}$ and hence would need to be re-evaluated for the whole volume in each iteration. Additionally, including the alignment update in this straight forward manner would mean that the convergence guarantees that the convex optimization algorithm offers are lost. Therefore, we propose an optimization strategy that addresses these issues by alternating between optimizing for the geometry and aligning the reconstructed surface to the prior.

For the alignment we only take into account the meaningful surfaces, namely the ones which are visible and hence originate from a transition between *free space* and *occupied space*. To avoid bad local minima, we execute the alignment before full convergence and only take into account surfaces which are already present by thresholding the magnitude of the transition gradient x_s^{ij}. We ran an experiment where we optimize for the alignment every 25, 50, 100, 250 and 500 iterations and then measure the distance to the mean shape of the statistical model to evaluate the alignment quality. As shown in Fig. 4 the alignment converges quickly when the alignment is performed often, the longer the interval between the alignments the slower the convergence. If the alignment is performed after many iterations the optimization gets stuck in a bad extremal point. Please note that the geometry at every alignment step is different and therefore the average distance for a better alignment can be higher when more geometry is reconstructed. With these points in mind, we propose to already run the alignment as soon as some geometry is reconstructed and only let the reconstruction converge once the alignment does not change any more. To additionally make the alignment more robust we start the reconstruction with a weak shape prior which only captures the strongest features of the shape and gradually change the prior after each alternation to the desired one for the reconstruction. When directly starting with the final shape prior the experiment given in Fig. 4(a) does not manage to find the right alignment in 3 out of the 5 runs. Taking into account all this leads to an algorithm which robustly finds an accurate alignment



(a)

(b)

Fig. 4. (a) Plot of average distance from mean face for different alignment intervals during the optimization. The optimized model is aligned every 25, 50, 100, 250 and 500 iterations for a total of 1000 iterations. (b) Alignment to the shape prior as described in Sect. 4.1. Top: Visualization of signed alignment error in centimetres. From left to right: face before alignment, error visualization on mean face before alignment, aligned face, error visualization on mean face after alignment. Bottom: Energy function plot of translation, rotation and scale components of seven degree of freedom alignment.

between the input data and the shape prior fully automatically starting from an initial rough estimate of the alignment. Next, we detail our alignment with respect to the surface.

Recall that label 0 denotes free space and labels $i > 0$ occupied space labels (skin, hair, beard, eyebrows and clothing, respectively). The goal is to minimize energy Eq. 1 with respect to the alignment $\mathcal{T}$ but only taking into account visible surfaces, e.g. occupied space ↔ free space transitions. We observe that as soon as we keep the reconstruction fixed, meaning the function that maps given input data to the reconstruction, a change in the alignment transformation $\mathcal{T}$ transforms the input data and hence also the solution for the x_s^i and x_s^{ij} with the same transformation. To make this dependency explicit in the notation we write $\tilde{x}_s^i(\mathcal{T})$ and $\tilde{x}_s^{ij}(\mathcal{T})$, to denote the assignments for the x_s^i and x_s^{ij} that we get for a fixed reconstruction under the alignment transformation $\mathcal{T}$. In terms of energy this means that the unary term is constant under a change of the alignment transformation $\mathcal{T}$ (note that here we ignore the effects of the discretization, which also agrees with the continuous origin of the formulation). The remaining energy for the alignment optimization step reads as

$$E(\mathcal{T}) = \sum_{s \in \Omega, i > 0} \frac{1}{\alpha^2} \phi_s^{0,i}(\mathcal{T}, \tilde{x}_s^{0,i}(\mathcal{T}) - \tilde{x}_s^{i,0}(\mathcal{T})). \tag{8}$$

Besides the dependency of the fixed reconstruction on $\mathcal{T}$ also the smoothness term $\phi_s^{0,i}$ is dependent on $\mathcal{T}$. This is due to the semantic part of the data cost which is included in the smoothness term. For the alignment this is not of big importance as its influence is minimal and it does not add significant complexity to the optimization. In the following we will transform the above energy as an energy over the surface. Besides the smaller complexity this also directly addresses issues with the discretization.

First, we state the relation between the gradient of x_s^i and x_s^{ij} (c.f. [38]):

$$\nabla x_s^i = \sum_j x_s^{ji} - x_s^{ij}. \tag{9}$$

Only taking into account the transitions between occupied space and free space, and ignoring discretization and relaxation, we have $x_s^{ji} = x_s^{ij} = 0$, $\forall j > 0$ and we arrive at $\nabla x_s^i = x_s^{0,i} - x_s^{i,0}$. Considering the original continuous formulation and again ignoring the relaxation, meaning the x_s^i are binary, we can rewrite the integral over the volume as an integral over the surface [9]

$$E(\mathcal{T}) = \int_{\Omega} \frac{1}{\alpha^2} \phi_s^{0,i}(\mathcal{T}, \nabla \tilde{x}_s^i(\mathcal{T})) ds = \int_{\partial \mathcal{F}^i} \frac{1}{\alpha^2} \phi_s^{0,i}(\mathcal{T}, n_s^i(\mathcal{T})) dA, \tag{10}$$

with n_s^i a unit length normal direction on the boundary between free space and label i ($\partial \mathcal{F}^i = \{s : x_s^{0,i} - x_s^{i,0} > 0\}$) at position s. This relation enables us to define the surface regularization in terms of the volume on the left hand side and in terms of an integral over the surface on the right hand side.

Before we explain the alignment over the discrete surface we need to make a remark on how to extract it from the volume. The surface cannot be extracted through thresholding the x_s^i because the entire information about the surface normal direction would get lost. To preserve the surface orientation accurately it is common to extract the surface using marching cubes [21] directly on the non-thresholded x_s^i variables. The output of marching cubes is a triangular mesh representing the surface. We denote the set of all triangles of occupied label i by $\mathbb{T}^i$. The triangle normal and surface area are denoted by $n_t^i(\mathcal{T})$ and $A_t^i(\mathcal{T})$, respectively. The transformation $\mathcal{T}$ also maps the triangle t to a position s in the volume. In the continuous setting this would mean the smoothness term varies at different positions on the triangle. However in practice the smoothness term is only defined on a discrete voxel grid, therefore we use a single constant smoothness term for each triangle which is extracted from the volumetric shape prior by trilinearly interpolating the smoothness cost of the neighboring voxels to the centroid of the triangle. We denote this term by $\phi_t^i(\mathcal{T}, n_t^i(\mathcal{T}))$. Finally, we state the regularization term in its surface formulation over the triangle mesh:

$$E_{\text{mesh}}(\mathcal{T}) = \sum_{i:i>0, t \in \mathbb{T}^i} \phi_t^i(\mathcal{T}, n_t^i(\mathcal{T})) \frac{A_t^i(\mathcal{T})}{\alpha^2} = \sum_{i:i>0, t \in \mathbb{T}^i} \phi_t^i(\mathcal{T}, n_t^i(\mathcal{T})) A_t^i(\mathcal{I}). \tag{11}$$

In the second equation we used that a transformation $\mathcal{T}$ changes the surface area with the square of the scaling factor α. By inserting the identity transformation $\mathcal{I}$, the term α^2 cancels out from the fraction. Leading to the desired property that the alignement part of the energy is independent from the surface area.

For minimizing Eq. 11, we use the gradient descent based, L-BFGS line search approach, implemented in the Ceres solver [1]. In order to start with a weak shape prior which gradually gets stronger, the prior is weakened by increasing $-\log P(\leftrightarrow_s)$

by a constant and scaling the data term. This corresponds to adding non-informative random training data to all voxels.

We present a qualitative and quantitative evaluation of the alignment in Fig. 4(b). The first part shows the signed distance to the mean face before and after refinement of the initial coarse alignment. We recover translation, rotation and scale parameters that lead to a very satisfactory alignment. The mean face used as a reference for the evaluation is close to the location of the best alignment due to the fact that all head models in the shape prior have been aligned to the mean shape. The second part shows plots of our alignment energy. To this end, we took a fixed geometry and plot the energy with respect to the seven dimensions of the similarity transform. We observe that for each of the dimensions the energy has one single local minimum and looks very smooth. It is important to note that we can easily handle translations of 2.5 cm, rotations of 10 degrees in yaw, pitch and roll and scale variations of 20 %. Typical errors of landmark detectors lie well within those bounds [10].

5 Experimental Evaluation

Our input data are images of faces captured using a mobile phone or a compact camera. The typical dataset size is between 15 and 100 images, with a resolution of 640 × 480 pixels. This is depending on whether only frontal images are taken by the person her- or himself or another person is taking pictures all around.

We use two sets of training data. To train the shape prior we use geometric models of heads. This data is derived by randomly sampling 100 human heads from the statistical model of [23]. To train an image based semantic classifier we labeled 80 training images (labels: skin, hair, eyebrows, eyes, beard, clothing and background). We only used the beard label for persons wearing a beard. The eye label is only used to filter the depth maps which are typically unreliable in the eye region (these are often non-rigid during capture, e.g. tracking the camera). We trained a per-pixel semantic classifier using the publicly available code from [19]. The camera poses are estimated using structure-from-motion [32,35] using SIFT features from [33]. The depth maps are computed with the publicly available plane-sweeping stereo matching implementation [13]. We use the landmark detector of [28] and our optimization is implemented in C++.

We present our results in Fig. 5. For more datasets and additional comparisons (patch-based multi view stereo [11] + Poisson surface reconstruction [18]) we refer the reader to the supplementary material. We compare our reconstructions to a state-of-the-art depth map fusion method and a state-of-the-art method for fitting statistical shape models. In the depth map fusion comparison we fuse the depth maps with the TV-Flux fusion from [36], which in our implementation corresponds to regularizing the same unary term that we are using for our multi-label reconstructions with a total variation (TV) prior. In the statistical shape model comparison we fit the model of [23] into our raw input data (depth maps and semantic labels). This leads to a reconstruction of the skin label only. Our proposed approach computes a full semantically annotated reconstruction of the head. Both shape prior formulations manage to overcome the defects

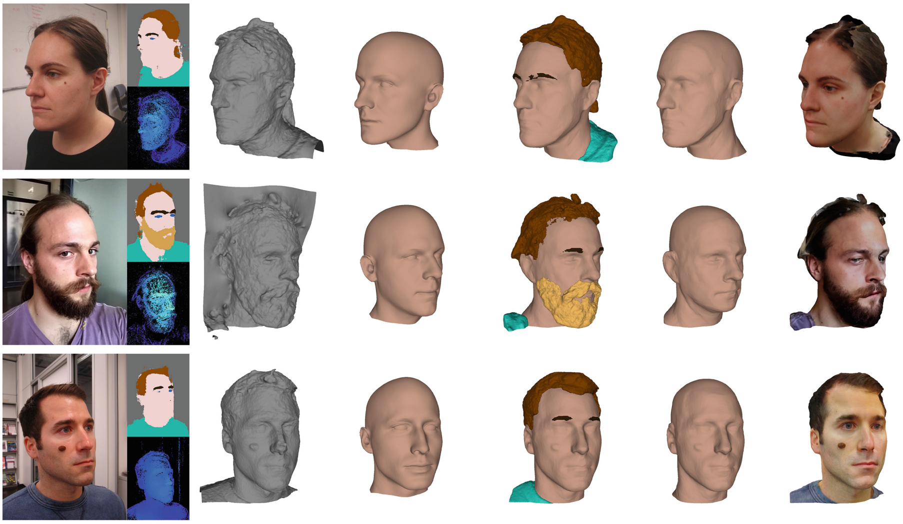


Fig. 5. From left to right: Input image; Input labels and depth; Depth map fusion (TV-Flux fusion from [36]); Statistical model of [23] fitted into our raw input data; Our semantic reconstruction; Our result *skin* class; Our model textured.

in the shapes of the observed geometry. The mole (simulated with a raisin) on the cheek of the person in the last row of Fig. 5 cannot be captured with the low dimensional shape model of [23], therefore it is completely invisible in the respective result. Using our method the mole gets correctly reconstructed even tough such shape details are not represented in the shape prior. One of the key advantages of the implicit shape prior over fitting a low dimensional statistical shape model, is that a deviation from the prior is possible if the data suggests it. In terms of semantic segmentation we are able to fuse the per image semantic classifications, which might be inconsistent in different images, to one single semantic segmentation which is consistent over the whole dataset. Additionally, the semantic segmentation is directly attached to the geometry. In summary, our method is able to reconstruct shape details, at the same time utilizes a strong shape prior for ambiguous input data, recovers hidden surfaces, and extracts one single consistent semantic segmentation for the whole dataset.

6 Conclusion

In this work we introduced a system that fully automatically computes a semantic 3D reconstruction of heads from images. The key novelty of the system is a fully automatic alignment of the shape prior to the input data. Our system reconstructs multiple semantic classes such as skin, hair, beard, clothing, and even handles thin layers of semantic classes such as eyebrows. We demonstrate the applicability of our method to challenging real-world data taken in uncontrolled environments. In future work, we plan to include the capability to handle glasses, potentially using connectivity priors [31]. Further generalizing the alignment to a non rigid transform to a space closer to the mean shape of a statistical shape model might lead to stronger implicit shape priors which are able to hallucinate more complex surfaces than the skin underneath the hair.

Acknowledgment. This project is supported by Grant 16703.1 PFES-ES of CTI Switzerland and the Swiss National Science Foundation under Project Nr. 143422.

References

1. Agarwal, S., Mierle, K., Others: Ceres solver. http://ceres-solver.org
2. Beeler, T., Bickel, B., Beardsley, P., Sumner, B., Gross, M.: High-quality single-shot capture of facial geometry. ACM Trans. Graph. (ToG) **29**(4), 40 (2010)
3. Beeler, T., Bickel, B., Noris, G., Marschner, S., Beardsley, P., Sumner, R.W., Gross, M.: Coupled 3D reconstruction of sparse facial hair and skin. ACM Trans. Graph. (ToG) **31**(4), 117 (2012)
4. Brunton, A., Salazar, A., Bolkart, T., Wuhrer, S.: Review of statistical shape spaces for 3D data with comparative analysis for human faces. Comput. Vis. Image Underst. (CVIU) **128**, 1–17 (2014)
5. Chambolle, A., Cremers, D., Pock, T.: A convex approach to minimal partitions. SIAM J. Imaging Sci. **5**(4), 1113–1158 (2012)
6. Curless, B., Levoy, M.: A volumetric method for building complex models from range images. In: International Conference on Computer graphics and interactive techniques (SIGGRAPH) (1996)
7. Dame, A., Prisacariu, V.A., Ren, C.Y., Reid, I.: Dense reconstruction using 3D object shape priors. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2013)
8. Echevarria, J.I., Bradley, D., Gutierrez, D., Beeler, T.: Capturing and stylizing hair for 3D fabrication. ACM Trans. Graphics (ToG) **33**(4), 125 (2014)
9. Esedoglu, S., Osher, S.J.: Decomposition of images by the anisotropic Rudin-Osher-Fatemi model. Commun. Pure Appl. Math. **57**(12), 1609–1626 (2004)
10. Fanelli, G., Dantone, M., Gall, J., Fossati, A., Van Gool, L.: Random forests for real time 3D face analysis. Int. J. Comput. Vis. (IJCV) **101**(3), 437–458 (2013)
11. Furukawa, Y., Ponce, J.: Accurate, dense, and robust multi-view stereopsis. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) **32**(8), 1362–1376 (2010)
12. Garrido, P., Valgaerts, L., Wu, C., Theobalt, C.: Reconstructing detailed dynamic face geometry from monocular video. ACM Trans. Graph. (ToG) **32**(6), 158 (2013)
13. Häne, C., Heng, L., Lee, G.H., Sizov, A., Pollefeys, M.: Real-time direct dense matching on fisheye images using plane-sweeping stereo. In: International Conference on 3D Vision (3DV) (2014)
14. Häne, C., Savinov, N., Pollefeys, M.: Class specific 3D object shape priors using surface normals. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
15. Häne, C., Zach, C., Cohen, A., Angst, R., Pollefeys, M.: Joint 3D scene reconstruction and class segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2013)
16. Ichim, A.E., Bouaziz, S., Pauly, M.: Dynamic 3D avatar creation from hand-held video input. ACM Trans. Graph. (ToG) **34**(4), 45 (2015)

17. Jourabloo, A., Liu, X.: Pose-invariant 3D face alignment. In: International Conference on Computer Vision (ICCV) (2015)
18. Kazhdan, M., Bolitho, M., Hoppe, H.: Poisson surface reconstruction. In: Eurographics Symposium on Geometry Processing (SGP) (2006)
19. Ladicky, L., Russell, C., Kohli, P., Torr, P.H.: Associative hierarchical random fields. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) **36**(6), 1056–1077 (2014)
20. Lempitsky, V., Boykov, Y.: Global optimization for shape fitting (2007)
21. Lorensen, W.E., Cline, H.E.: Marching cubes: a high resolution 3D surface construction algorithm. In: Conference on Computer Graphics and Interactive Techniques (SIGGRAPH) (1987)
22. Luo, L., Li, H., Paris, S., Weise, T., Pauly, M., Rusinkiewicz, S.: Multi-view hair capture using orientation fields. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
23. Paysan, P., Knothe, R., Amberg, B., Romdhani, S., Vetter, T.: A 3D face model for pose and illumination invariant face recognition (2009)
24. Pock, T., Chambolle, A.: Diagonal preconditioning for first order primal-dual algorithms in convex optimization. In: International Conference on Computer Vision (ICCV) (2011)
25. Preparata, F.P., Shamos, M.: Computational Geometry: An Introduction. Springer, New York (1985)
26. Prisacariu, V.A., Segal, A.V., Reid, I.: Simultaneous monocular 2D segmentation, 3D pose recovery and 3D reconstruction. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) ACCV 2012, Part I. LNCS, vol. 7724, pp. 593–606. Springer, Heidelberg (2013)
27. Ren, S., Cao, X., Wei, Y., Sun, J.: Face alignment at 3000 FPS via regressing local binary features. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
28. Saragih, J.M., Lucey, S., Cohn, J.F.: Deformable model fitting by regularized landmark mean-shift. Int. J. Comput. Vis. (IJCV) **91**(2), 200–215 (2011)
29. Savinov, N., Häne, C., Ladicky, L., Pollefeys, M.: Semantic 3D reconstruction with continuous regularization and ray potentials using a visibility consistency constraint. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
30. Savinov, N., Ladicky, L., Häne, C., Pollefeys, M.: Discrete optimization of ray potentials for semantic 3D reconstruction. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
31. Stühmer, J., Schröder, P., Cremers, D.: Tree shape priors with connectivity constraints using convex relaxation on general graphs. In: International Conference on Computer Vision Proceedings (ICCV) (2013)
32. Tanskanen, P., Kolev, K., Meier, L., Camposeco, F., Saurer, O., Pollefeys, M.: Live metric 3D reconstruction on mobile phones. In: International Conference on Computer Vision (ICCV) (2013)
33. Vedaldi, A., Fulkerson, B.: VLFeat: an open and portable library of computer vision algorithms (2008). http://www.vlfeat.org/
34. Wei, Y., Ofek, E., Quan, L., Shum, H.Y.: Modeling hair from multiple views. ACM Trans. Graphics (TOG) **24**(3), 816–820 (2005)
35. Wu, C.: VisualSFM: a visual structure from motion system (2011). http://ccwu.me/vsfm/
36. Zach, C.: Fast and high quality fusion of depth maps. In: International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT) (2008)

37. Zach, C., Häne, C., Pollefeys, M.: What is optimized in tight convex relaxations for multi-label problems? In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
38. Zach, C., Häne, C., Pollefeys, M.: What is optimized in convex relaxations for multilabel problems: Connecting discrete and continuously inspired map inference. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) **36**(1), 157–170 (2014)