Tensor-Based Non-Rigid Structure from Motion

Stella Graßhof and Sami Sebastian Brandt
IT University of Copenhagen, Denmark
stgr@itu.dk, sambr@itu.dk

Abstract

In this work we present a method that combines tensor-based face modelling and analysis and non-rigid structure-from-motion (NRSFM). The core idea is to see that the conventional tensor formulation for the face structure and expression analysis can be utilised while the structure component can be directly analysed as the non-rigid structure-from-motion problem. To the NRSFM problem part we further present a novel prior-free approach that factorises the 2D input shapes into affine projection matrices, rank-one 3D affine basis shapes, and the basis shape coefficients. The linear combination of the basis shapes thus yields the recovered 3D shapes upto an affine transformation. In contrast to most works in literature, no calibration information of the cameras or structure prior is required. Experiments on challenging face datasets show that our method, with and without the metric upgrade, is accurate and fast when compared to the state-of-the-art and is well suitable for dense reconstruction and face editing.

1. Introduction

Non-rigid structure-from-motion (NRSFM), the problem of reconstructing both the scene geometry and dynamic structure, is a classic problem in computer vision. NRSFM in general is a difficult problem, although there have been significant developments in the last two decades. The starting point for NRSFM can be seen as the work [11] proposing a low-rank approach with the assumption that the deformable 3D shape is a linear combination of rigid 3D basis shapes. This led to a matrix factorisation problem generalising [34]. The classic NRSFM problem has the characteristic that the decomposed motion matrix has a block-form structure. A general solution also needs to tackle the inherent geometric and structural ambiguities that have been a challenging problem to date.

There have been numerous approaches to the NRSFM problem. The majority have assumed a calibrated affine camera and utilised the well-known orthogonality constraints. Additional constraints include heuristic deformation minimisation [8], constraints arising from stereo rig [19], shape basis fixation [41], and factoring a multifocal tensor [26]. Physical and temporal priors have also been widely used such as rigidity [4, 13], camera trajectory smoothness [22], temporal smoothness [2, 35], and deformation [8, 14]. The problem has alternatively been viewed as manifold learning [14] that has naturally led to optimisation by alternation [31, 35, 36]. In addition, a coarse-to-fine solution was proposed in [4] that uses information on several scales. There have also been uncalibrated approaches [9, 10] that assume statistical independence of the shape bases to solve the structural and geometric ambiguities. In [17], the structural ambiguity was ignored by using the observation that the reconstruction is not ambiguous unlike the shape basis.

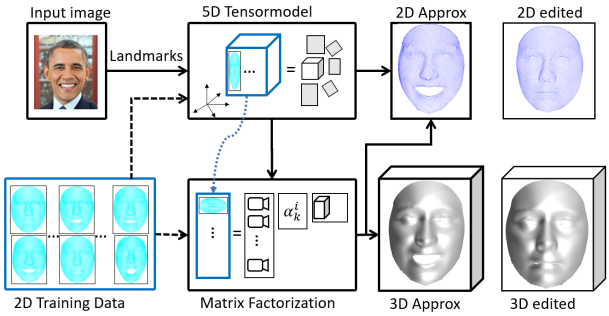


Figure 1: Overview for 3D reconstruction and editing. The 2D training data is factorised by a 5D tensor-based, and a matrix-based approach. The tensor model yields dense 2D shapes from 2D landmarks, and editing the parameters gives an adapted 2D shape. The 3D shapes are obtained by a factorisation defined by the matrix-based model.

Most recently, [29] posed the NRSFM problem as a multi-layer block sparse dictionary learning problem which was converted into a form of a deep neural network. In [33], a dense auto-decoder-based deformation model with Fourier domain constraints was trained on dense 2D point tracks. In [30], a union of local linear subspaces approach was proposed that summarises the behaviour of local measurement by points on the Grassmannian manifold while

the 3D shape was represented using the low-rank constraint. Specifically 3D faces have been modelled by factorisation approaches [5, 20]. As a generalisation of the matrix-based SVD the Higher-Order SVD (HOSVD) was introduced in [18] which yields subspaces directly related to the data dimensions. The HOSVD has since been proven to be a useful tool to model and analyse faces [37, 38, 16, 24, 25, 23]. However, in most works the HOSVD is employed on 3D faces [12, 6, 7, 1, 24, 25, 23]. Additionally, there have been attempts in the shape-from-shading community to estimate a tensor structure from unstructured data, i.e., if no labels for the tensor dimensions are available [39, 40].

Even though the neural networks (NNs) have received a lot of attention in the community their application in NRSFM is not entirely problem-free. They tend to overfit and they do not provide a direct control of the model complexity. Additionally, with the exception of [29], previous methods on 3D reconstruction of faces and human body shapes require strong 3D supervision, and are unable to interpolate or create new shapes with varying expressions.

This work hence has two major objectives. First, by applying a tensor model for faces, we aim at utilising the structure of the database where shape, data dimensionality, viewing angle, identity, emotion naturally form the modes of the data tensor. The approach [23] therefore provides a straightforward way to parameterise and edit faces. We show that their results, obtained with 3D data, can also be derived from 2D projection data. Second, after realising that the face projection data naturally yields the non-rigid structure from motion problem in one of the matrix unfoldings of the data tensor, we aim at simultaneously solving it as part of the tensor model. Moreover, we reformulate the non-rigid, low-rank model and, instead of trying to solve the harder problem of finding the underlying rank-three 3D shape basis, we individually analyse the singular vectors that form the shape matrix and back-project them onto 3D to create rank-one shapes.[1] An overview of our approach is illustrated in Fig. 1 for 3D reconstruction and shape editing.

Our contributions are summarised as follows.

- We combine (1) tensor-based modelling of faces and expressions and the (2) non-rigid structure-from-motion into one problem. The key observation is that a matrix unfolding of the tensor yields the standard measurement matrix of the NRSFM problem.
- We propose a novel NRSFM method that is simpler, more accurate, and computationally more efficient than previous methods. Our method is faster than all the other factorisation approaches, and computationally light compared to NN approaches, because it does not require a large database and it has few parameters. It is also well suited for dense data.
- We apply the well known *stratified approach* for the NRSFM problem, i.e., we use uncalibrated, affine cameras. The advantage is that we avoid the cumbersome orthogonal constraints in the non-rigid factorisation step that makes our method simpler and more general. The metric upgrade for the reconstruction can directly achieved by using camera calibration information or the standard autocalibration methods.
- We suggest using a rank-one shape basis by back-projecting each singular vector of the factorisation model onto the 3D space. We hence avoid the problem of grouping the singular vectors and do not need to explicitly enforce the block-structure of the motion matrix. We provide an option to retrieve the basis shapes so that they become as independent as possible.
- We retrieve the same subspace structure for 2D faces, which was found for 3D faces in [24, 23].
- We propose a *generative* model to explicitly parameterise and edit 3D shapes using the semantically meaningful subspaces of the 2D canonical tensor model.

[1] Note that this is different from [17] where full-rank basis shapes were represented by $3N$ vectors—we instead assume that the basis shapes are degenerate in the meaning that each of them is represented by a $3 \times N$ rank-one matrix, i.e., the matrix has three linearly dependent rows.

2. Tensor Representation

Let $\mathcal{X} \in \mathbb{R}^{N \times D \times F \times P \times E}$, $D = 2$, be a data tensor of 2D faces, where N is the number of corresponding points, F number of 2D projections of each 3D face with the fixed expression and identity, P number of persons, and E number of expressions. It is assumed that all the faces in $\mathcal{X}$ have been centered so that the each 2D face has the mean coordinate in the origin. The tensor $\mathcal{X}$ is then decomposed by the Higher-Order-SVD (HOSVD) [18] as

$$\mathcal{X} \approx \widehat{\mathcal{X}} = \mathcal{S} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)} \times_3 \mathbf{U}^{(3)} \times_4 \mathbf{U}^{(4)} \times_5 \mathbf{U}^{(5)}, \tag{1}$$

where $\times_d$ is the d-way product, $\mathcal{S} \in \mathbb{R}^{\widetilde{N} \times D \times \widetilde{F} \times \widetilde{P} \times \widetilde{E}}$ is the core tensor, and $\mathbf{U}^{(d)} \in \mathbb{R}^{d \times \widetilde{d}}$, $d = 1, 2, \ldots, 5$; are semi-orthogonal matrices which consist of the singular vectors corresponding to the d-mode unfolded tensor, with $\widetilde{d} \leq d$ representing the number of retained elements of the dimension d, i.e., the smallest singular values and vectors have been truncated. The tensor $\mathcal{X} = \mathcal{X}_0 + \Delta\mathcal{X}$ is divided into rigid $\mathcal{X}_0$ and non-rigid $\Delta\mathcal{X}$ parts by separating the three largest mode-1 singular values and vectors from the remaining ones, respectively. The rigid part is thus obtained as

$$\mathcal{X}_0 \approx \widehat{\mathcal{X}}_0 = \mathcal{S}_0 \times_1 \mathbf{U}_0^{(1)} \times_2 \mathbf{U}^{(2)} \times_3 \mathbf{U}^{(3)} \times_4 \mathbf{U}^{(4)} \times_5 \mathbf{U}^{(5)}, \tag{2}$$

where $\mathcal{S}_0 \in 3 \times D \times \widetilde{F} \times \widetilde{P} \times \widetilde{E}$ is the core tensor, and the $N \times 3$ semi-orthogonal matrix $\mathbf{U}_0^{(1)}$ contain the first three 1-mode singular values and vectors. The non-rigid part $\Delta\mathcal{X}$

is formed as

$$\Delta\mathcal{X} \approx \Delta\widehat{\mathcal{X}} = \mathcal{S}_1 \times_1 \mathbf{U}_1^{(1)} \times_2 \mathbf{U}^{(2)} \times_3 \mathbf{U}^{(3)} \times_4 \mathbf{U}^{(4)} \times_5 \mathbf{U}^{(5)}, \tag{3}$$

where $\mathcal{S}_1 \in \mathbb{R}^{(\widetilde{N}-3)\times D\times \widetilde{F}\times \widetilde{P}\times \widetilde{E}}$ is the core tensor, and $\mathbf{U}_1^{(1)} \in \mathbb{R}^{N\times(\widetilde{N}-3)}$.

3. Rank-One Basis Shapes

3.1. Rigid Component

The rationale of dividing the tensor $\mathcal{X}$ into the rigid and non-rigid part comes from the interpretation of the 1-mode matrix unfolding $\mathbf{X}^{(1)}$ of the tensor which is the transposed measurement matrix of the classic Tomasi–Kandade [34] factorisation method. Hence, the first three left singular vectors of $\mathbf{X}^{(1)}$ represent the least squares estimate for the affine 3D structure $\mathbf{B}_0$, obtained by the singular value decomposition as

$$\widehat{\mathbf{X}}_0^{(1)^{\mathrm{T}}} = \underbrace{\left(\frac{1}{\sqrt{N}}\mathbf{V}_0^{(1)}\mathbf{\Sigma}_0^{(1)}\right)}_{\mathbf{M}_0}\underbrace{\left(\sqrt{N}\mathbf{U}_0^{(1)^{\mathrm{T}}}\right)}_{\mathbf{B}_0} = \mathbf{M}_0\mathbf{B}_0, \tag{4}$$

where the $2I \times 3$ matrix $\mathbf{M}_0 = (\mathbf{M}^1, \mathbf{M}^2, \ldots, \mathbf{M}^I)$ contains estimates for all the 2×3 inhomogeneous affine projection matrices for all the $I = FPE$ views. In other words, the non-rigid variation in the decomposition (1) is constructed centred at the mean rigid 3D shape, that is, the 3D point, corresponding to a non-rigid object and projected to the image i, is $\mathbf{z}_n^i = \mathbf{b}_{0n} + \Delta\mathbf{z}_n^i$, where $\mathbf{b}_{0n}$ is the rigid, mean shape and $\Delta\mathbf{z}_n^i$ is the non-rigid component. The other modal components in the rigid approximation $\mathcal{X}_0$, apart from the mode-1, constitute the variations in the affine projection matrix e.g. the face widening due to smiling.

3.2. Non-Rigid Component

In contrast to the standard factorisation model which is based on the assumption that a non-rigid shape is a linear combination of 3-dimensional basis shapes, we additionally assume that the non-rigid basis shapes are 3-dimensional *rank-one* shapes—*not* $3N$-vectors as e.g. in Dai et al. [17]. In effect, the non-rigid components of the 3D shapes are represented as $\Delta\mathbf{z}_n^i = \sum_{k=1}^{\widetilde{N}-3} \alpha_k^i \mathbf{b}_{kn}$, where α_k^i is a scalar and $\mathrm{rank}(\mathbf{B}_k) = 1$ for $k \neq 0$, where $\mathbf{B}_k = (\mathbf{b}_{k1}\ \mathbf{b}_{k2}\ \cdots \mathbf{b}_{kN}) \in \mathbb{R}^{3\times N}$. Assuming that all the structure components share the same projection matrix on to a fixed image, the non-rigid 3D component $\Delta\mathbf{z}_n^i$ maps to the non-rigid 2D parts $\Delta\hat{\mathbf{x}}_n^i$ stored in $\Delta\mathcal{X}$, where

$$\Delta\hat{\mathbf{x}}_n^i = \mathbf{M}^i\Delta\mathbf{z}_n^i = \mathbf{M}^i\left(\sum_{k=1}^{\widetilde{N}-3}\alpha_k^i\mathbf{b}_{kn}\right). \tag{5}$$

The mode-1 unfolding of the tensor $\Delta\mathcal{X}$ thus factorises into the a weighted sum of $\widetilde{N}-3$ 3D rank-one basis shapes $\mathbf{B}_k$, with $\alpha_k^i \in \mathbb{R}$, and I projection matrices $\mathbf{M}^i \in \mathbb{R}^{2\times 3}$, as

$$\Delta\mathbf{X}^{(1)^{\mathrm{T}}} \approx \underbrace{\begin{pmatrix} \alpha_1^1\mathbf{M}^1 & \alpha_2^1\mathbf{M}^1 & \cdots & \alpha_{\widetilde{N}-3}^1\mathbf{M}^1 \\ \alpha_1^2\mathbf{M}^2 & \alpha_2^2\mathbf{M}^2 & \cdots & \alpha_{\widetilde{N}-3}^2\mathbf{M}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_1^I\mathbf{M}^I & \alpha_2^I\mathbf{M}^I & \cdots & \alpha_{\widetilde{N}-3}^I\mathbf{M}^I \end{pmatrix}}_{=\breve{\mathbf{M}}\in\mathbb{R}^{2I\times 3\widetilde{N}-9}} \underbrace{\begin{pmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \\ \vdots \\ \mathbf{B}_{\widetilde{N}-3} \end{pmatrix}}_{=\breve{\mathbf{B}}\in\mathbb{R}^{(3\widetilde{N}-9)\times N}}. \tag{6}$$

Clearly, the basis matrix $\breve{\mathbf{B}}$, and hence $\Delta\mathbf{X}^{(1)^{\mathrm{T}}}$, has the rank $\tilde{N}-3$. Additionally, the motion matrix $\breve{\mathbf{M}}$ has the block structure shown above.

Since the SVD yields the closest approximation in the least squares sense under the rank constraint, the auxiliary estimates for the structure and motion matrix are obtained as

$$\Delta\widehat{\mathbf{X}}^{(1)^{\mathrm{T}}} = \underbrace{\left(\frac{1}{\sqrt{N}}\mathbf{V}^{(1)}\mathbf{\Sigma}^{(1)}\right)}_{=\mathbf{M}}\underbrace{\left(\sqrt{N}\mathbf{U}^{(1)^{\mathrm{T}}}\right)}_{=\mathbf{B}} = \mathbf{M}\mathbf{B}, \tag{7}$$

The remaining problem is to find the $(3\tilde{N}-9)\times(\tilde{N}-3)$ operator $\mathbf{A}$ so that

$$\mathbf{M}\mathbf{B} = \mathbf{M}\mathbf{A}^{\dagger}\mathbf{A}\mathbf{B} = \widehat{\mathbf{M}}\widehat{\mathbf{B}}, \tag{8}$$

where $\widehat{\mathbf{B}} = \mathbf{A}\mathbf{B}$ and $\widehat{\mathbf{M}} = \mathbf{M}\mathbf{A}^{\dagger}$, and $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{B}}$ are the estimates matching the form (6). In the following two subsections, we discuss two different approaches for finding $\mathbf{A}$.

3.3. Principal Component Analysis (PCA)

Let $\mathbf{C}^{(1)} = \mathbf{C}_0^{(1)} + \Delta\mathbf{C}^{(1)}$ denote the mode-1 *covariance* equivalent to

$$\mathbf{C}^{(1)} = \mathbf{X}^{(1)}\mathbf{X}^{(1)^{\mathrm{T}}} = \mathbf{U}^{(1)}{\mathbf{\Sigma}^{(1)}}^2\mathbf{U}^{(1)^{\mathrm{T}}}. \tag{9}$$

where $\mathbf{C}_0^{(1)} = \mathbf{U}_0^{(1)}{\mathbf{\Sigma}_0^{(1)}}^2\mathbf{U}_0^{(1)^{\mathrm{T}}}$ and $\Delta\mathbf{C}^{(1)} = \mathbf{U}_1^{(1)}{\mathbf{\Sigma}_1^{(1)}}^2\mathbf{U}_1^{(1)^{\mathrm{T}}}$. We see that the mode-1 left singular vectors, or the rows in $\mathbf{B}_0$ and $\mathbf{B}$, are the *principal components* of the columns in the mode-1 matrix unfolding of the data tensor while the squared mode-1 singular values squared are the corresponding variances. In other words, $\Delta\mathbf{C}^{(1)}$ contain all the principal components, except the first three since the mean rigid shape has been factored to $\mathbf{C}_0^{(1)}$. Hence, $\mathbf{B}$ has $\tilde{N}-3$ linearly independent rows that form the basis to the non-rigid structure. Each of them are mapped to the three-dimensional space by back-projection to form rank-one basis shape $\hat{\mathbf{B}}_k = \mathbf{d}_k\mathbf{b}_k^{\mathrm{T}}$, where $\mathbf{b}_k^{\mathrm{T}}$ denotes the row k of $\mathbf{B}$, and $\mathbf{d}_k$, is the 3×1 unit vector in which the component k is back-projected into the 3D space. The direction

$\mathbf{d}_k = \mathbf{R}_k\mathbf{e}_1$ results from the 3D rotation $\mathbf{R}_k$ that maps the one-dimensional basis $\mathbf{b}_k^{\mathrm{T}}$, first back-projected on the x-axis direction, to the rigid 3D shape. $\mathbf{A}$ in (8) is hence equivalent to the form $\mathbf{A} = \mathbf{D}$, where $\mathbf{D}$ is the $(3\tilde{N} - 9) \times (\tilde{N} - 3)$ block diagonal matrix with 3×1 blocks $\mathbf{d}_k$.

3.4. Independent Component Analysis (ICA)

More generally, the operator $\mathbf{A}$ can be written in the form $\mathbf{A} = \mathbf{DG}$, where $\mathbf{G}$ is a $(\tilde{N}-3)\times(\tilde{N}-3)$ orthogonal matrix. Although no grouping of the rank-one components is strictly necessary, we also consider estimating the rank-one shapes by setting $\mathbf{G}$ so that the components are as *statistically independent factors* as possible. This will allow us to analyse statistically linked shape components, such as lip movements, by isolating them from the other deformations. We will do this by Independent Component Analysis (ICA).

ICA is a method for blind source separation that intends to decompose the underlying signals into statistically independent factors by using higher order statistics of multidimensional observations characterised by the random vector $\mathbf{Z}$, and can be defined by minimising the mutual information

$$I(\mathbf{Z}) = \sum_j H(Z_j) - H(\mathbf{Z}), \tag{10}$$

where H refers to differential entropy and $\mathbf{Y} = \mathbf{A}_{\mathrm{ICA}}\mathbf{Z}$ to a random vector corresponding to the columns in $\mathbf{B}$. If the vectors are mean centred and white, it implies that the mixing matrix $\mathbf{A}_{\mathrm{ICA}} = \mathbf{G}^{\mathrm{T}}$ will be an orthogonal matrix, hence,

$$\mathbf{B}_{\mathrm{ICA}} \equiv \mathbf{A}_{\mathrm{ICA}}^{\mathrm{T}}\mathbf{B} = \mathbf{GB}, \tag{11}$$

where the rows of $\mathbf{B}_{\mathrm{ICA}}$ will be in as statistically independent as possible. Here, we compute the orthogonal, separation matrix $\mathbf{G}$ as described in [27].

3.5. Recovery of Rank-One Basis Shapes

Let $\mathbf{b}_k$ denote the row k in $\mathbf{B}_{\mathrm{PCA}} \equiv \mathbf{B}$ or $\mathbf{B}_{\mathrm{ICA}}$, depending whether the PCA or ICA model is selected, respectively. We are searching for the minimiser to the energy functional

$$E(\mathbf{d}, \alpha) = \sum_i \|\Delta\mathbf{X}^{(1)^i} - \sum_k \alpha_k^i \mathbf{B}_k^i\|_{\mathrm{Fro}}^2, \tag{12}$$

subject to $\|\mathbf{d}_k\|_2 = 1$, for all k, where $\mathbf{B}_k^i = \mathbf{M}^i\mathbf{d}_k\mathbf{b}_k^{\mathrm{T}}$ are the rank-one operators referring to the rank-one basis shapes, and α_k^i are the corresponding basis coefficients that can be computed by orthogonally projecting the differential measurement matrix blocks $\Delta\mathbf{X}^{(1)^i}$ onto the rank-one operators. We first note a useful property, stated as follows.

Lemma 3.1 *$\mathbf{B}_k^i \perp \mathbf{B}_{k'}^{i'}$ in the operator inner product, $k \neq k'$, for all i, i'.*

Algorithm 1 Non-rigid Structure From Motion by rank-one Basis Shapes

1. Form the translation corrected data tensor $\mathcal{X}$, as in (6). Initialise the parameters $\mathbf{d}$, α.
2. Decompose $\mathcal{X}$ into the rigid $\mathcal{X}_0$ and non-rigid $\Delta\mathcal{X}$ part as in (2) and (3), respectively.
3. Factorise the non-rigid part as $\Delta\mathbf{X}^{(1)^{\mathrm{T}}} \approx \mathbf{MB}$, where $\mathbf{M} = \frac{1}{\sqrt{N}}\mathbf{V}^{(1)}\mathbf{\Sigma}^{(1)}$ and $\mathbf{B} = \sqrt{N}\mathbf{U}^{(1)^{\mathrm{T}}}$.
4. Do either
   (a) Compute the PCA basis by assuming $\mathbf{G} = \mathbf{I}$ and so that $\mathbf{B}_{\mathrm{PCA}} = \mathbf{B}$; or
   (b) Find the orthogonal transformation $\mathbf{G}$ and ICA basis by FastICA [27] so that $\mathbf{B}_{\mathrm{ICA}} = \mathbf{GB}$.
5. Update the component affine back-projections $\mathbf{d}_k$, $k = 1, 2, \ldots, K$, by minimising (12) over $\mathbf{d}$ subject to $\|\mathbf{d}_k\|_2 = 1$, for all k.
6. Form the rank-one basis shapes $\mathbf{B}_k^i = \mathbf{M}_0^i\mathbf{d}_k\mathbf{b}_k^{\mathrm{T}}$, $i = 1, 2, \ldots, I$, where $\mathbf{b}_k^{\mathrm{T}}$ is the kth row of $\mathbf{B}_{\mathrm{PCA}}$ or $\mathbf{B}_{\mathrm{ICA}}$, $k = 1, 2, \ldots, K$.
7. Update the basis coefficients by orthogonal projection $\alpha_k^i = \langle\Delta\mathbf{X}^{(1)^i}, \mathbf{B}_k^i\rangle / \langle\mathbf{B}_k^i, \mathbf{B}_k^i\rangle$, $i = 1, 2, \ldots, I$,
8. Iterate from Step 5 until convergence.

Proof. We may write

$$\begin{aligned}\langle\mathbf{B}_k^i, \mathbf{B}_{k'}^{i'}\rangle &= \langle\mathrm{vec}\{\mathbf{B}_k^i\}, \mathrm{vec}\{\mathbf{B}_{k'}^{i'}\}\rangle \\ &= \langle\mathbf{M}^i\mathbf{d}_k, \mathbf{M}^{i'}\mathbf{d}_{k'}\rangle\langle\mathbf{b}_k, \mathbf{b}_{k'}\rangle,\end{aligned} \tag{13}$$

which vanishes for $k \neq k$' since $\mathbf{b}_k \perp \mathbf{b}_{k'}$. □

Now, we are ready to show how we minimise (12). The method is given in Algorithm 1.

4. Canonical Tensor Model

In [24], a tensor model of 3D face shapes based on the factorisation of a 3D data tensor was presented. We adopt this approach and apply the HOSVD to a 5D data tensor containing 2D shapes. Inspired by [24], a 2D face shape $\widehat{\mathbf{f}} \in \mathbb{R}^{N\times D}$ can be represented using (1) as

$$\begin{aligned}\widehat{\mathbf{f}} = \mathcal{S} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)} \\ \times_3 \mathbf{p}_3^{\mathrm{T}}\mathbf{U}^{(3)} \times_4 \mathbf{p}_4^{\mathrm{T}}\mathbf{U}^{(4)} \times_5 \mathbf{p}_5^{\mathrm{T}}\mathbf{U}^{(5)},\end{aligned} \tag{14}$$

where $\mathbf{p}_k$, $k = 3, 4, 5$ represent the canonical parameter vectors, whose lengths correspond to their respective subspace $\mathbf{U}^{(k)}$. In analogue to [24], an unknown shape $\mathbf{f}$ can be approximated by $\widehat{\mathbf{f}}$ by minimising

$$\min_{\mathbf{p}_k} \|\widehat{\mathbf{f}} - \mathbf{f}\|_F^2 + \sum_{k=3}^{5} \lambda_k\|\mathbf{p}_k\|_2^2 + \lambda_{k,s}\left(\mathbf{p}_k^{\mathrm{T}}\mathbf{1} - 1\right)^2, \tag{15}$$

where $\lambda_k, \lambda_{k,s} \in \mathbb{R}^+$ are weights which must be manually set. This minimisation problem can be conveniently solved in an alternating scheme, see [24, 23].

In analogue to (2) and (3), one face shape $\widehat{\mathbf{f}}$ can be represented as the sum $\widehat{\mathbf{f}} = \widehat{\mathbf{f}}_0 + \Delta\widehat{\mathbf{f}}$, where the rigid part

$$\widehat{\mathbf{f}}_0 = \mathcal{S}_0 \times_1 \mathbf{U}_0^{(1)} \times_2 \mathbf{U}^{(2)} \times_3 \mathbf{u}_3^{\mathrm{T}} \times_4 \mathbf{u}_4^{\mathrm{T}} \times_5 \mathbf{u}_5^{\mathrm{T}}, \quad (16)$$

and the nonrigid part

$$\Delta\widehat{\mathbf{f}} = \mathcal{S}_1 \times_1 \mathbf{U}_1^{(1)} \times_2 \mathbf{U}^{(2)} \times_3 \mathbf{u}_3^{\mathrm{T}} \times_4 \mathbf{u}_4^{\mathrm{T}} \times_5 \mathbf{u}_5^{\mathrm{T}}, \quad (17)$$

and $\mathbf{u}_k^{\mathrm{T}} = \mathbf{p}_k^{\mathrm{T}} \mathbf{U}^{(k)}$, $k = 3, 4, 5$.

5. 3D Shape Synthesis and Editing

So far we have presented two factorisation approaches: a matrix-based (Sec. 3), and a tensor-based (Sec. 4). The two major differences between them are that (1) only the matrix-based approach provides 3D estimates for each 2D shape, as a weighted sum of 3D basis shapes, and (2) only the canonical tensor model-based approach offers intuitive synthesis of new 2D shapes by semantically meaningful parameters related to the subspaces. Here, we combine both approaches to enable synthesis of new 3D shapes from 2D by parameter editing, e.g. to change the expression.

First, we factorise the 2D data by both approaches. Second, we use the tensor model (14) to create a new 2D shape by either: (1) choosing the parameter vectors $\mathbf{p}_k$ freely, or (2) estimate them to approximate an arbitrary 2D shape $\mathbf{x}'$ by solving (15), after global alignment, and then change the parameter vectors to a desired identity or expression, yielding a transfer of person or expression, respectively. In both cases the tensor model provides a 2D shape $\widehat{\mathbf{x}}'$.

Third, we employ the matrix-factorisation to retrieve the corresponding 3D estimate $\widehat{\mathbf{z}}'$ as follows. For each 2D training sample $\mathbf{x}^i$ its 3D estimate $\mathbf{z}^i$ is the weighted sum of 3D rank-one basis shapes

$$\widehat{\mathbf{z}}^i = \mathbf{B}_0 + \sum_k \alpha_k^i \mathbf{d}_k \mathbf{b}_k^{\mathrm{T}}. \quad (18)$$

The basis coefficients α_k^i are computed by orthogonal projection in step 7 of Alg. 1, and employs the estimated projection matrix $\mathbf{M}_0^i$ of the sample i, which is unknown for a new shape $\widehat{\mathbf{x}}'$, but can be estimated by the affine camera resection algorithm as $\mathbf{M}_0'$. The basis coefficients α_k' are then estimated that yield the 3D estimate $\widehat{\mathbf{z}}'$ corresponding to the new shape $\widehat{\mathbf{x}}'$ as $\widehat{\mathbf{z}}' = \mathbf{B}_0 + \sum_k \alpha_k' \mathbf{d}_k \mathbf{b}_k^{\mathrm{T}}$.

We use the proposed approach to synthesise the six prototypical emotions for the mean person and rotation in 2D, shown in Fig. 4(a)-(g), and their 3D estimates shown in Fig. 4(h)-(n), thereby retrieve dense 3D shapes from sparse 2D points, as shown in Fig. 1.

6. Databases

6.1. LS3D-W Balanced

The Large Scale 3D Faces in-the-Wild dataset (LS3D-W) [15] is a facial landmark dataset, which contains ca. 230,000 images, each annotated with 68 2D points. [15] also defines the *LS3D-W Balanced*, as a subset of the LS3D-W, a total of 7200 images, and includes a *balanced* number of varying yaw angles. The faces vary in expression and are in random orientation, and order, hence no temporal information or underlying substructure is provided.

6.2. Binghamton 3D Facial Expression Database

The Binghamton 3D Facial Expression Database (BU3DFE) [42] contains 2500 3D face scans, and corresponding images. 100 persons (56% female, 44% male) in 25 facial expressions: neutral, or one of the six basic emotions (anger, disgust, fear, happiness, sadness, and surprise) in four increasing expression intensity levels. For each face scan 83 3D facial landmarks are provided, and we added the nose tip and top of forehead, resulting in 85 points. These were used to estimate 7308 dense point correspondences between the dense scans by an adapted version of [21]. Additionally, [3] yields 68 2D landmarks for each frontal face image. Hence, we obtain the following three 2D datasets:

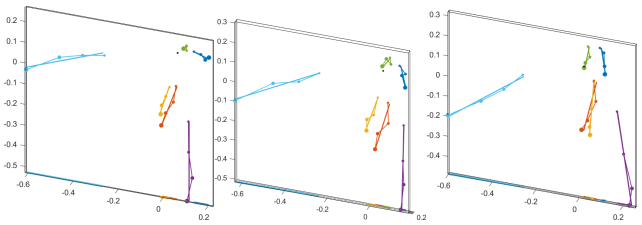
- BU3DFE-68: the 68 2D landmarks retrieved by [3].
- BU3DFE-85: the 85 sparse 3D landmarks rotated by 3 yaw angles $\alpha_y \in \left\{-\frac{\pi}{8}, 0, \frac{\pi}{8}\right\}$, projected to 2D.
- BU3DFE-7k: the estimated 7308 3D points rotated by 3 yaw angles $\alpha_y \in \left\{-\frac{\pi}{8}, 0, \frac{\pi}{8}\right\}$, projected to 2D.

7. Experiments

We compare the two proposed variants of our approach to Dai et al.'s pseudoinverse (PI) [17], and Block matrix Method (BMM) [17], and Kong and Lucey's Priorless decomposition (K&L) [28], and Brandt et al.'s ISA [10]. For the factorisation we selected $\tilde{N} = 15$ components for all experiments and used the equivalent truncation point for all the methods so that the results are directly comparable.

7.1. Expression Space

The multilinear tensor model of 3D faces in [24], based on the BU3DFE database [42], revealed a planar star-shaped substructure in the expression subspace. and a similar structure on the basis of 2D database was found in [23]. To complement these findings, we investigated 2D data based on the BU3DFE dataset [42], described in Sec. 6.2. The resulting expression space $\mathbf{U}^{(5)}$ from (1) reveals the same substructure for all of the three datasets, see Fig. 2.



(a) BU3DFE-68 (b) BU3DFE-85 (c) BU3DFE-7k

Figure 2: Visualisation of the first three singular vectors of the expression space $\mathbf{U}^{(5)}$ from (1) resulting from the factorisation of the datasets (a) BU3DFE-68, (b) BU3DFE-85, and (c) BU3DFE-7k; c.f. [23].

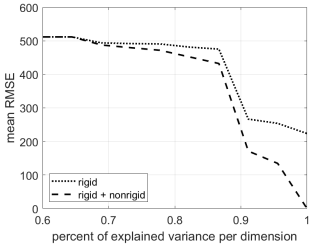


Figure 3: Approximation error for varying percentage of explained variance (PoV).

7.2. Static Rigid vs. Flexible Nonrigid

In our approach, the original data tensor $\mathcal{X}$ is represented as sum of rigid $\mathcal{X}_0$ and nonrigid $\Delta\mathcal{X}$ components. Assuming that the rigid part does not vary among persons or expressions, changing the parameter vectors of the tensor model (14) should not change it. Therefore, we synthesise the rigid and nonrigid 2D representations of the basis emotions by varying $\mathbf{u}_5^{\mathrm{T}}$ as one row of $\mathbf{U}^{(5)}$, referred to as $\widehat{\mathbf{f}}_0(\mathbf{u}_{\text{emotion}})$, or $\widehat{\mathbf{f}}(\mathbf{u}_{\text{emotion}})$, while $\mathbf{u}_k^{\mathrm{T}}$, $k = 3, 4$ are the row-wise mean of $\mathbf{U}^{(k)}$, i.e., average rotation or person. The resulting 2D faces, shown in Fig. 4(a)-(g), can be projected to their corresponding 3D representation $\widehat{\mathbf{f}}^{3\mathrm{D}}$ illustrated in Fig. 4(h)-(n), see Sec. 5. Here different heights in 2D stem from 2D affine projections. As expected, varying the emotion does not change the rigid part (see Supplementary Material), which equals to the rigid basis shape, shown in column one. Please note that neither the 2D faces nor the 3D faces in Fig. 4 relate to actual training samples.

For quantitative evaluation, we approximate the original shapes $\mathbf{f}_i \subset \mathcal{X}$, by their known parameter vectors $\mathbf{u}_k$ (14), and compute the distance between true and estimated shapes. We repeat the experiment with varying percentages of explained variances, i.e., cropping factors of subspaces. Fig. 3 shows that the error based on the nonrigid part is always below the error of the rigid part, that is, $\frac{1}{I}\sum_{i=1}^{I}\|\mathbf{f}_i - \widehat{\mathbf{f}}_i\|_2 < \frac{1}{I}\sum_{i=1}^{I}\|\mathbf{f}_i - \widehat{\mathbf{f}}_{0,i}\|_2$.

Table 1: 3D MSE error based on (19) of different methods.

	BU3DFE-85	BU3DFE-7k
PI [17]	*	*
BMM [17]	*	*
K&L [28]	0.1666	0.3521
ISA [10]	0.0290	0.0263
BPCA	**0.0098**	0.0090
BICA	0.0157	**0.0088**
BPCA+QA	0.0286	0.0140
BICA+QA	0.0340	0.0130

* no result within 5 days

Table 2: Relative reprojection error, reported as inverse SNR (20), for different NRSFM methods.

	LS3D-W	BU3DFE-68	BU3DFE-85	BU3DFE-7k
PI [17]	0.00126	0.00231	*	*
BMM [17]	0.00100	0.00231	*	*
K&L [28]	0.00013	0.00141	0.00214	0.00121
ISA [10]	0.00016	0.00130	0.00216	0.00120
BPCA	0.00012	**0.00041**	**0.00121**	**0.00066**
BICA	**0.00011**	0.00062	0.00134	0.00072

* no result within 5 days

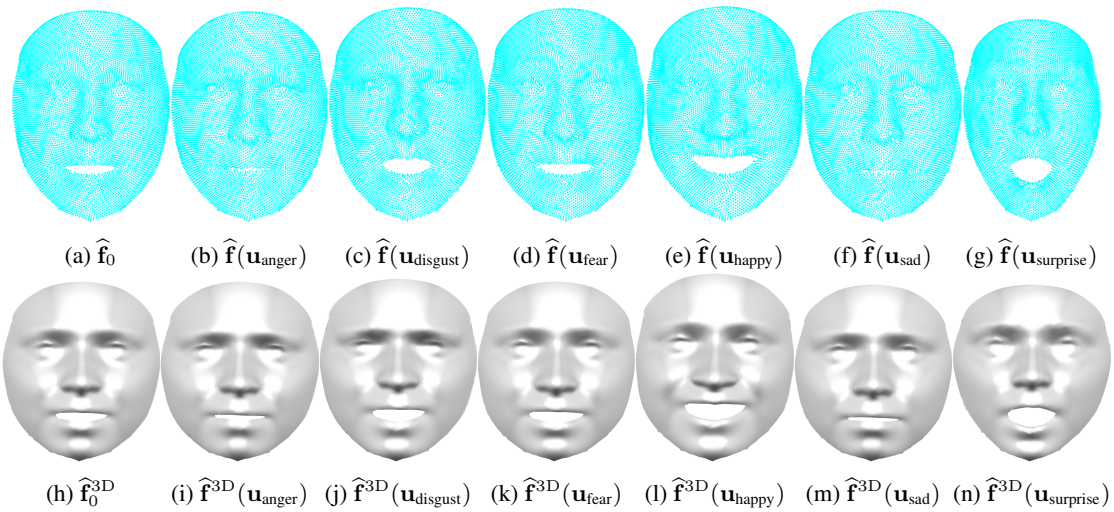
7.3. 3D Reconstruction

In this section, we report the 3D reconstruction results for LS3D-W Balanced, see Sec. 6.1, and the three datasets based on BU3DFE, see Sec. 6.2, with different methods. Specifically, our method factorises the input into a motion matrix and 3D rank-one basis shapes, as illustrated for the dense dataset BU3DFE-7k in Fig. 6.

All the methods provide 3D estimates $\widehat{\mathbf{Z}} \in \mathbb{R}^{3I\times N}$ for the 2D input shapes $\mathbf{X} \in \mathbb{R}^{2I\times N}$, which can be compared to normalised ground truth (GT) 3D shapes $\mathbf{Z}$, if available. Dai et al.'s and Kong and Lucey's methods yield the result up to an unknown similarity transform, as to the GT, while ISA and our methods yield the result up to an unknown affine transform. Thus we report the 3D error between the aligned 3D shapes $\mathbf{z}^i \in \mathbf{Z}_{\text{align}}$, and $\widehat{\mathbf{z}}^i \in \widehat{\mathbf{Z}}_{\text{align}}$, defined as

$$\mathrm{MSE}_{3\mathrm{D}} = \frac{1}{3NI}\sum_{i=1}^{I}\|\mathbf{z}^i - \widehat{\mathbf{z}}^i\|_{\mathrm{Fro}}^2. \qquad (19)$$

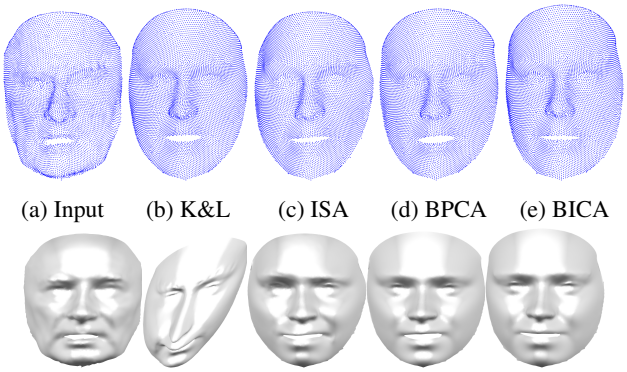
We also evaluate our affine reconstruction results upgraded to metric by Quan's affine autocalibration (QA) method [32], and thereafter registered by Procrustes alignment. The results are collected in Tab. 1.



(a) $\widehat{\mathbf{f}}_0$ (b) $\widehat{\mathbf{f}}(\mathbf{u}_{\text{anger}})$ (c) $\widehat{\mathbf{f}}(\mathbf{u}_{\text{disgust}})$ (d) $\widehat{\mathbf{f}}(\mathbf{u}_{\text{fear}})$ (e) $\widehat{\mathbf{f}}(\mathbf{u}_{\text{happy}})$ (f) $\widehat{\mathbf{f}}(\mathbf{u}_{\text{sad}})$ (g) $\widehat{\mathbf{f}}(\mathbf{u}_{\text{surprise}})$

(h) $\widehat{\mathbf{f}}_0^{\text{3D}}$ (i) $\widehat{\mathbf{f}}^{\text{3D}}(\mathbf{u}_{\text{anger}})$ (j) $\widehat{\mathbf{f}}^{\text{3D}}(\mathbf{u}_{\text{disgust}})$ (k) $\widehat{\mathbf{f}}^{\text{3D}}(\mathbf{u}_{\text{fear}})$ (l) $\widehat{\mathbf{f}}^{\text{3D}}(\mathbf{u}_{\text{happy}})$ (m) $\widehat{\mathbf{f}}^{\text{3D}}(\mathbf{u}_{\text{sad}})$ (n) $\widehat{\mathbf{f}}^{\text{3D}}(\mathbf{u}_{\text{surprise}})$

Figure 4: The six prototypical emotions, synthesised by the tensor model (14) for the average rotation, and average person, with varying expression $\mathbf{u}_5$. First row: 2D, Second row: 3D. (a) shows the rigid 2D shape $\widehat{\mathbf{f}}_0$ (16), which looks the same with varying emotions (see Supplementary Material), (b)-(g) 2D shapes with the nonrigid part $\widehat{\mathbf{f}} = \widehat{\mathbf{f}}_0 + \Delta\widehat{\mathbf{f}}$ (17). The synthesised 3D shapes (see Sec. 5) are shown in (h) for rigid, and (i)-(n) with the nonrigid part. Please note that none of these faces has a corresponding 2D face in the training data, and all of them have been created solely from 2D points.



(a) Input (b) K&L (c) ISA (d) BPCA (e) BICA

(f) GT (g) K&L (h) ISA (i) BPCA (j) BICA

Figure 5: Selected examples of the 3D reconstruction from (a) 2D input with known (f) 3D ground truth 3D. Estimates in 2D and 3D are presented for the methods: (b), (g) K&L, (c), (h) ISA, (d), (i) BPCA, and (e), (j) BICA.

Additionally, we compute the distance between the 2D input $\mathbf{x}^i \in \mathbf{X}$ and reprojected estimated 3D reconstruction $\widehat{\mathbf{x}}^i \in \widehat{\mathbf{X}}$. We use the *relative reprojection error* defined by the Inverse Signal to Noise Ratio (iSNR) [10] as

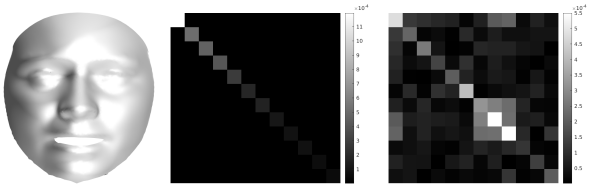
$$\text{iSNR} = \frac{\|\epsilon - \overline{\epsilon}\|_{\text{Fro}}^2}{\|\mathbf{X} - \overline{\mathbf{X}}\|_{\text{Fro}}^2}, \tag{20}$$

where $\epsilon = \widehat{\mathbf{X}} - \mathbf{X}$, and $\overline{\mathbf{X}}$ refers to the mean. The resulting mean iSNR are displayed in Tab. 2. The affine autocalibration does not affect the reprojection error, hence is not repeatedly reported.

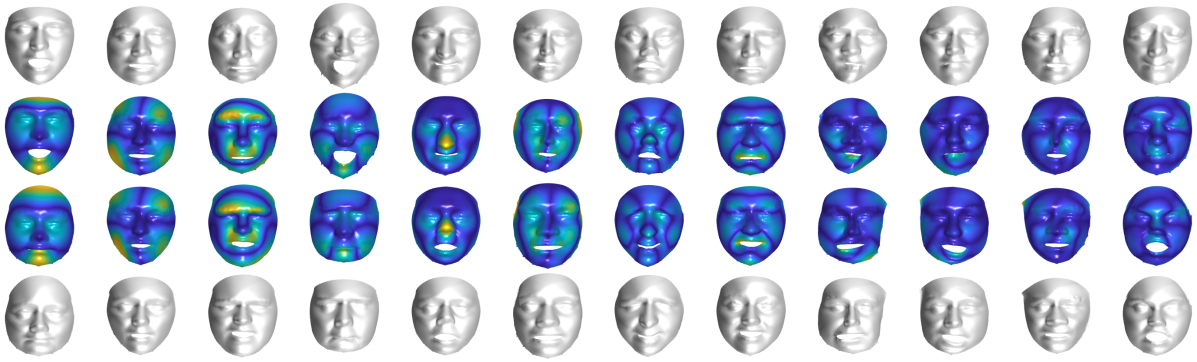
In general, it can be seen that all our proposed methods BPCA, BICA, BPCA+QA, and BICA+QA lead to satisfactory results in both low 2D and 3D errors. The variants with the metric upgrade yield slightly lower score, when compared to the affine reconstructions, due to the inevitable autocalibration error. While our methods finish in moderate running time, the methods BMM and PI tend to take several days, even on sparse datasets. Therefore, we did not evaluate them on two of the four datasets. The method [28] performs similarly as well as our approaches in terms of time, but yields moderately higher errors in 2D, and clearly higher 3D errors, see Tab. 1. The 3D errors based on ISA [10] are similar to our methods. The quantitative findings are supported by qualitative evaluation (Fig. 5) which shows that our dense 3D shape reconstructions by BPCA and BICA, match the GT shape equally well as ISA, and substantially better than K&L [28], which yields relatively flat 3D shapes (see also Supplementary Material).

8. Conclusion

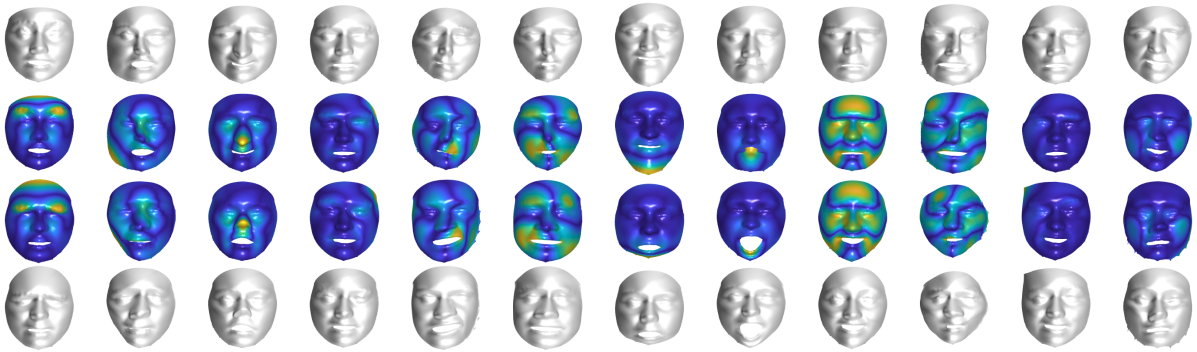
In this work, we combined a tensor model, similar to [23], with a matrix-based factorisation addressing the non-rigid structure-from-motion problem. By construction, the tensor model naturally unfolds to the NRSFM measurement matrix which sets the starting point for our method. We then showed how the non-rigid structure-from-motion problem can be solved by introducing rank-one basis shapes, which simply are 3D back-projections of the principal com-



(a) $\mathbf{B}_0$ (b) BPCA covariance matrix (c) BICA covariance matrix



(d) Rank-one basis shapes of the BPCA method, where the kth column represents deviations from the basis shape computed by $\mathbf{B}_0 \pm \omega_k \hat{\mathbf{B}}_k$, $\omega_k \in \mathbb{R}$.



(e) Rank-one basis shapes of the BICA method, where the kth column represents deviations from the basis shape computed by $\mathbf{B}_0 \pm \omega_k \hat{\mathbf{B}}_k$, $\omega_k \in \mathbb{R}$.

Figure 6: Illustration of the 12 rank-one basis shape $\mathbf{B}_0 \pm \omega_k \hat{\mathbf{B}}_k$ retrieved by our proposed methods. (a) shows the identical rigid basis shapes $\mathbf{B}_0$ for BPCA, and BICA. The absolute values of the covariance matrices of the nonrigid basis shapes are shown in (b) for BPCA, and (c) for BICA. The synthesised 3D basis shapes are shown in (d) for BPCA, and (e) for BICA. Each column represents the deviation from the basis shape based on the kth rank-one basis shape. Each shape is displayed in grey and with colour-coded distance to the basis shape. (Dark blue is zero distance, yellow represents a high distance.)

ponents of the measurement matrix, or alternatively, back-projections of its statistically independent components. In contrast to almost all the earlier methods in NRSFM, no cumbersome orthogonal constraints are required with our method since it is based on *stratified approach*. That is the reconstruction is first found up to an unknown affine transform after which it can be converted to metric by using the camera calibration information. The experiments showed that our approach suits well for dense correspondences and is better than the state-of-the-art methods in reconstruction error and computational efficiency, both in 2D and 3D. The other modes of the tensor provide an intuitive folding into the semantically meaningful subspaces. This facilitates the creation of new identities or expressions by editing the model parameters. Even though we build our method around the tensor model solely on 2D data, the affine, rank-one 3D basis shape formulation of the problem is simple and efficient, and renders as a promising tool for further development of dense NRSFM methods. In the future, we hence plan to incorporate the proposed tensor formulation into a neural network architecture.

References

[1] Victoria Fernandez Abrevaya, Stefanie Wuhrer, and Edmond Boyer. Multilinear Autoencoder for 3D Face Model Learning. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9, Mar. 2018.

[2] I. Akhter, Y. Sheikh, and S. Khan. In defense of orthonormality constraints for nonrigid structure from motion. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1534–1541, June 2009. ISSN: 1063-6919.

[3] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L. Morency. OpenFace 2.0: Facial Behavior Analysis Toolkit. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 59–66, May 2018.

[4] A. Bartoli, V. Gay-Bellile, U. Castellani, J. Peyras, S. Olsen, and P. Sayd. Coarse-to-fine low-rank structure-from-motion. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2008. ISSN: 1063-6919.

[5] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, SIGGRAPH '99, pages 187–194, New York, NY, USA, 1999. ACM Press/Addison-Wesley Publishing Co.

[6] Timo Bolkart and Stefanie Wuhrer. A Groupwise Multilinear Correspondence Optimization for 3D Faces. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3604–3612, Dec. 2015. ISSN: 2380-7504.

[7] Timo Bolkart and Stefanie Wuhrer. A Robust Multilinear Model Learning Framework for 3D Faces. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4911–4919, June 2016. ISSN: 1063-6919.

[8] W. Brand. Morphable 3D models from video. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 2, pages II–II, Dec. 2001. ISSN: 1063-6919.

[9] S.S. Brandt, P. Koskenkorva, Juho Kannala, and A. Heyden. Uncalibrated non-rigid factorisation with automatic shape basis selection. In *2009 IEEE 12th International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 352–359, Sept. 2009.

[10] S. S. Brandt, H. Ackermann, and S. Grasshof. Uncalibrated Non-Rigid Factorisation by Independent Subspace Analysis. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 569–578, Oct. 2019. ISSN: 2473-9944.

[11] C. Bregler, A. Hertzmann, and H. Biermann. Recovering non-rigid 3D shape from image streams. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No.PR00662)*, volume 2, pages 690–696 vol.2, 2000.

[12] Alan Brunton, Timo Bolkart, and Stefanie Wuhrer. Multilinear Wavelets: A Statistical Shape Space for Human Faces. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, Lecture Notes in Computer Science, pages 297–312, Cham, 2014. Springer International Publishing.

[13] A. Del Bue, X. Llad, and L. Agapito. Non-Rigid Metric Shape and Motion Recovery from Uncalibrated Images Using Priors. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 1191–1198, June 2006. ISSN: 1063-6919.

[14] A. Del Bue, J. Xavier, L. Agapito, and M. Paladini. Bilinear Modeling via Augmented Lagrange Multipliers (BALM). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(8):1496–1508, Aug. 2012.

[15] A. Bulat and G. Tzimiropoulos. How Far are We from Solving the 2D 3D Face Alignment Problem? (and a Dataset of 230,000 3D Facial Landmarks). In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1021–1030, Oct. 2017. ISSN: 2380-7504.

[16] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. FaceWarehouse: A 3D Facial Expression Database for Visual Computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, Mar. 2014.

[17] Y. Dai, H. Li, and M. He. A simple prior-free method for non-rigid structure-from-motion factorization. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2018–2025, June 2012. ISSN: 1063-6919.

[18] Lieven De Lathauwer and Bart De Moor. A multi-linear singular value decomposition. *Society for Industrial and Applied Mathematics*, 21:1253–1278, 03 2000.

[19] Alessio Del Bue and Lourdes Agapito. Non-rigid 3D shape recovery using stereo factorization. *Asian Conference of Computer Vision*, 1:25–30, Jan. 2004.

[20] T. Gerig, Andreas Forster, Clemens Blumer, B. Egger, M. Lüthi, Sandro Schönborn, and T. Vetter. Morphable Face Models - An Open Framework. *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 2018.

[21] V. Golyanik, B. Taetz, G. Reis, and D. Stricker. Extended coherent point drift algorithm with correspondence priors and optimal subsampling. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9, Mar. 2016.

[22] P. F. U. Gotardo and A. M. Martinez. Computing Smooth Time Trajectories for Camera and Deformable Shape in Structure from Motion with Occlusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10):2051–2065, Oct. 2011.

[23] Stella Grasshof, Hanno Ackermann, Sami Sebastian Brandt, and Jörn Ostermann. Multilinear Modelling of Faces and Expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3540–3554, Oct. 2021. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[24] Stella Graßhof, Hanno Ackermann, Sami S. Brandt, and Jörn Ostermann. Apathy Is the Root of All Expressions. In *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, pages 658–665, May 2017.

[25] Stella Graßhof, Hanno Ackermann, Felix Kuhnke, Jörn Ostermann, and Sami S. Brandt. Projective structure from facial motion. In *2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA)*, pages 298–301, May 2017.

[26] Richard Hartley and René Vidal. Perspective Nonrigid Shape and Motion Recovery. In David Forsyth, Philip Torr, and

Andrew Zisserman, editors, *Computer Vision – ECCV 2008*, Lecture Notes in Computer Science, pages 276–289, Berlin, Heidelberg, 2008. Springer.

[27] A. Hyvärinen and E. Oja. A Fast Fixed-Point Algorithm for Independent Component Analysis. *Neural Computation*, 9(7):1483–1492, July 1997. Conference Name: Neural Computation.

[28] C. Kong and S. Lucey. Prior-Less Compressible Structure from Motion. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4123–4131, June 2016. ISSN: 1063-6919.

[29] Chen Kong and Simon Lucey. Deep Non-Rigid Structure From Motion. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1558–1567, Oct. 2019. ISSN: 2380-7504.

[30] Suryansh Kumar, Luc Van Gool, Carlos E. P. de Oliveira, Anoop Cherian, Yuchao Dai, and Hongdong Li. Dense Non-Rigid Structure from Motion: A Manifold Viewpoint. *arXiv:2006.09197 [cs]*, June 2020. arXiv: 2006.09197.

[31] Marco Paladini, Alessio Del Bue, João Xavier, Lourdes Agapito, Marko Stošić, and Marija Dodig. Optimal Metric Projections for Deformable and Articulated Structure-from-Motion. *International Journal of Computer Vision*, 96(2):252–276, Jan. 2012.

[32] Long Quan. Self-calibration of an affine camera from multiple views. *International Journal of Computer Vision*, 19(1):93–105, July 1996.

[33] Vikramjit Sidhu, Edgar Tretschk, Vladislav Golyanik, Antonio Agudo, and Christian Theobalt. Neural dense non-rigid structure from motion with latent space constraints. In *European Conference on Computer Vision (ECCV)*, 2020.

[34] Carlo Tomasi and Takeo Kanade. Shape and motion from image streams under orthography: a factorization method. *International Journal of Computer Vision*, 9(2):137–154, Nov. 1992.

[35] L. Torresani, A. Hertzmann, and C. Bregler. Nonrigid Structure-from-Motion: Estimating Shape and Motion with Hierarchical Priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(5):878–892, May 2008.

[36] L. Torresani, D. B. Yang, E. J. Alexander, and C. Bregler. Tracking and modeling non-rigid objects with rank constraints. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I, Dec. 2001. ISSN: 1063-6919.

[37] M. Alex O. Vasilescu and Demetri Terzopoulos. Multilinear Analysis of Image Ensembles: TensorFaces. In Anders Heyden, Gunnar Sparr, Mads Nielsen, and Peter Johansen, editors, *Computer Vision — ECCV 2002*, number 2350 in Lecture Notes in Computer Science, pages 447–460. Springer Berlin Heidelberg, 2002.

[38] Daniel Vlasic, Matthew Brand, Hanspeter Pfister, and Jovan Popović. Face transfer with multilinear models. *ACM Transactions on Graphics*, 24(3):426–433, July 2005.

[39] Mengjiao Wang, Yannis Panagakis, Patrick Snape, and Stefanos Zafeiriou. Learning the Multilinear Structure of Visual Data. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6053–6061, July 2017. ISSN: 1063-6919.

[40] Mengjiao Wang, Zhixin Shu, Shiyang Cheng, Yannis Panagakis, Dimitris Samaras, and Stefanos Zafeiriou. An Adversarial Neuro-Tensorial Approach For Learning Disentangled Representations. *International Journal of Computer Vision*, 127(6-7):743–762, June 2019. arXiv: 1711.10402.

[41] Jing Xiao, Jin-Xiang Chai, and Takeo Kanade. A Closed-Form Solution to Non-rigid Shape and Motion Recovery. In Tomás Pajdla and Jiří Matas, editors, *Computer Vision - ECCV 2004*, Lecture Notes in Computer Science, pages 573–587, Berlin, Heidelberg, 2004. Springer.

[42] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and M.J. Rosato. A 3D facial expression database for facial behavior research. In *7th International Conference on Automatic Face and Gesture Recognition, 2006. FGR 2006*, pages 211–216, Apr. 2006.