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 up to 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 multifo-cal 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.

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 \cite{5} \cite{20}. As a generalisation of the matrix-based SVD the Higher-Order SVD (HOSVD) was introduced in \cite{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 \cite{37,38,16,24,25,23}. However, in most works the HOSVD is employed on 3D faces \cite{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 \cite{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 \cite{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 \cite{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 and back-project them onto 3D to create rank-one shapes\footnote{Note that this is different from \cite{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.}.

An overview of our approach is illustrated in Fig. 1 for 3D reconstruction and shape editing.

Our contributions are summarised as follows.

- 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 \cite{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.

2. Tensor Representation

Let \(X \in \mathbb{R}^{N \times D \times F \times P \times E}\), \(D = 2\), 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 \(X\) have been centered so that the each 2D face has the mean coordinate. The tensor \(X\) is then decomposed by the Higher-Order-SVD (HOSVD) \cite{18} as

\[
X \approx \hat{X} = S \times_1 U_1^{(1)} \times_2 U_2^{(2)} \times_3 U_3^{(3)} \times_4 U_4^{(4)} \times_5 U_5^{(5)},
\]

(1)

where \(\times_d\) is the \(d\)-way product, \(S \in \mathbb{R}^{N \times D \times F \times P \times E}\) is the core tensor, and \(U_d^{(d)} \in \mathbb{R}^{d \times d}\), \(d = 1, 2, \ldots, 5\); are semi-orthogonal matrices which consist of the singular vectors corresponding to the \(d\)-mode unfolded tensor, with \(\hat{d} \leq d\) representing the number of retained elements of the dimension \(d\), i.e., the smallest singular values and vectors have been truncated. The tensor \(X = X_0 + \Delta X\) is divided into rigid \(X_0\) and non-rigid \(\Delta X\) parts by separating the three largest mode-1 singular values and vectors from the remaining ones, respectively. The rigid part is thus obtained as

\[
X_0 \approx \hat{X}_0 = S_0 \times_1 U_0^{(1)} \times_2 U_2^{(2)} \times_3 U_3^{(3)} \times_4 U_4^{(4)} \times_5 U_5^{(5)},
\]

(2)

where \(S_0 \in 3 \times D \times F \times P \times E\) is the core tensor, and the \(N \times 3\) semi-orthogonal matrix \(U_0^{(1)}\) contain the first three 1-mode singular values and vectors. The non-rigid part \(\Delta X\)
is formed as
$$\Delta \mathcal{X} \approx \Delta \hat{\mathcal{X}} = S_1 \times_1 U_1^{(1)} \times_2 U_2^{(2)} \times_3 U_3^{(3)} \times_4 U_4^{(4)} \times_5 U_5^{(5)},$$
where $S_1 \in \mathbb{R}^{(N-3) \times D \times F \times P \times E}$ is the core tensor, and $U_1^{(1)} \in \mathbb{R}^{N \times (N-3)}$ is the first three left singular vectors of $\mathcal{X}^{(1)}$ representing the least squares estimate for the affine 3D structure $B_0$, obtained by the singular value decomposition as
$$\hat{\mathcal{X}}_0^{(1)T} = \frac{1}{\sqrt{N}} V_0^{(1)} \Sigma_0^{(1)} (\sqrt{N} U_0^{(1)T}) = M_0 B_0,$$
where the $2I \times 3$ matrix $M_0 = (M^1, M^2, \ldots, M^I)$ contains estimates for all the $2 \times 3$ inhomogeneous affine projection matrices for all the $I = FPE$ views. In other words, the non-rigid variation in the decomposition 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 $z^n_i = b_{0n} + \Delta z^n_i$, where $b_{0n}$ is the rigid, mean shape and $\Delta z^n_i$ is the non-rigid component. The other modal components in the rigid approximation $X_0$, 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 $\Delta z^n_i = \sum_{k=1}^{N-3} \alpha_k b_{kn}$, where $\alpha_k$ is a scalar and $\mathbf{rank}(B_k) = 1$ for $k \neq 0$, where $B_k = (b_{k1}, b_{k2}, \ldots, b_{kN}) \in \mathbb{R}^{3 \times N}$. Assuming that all the structure components share the same projection matrix on to a fixed image, the non-rigid 3D component $\Delta z^n_i$ maps to the non-rigid 2D parts $\Delta x^n_i$ stored in $\Delta \mathcal{X}$, where
$$\Delta x^n_i = M^i \Delta z^n_i = M^i \left( \sum_{k=1}^{N-3} \alpha_k b_{kn} \right).$$
The mode-1 unfolding of the tensor $\Delta \mathcal{X}$ thus factorises into the a weighted sum of $N-3$ 3D rank-one basis shapes $B_k$, with $\alpha_k \in \mathbb{R}$, and $I$ projection matrices $M^i \in \mathbb{R}^{2 \times 3}$, as
$$\Delta \mathcal{X}^{(1)T} \approx \begin{pmatrix} \alpha_1 M^1 & \alpha_2 M^2 & \ldots & \alpha_{N-3} M^{N-3} \\ \alpha_1^T M^2 & \alpha_2^T M^2 & \ldots & \alpha_{N-3}^T M^{N-3} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_1^T M^{N-3} & \alpha_2^T M^{N-3} & \ldots & \alpha_{N-3}^T M^{N-3} \end{pmatrix} \begin{pmatrix} B_1 \\ B_2 \\ \vdots \\ B_{N-3} \end{pmatrix} = \begin{pmatrix} B \end{pmatrix}_{(N-9) \times N}.$$ 

Clearly, the basis matrix $B$, and hence $\Delta \mathcal{X}^{(1)T}$, has the rank $N - 3$. Additionally, the motion matrix $M$ 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
$$\delta \hat{\mathcal{X}}^{(1)T} = \frac{1}{\sqrt{N}} V^{(1)} \Sigma^{(1)} (\sqrt{N} U^{(1)T}) = MB,$$
where $\hat{B} = AB$ and $\hat{M} = MA^T$, and $\hat{M}$ and $\hat{B}$ are the estimates matching the form $[9]$. In the following two subsections, we discuss two different approaches for finding $A$.

3.3. Principal Component Analysis (PCA)

Let $C^{(1)} = C_0^{(1)} + \Delta C^{(1)}$ denote the mode-1 covariance equivalent to
$$C^{(1)} = \mathcal{X}^{(1)T} \mathcal{X}^{(1)} = U^{(1)} \Sigma_0^{(1)2} U^{(1)T} + \Delta C^{(1)} = U^{(1)} \Sigma_0^{(1)2} U^{(1)T} + U^{(1)} \Sigma_1^{(1)2} U^{(1)T},$$
where $C_0^{(1)} = U_0^{(1)} \Sigma_0^{(1)2} U_0^{(1)T}$ and $\Delta C^{(1)} = U_1^{(1)} \Sigma_1^{(1)2} U_1^{(1)T}$. We see that the mode-1 left singular vectors, or the rows in $B_0$ and $B$, 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, $\Delta C^{(1)}$ contain all the principal components, except the first three since the mean rigid shape has been factored to $C_0^{(1)}$. Hence, $B$ has $N - 3$ 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 $B_k = d_k b_k^T$, where $b_k^T$ denotes the row $k$ of $B$, and $d_k$, is the $3 \times 1$ 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 \( \mathbf{b}_k^T \) 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 \((3N - 9) \times (N - 3)\) block diagonal matrix with \( 3 \times 1 \) blocks \( \mathbf{d}_k \).

### 3.4. Independent Component Analysis (ICA)

More generally, the operator \( \mathbf{A} \) can be written in the form \( \mathbf{A} = \mathbf{D} \mathbf{G} \), where \( \mathbf{G} \) is a \((N - 3) \times (N - 3)\) 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

\[
I(\mathbf{Z}) = \sum_j H(\mathbf{Z}_j) - H(\mathbf{Z}),
\]

where \( H \) refers to differential entropy and \( \mathbf{Y} = \mathbf{A}\text{ICA}\mathbf{Z} \) 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 \( \mathbf{A}\text{ICA} = \mathbf{G}^T \) will be an orthogonal matrix, hence,

\[
\mathbf{B}\text{ICA} \equiv \mathbf{A}_{\text{ICA}}^T \mathbf{B} = \mathbf{G} \mathbf{B},
\]

where the rows of \( \mathbf{B}\text{ICA} \) 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 \( \mathbf{B}\text{PCA} \equiv \mathbf{B} \) or \( \mathbf{B}\text{ICA} \), depending whether the PCA or ICA model is selected, respectively. We are searching for the minimiser to the energy functional

\[
E(\mathbf{d}, \alpha) = \sum_i ||\Delta \mathbf{X}(1)^i - \Sigma_k \alpha_k \mathbf{B}_k^i||^2_{\text{Fro}},
\]

subject to \( ||\mathbf{d}_k||_2 = 1 \), for all \( k \), where \( \mathbf{B}_k^i = \mathbf{M}^i \mathbf{d}_k \mathbf{b}_k^T \) are the rank-one operators referring to the rank-one basis shapes, and \( \alpha_k \) 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 \( \mathbf{X} \), as in (6).
2. Initialise the parameters \( \mathbf{d} \), \( \alpha \).
3. Decompose \( \mathbf{X} \) into the rigid \( \mathbf{X}_0 \) and non-rigid \( \Delta \mathbf{X} \) part as in (2) and (3), respectively.
4. Factorise the non-rigid part as \( \Delta \mathbf{X}(1)^i \approx \mathbf{M} \mathbf{V}(1)^i \mathbf{U}(1)^i \), where \( \mathbf{M} = \frac{1}{\sqrt{N}} \mathbf{V}(1)^i \mathbf{\Sigma}(1)^i \mathbf{B} = \sqrt{N} \mathbf{U}(1)^i \).
5. Do either
   
   (a) Compute the PCA basis by assuming \( \mathbf{G} = \mathbf{I} \) and so that \( \mathbf{B}_{\text{PCA}} = \mathbf{B} \) or
   
   (b) Find the orthogonal transformation \( \mathbf{G} \) and ICA basis by FastICA [27] so that \( \mathbf{B}_{\text{ICA}} = \mathbf{G} \mathbf{B} \).
6. Form the rank-one basis shapes \( \mathbf{B}_k^i = \mathbf{M}^i \mathbf{d}_k \mathbf{b}_k^T, i = 1, 2, \ldots, I \), where \( \mathbf{b}_k^T \) is the \( k \)th row of \( \mathbf{B}_{\text{PCA}} \) or \( \mathbf{B}_{\text{ICA}} \), \( k = 1, 2, \ldots, K \).
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

\[
\langle \mathbf{B}_k^i, \mathbf{B}_{k'}^{i'} \rangle = \langle \text{vec} \{ \mathbf{B}_k^i \}, \text{vec} \{ \mathbf{B}_{k'}^{i'} \} \rangle = \langle \mathbf{M}^i \mathbf{d}_k, \mathbf{M}^{i'} \mathbf{d}_{k'} \rangle \langle \mathbf{b}_k, \mathbf{b}_{k'} \rangle,
\]

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 \( \mathbf{f} \in \mathbb{R}^{N \times D} \) can be represented using (1) as

\[
\mathbf{f} = \mathbf{S} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)}
\]

\[
\times_3 \mathbf{P}_3^T \mathbf{U}^{(3)} \times_4 \mathbf{P}_4^T \mathbf{U}^{(4)} \times_5 \mathbf{P}_5^T \mathbf{U}^{(5)},
\]

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 \( \hat{\mathbf{f}} \) by minimising

\[
\min_{\mathbf{P}_k} ||\mathbf{f} - \hat{\mathbf{f}}||^2 + \sum_{k=3}^5 \lambda_k ||\mathbf{p}_k||^2 + \lambda_{k,s} (\mathbf{p}_k^T \mathbf{1} - 1)^2,
\]
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 \cite{24,23}.

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

$$ \tilde{f}_0 = S_0 \times_1 U_0^{(1)} \times_2 U^{(2)} \times_3 u_3^T \times_4 u_4^T \times_5 u_5^T, \quad (16) $$

and the nonrigid part

$$ \Delta \hat{f} = S_1 \times_1 U_1^{(1)} \times_2 U^{(2)} \times_3 u_3^T \times_4 u_4^T \times_5 u_5^T, \quad (17) $$

and $u_k^T = p_k^T U^{(k)}, k = 3, 4, 5$.

5. 3D Shape Synthesis and Editing

So far we have presented two factorisation approaches: a matrix-based (Sec. 5), 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 \cite{14} to create a new 2D shape by either: (1) choosing the parameter vectors $p_k$ freely, or (2) estimate them to approximate an arbitrary 2D shape $x'$ by solving \cite{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 $x'$.

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

$$ \hat{z}' = B_0 + \sum_k \alpha_k' d_k b_k^T, \quad (18) $$

The basis coefficients $\alpha_k'$ are computed by orthogonal projection in step 7 of Alg. \cite{1} and employs the estimated projection matrix $M_0$ of the sample $i$, which is unknown for a new shape $\hat{x}'$, but can be estimated by the affine camera resection algorithm as $M_0$. The basis coefficients $\alpha_k'$ are then estimated that yield the 3D estimate $\hat{z}'$ corresponding to the new shape $\hat{x}'$ as $\hat{z}' = B_0 + \sum_k \alpha_k' d_k b_k^T$.

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. 7.

6. Databases

6.1. LS3D-W Balanced

The Large Scale 3D Faces in-the-Wild dataset (LS3D-W) \cite{13} is a facial landmark dataset, which contains ca. 230,000 images, each annotated with 68 2D points. \cite{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) \cite{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 \cite{21}. Additionally, \cite{13} 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 \cite{3}.
- BU3DFE-85: the 85 sparse 3D landmarks rotated by 3 yaw angles $\alpha_y \in \{-\frac{\pi}{8}, 0, \frac{\pi}{8}\}$, projected to 2D.
- BU3DFE-7k: the estimated 7308 3D points rotated by 3 yaw angles $\alpha_y \in \{-\frac{\pi}{8}, 0, \frac{\pi}{8}\}$, projected to 2D.

7. Experiments

We compare the two proposed variants of our approach to Dai et al.’s pseudoinverse (PI) \cite{17}, and Block matrix Method (BMM) \cite{17}, and Kong and Lucey’s Priorless decomposition (K&L) \cite{28}, and Brandt et al.’s ISA \cite{10}. For the factorisation we selected $N = 15$ 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 \cite{24}, based on the BU3DFE database \cite{42}, revealed a planar star-shaped substructure in the expression subspace. and a similar structure on the basis of 2D database was found in \cite{23}. To complement these findings, we investigated 2D data based on the BU3DFE dataset \cite{42}, described in Sec. 6.2. The resulting expression space $U^{(5)}$ from (1) reveals the same substructure for all of the three datasets, see Fig. 2.
7.2. Static Rigid vs. Flexible Nonrigid

In our approach, the original data tensor $X$ is represented as sum of rigid $X_0$ and nonrigid $\Delta X$ 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 $u^T_5$ as one row of $U^{(5)}$, referred to as $bf_0^{emotion}$, or $bf_4^{emotion}$, while $u^T_k$, $k=3,4$ are the row-wise mean of $U^{(k)}$, i.e., average rotation or person. The resulting 2D faces, shown in Fig. 4(a)-(g), can be projected to their corresponding 3D representation $bf^{3D}$ 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 $f_i \in X$, by their known parameter vectors $u_k \in (14)$, 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, $\frac{1}{T} \sum_{i=1}^{T} \|f_i - \hat{f}_i\|_2 < \frac{1}{T} \sum_{i=1}^{T} \|f_i - \hat{f}_{0,i}\|_2$.

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.

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

<table>
<thead>
<tr>
<th>Method</th>
<th>BU3DFE-85</th>
<th>BU3DFE-7k</th>
</tr>
</thead>
<tbody>
<tr>
<td>PI [17]</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>BMM [17]</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>K&amp;L [28]</td>
<td>0.1666</td>
<td>0.3521</td>
</tr>
<tr>
<td>ISA [10]</td>
<td>0.0290</td>
<td>0.0263</td>
</tr>
<tr>
<td>BPCA</td>
<td>0.0098</td>
<td>0.0090</td>
</tr>
<tr>
<td>BICA</td>
<td>0.0157</td>
<td>0.0088</td>
</tr>
<tr>
<td>BPCA+QA</td>
<td>0.0286</td>
<td>0.0140</td>
</tr>
<tr>
<td>BICA+QA</td>
<td>0.0340</td>
<td>0.0130</td>
</tr>
</tbody>
</table>

* no result within 5 days

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

<table>
<thead>
<tr>
<th>Method</th>
<th>LS3D-W</th>
<th>BU3DFE-85</th>
<th>BU3DFE-7k</th>
</tr>
</thead>
<tbody>
<tr>
<td>PI [17]</td>
<td>0.00126</td>
<td>0.00231</td>
<td>*</td>
</tr>
<tr>
<td>BMM [17]</td>
<td>0.00100</td>
<td>0.00231</td>
<td>*</td>
</tr>
<tr>
<td>K&amp;L [28]</td>
<td>0.00013</td>
<td>0.00141</td>
<td>0.00214</td>
</tr>
<tr>
<td>ISA [10]</td>
<td>0.00016</td>
<td>0.00130</td>
<td>0.00216</td>
</tr>
<tr>
<td>BPCA</td>
<td>0.00012</td>
<td>0.00041</td>
<td>0.00121</td>
</tr>
<tr>
<td>BICA</td>
<td>0.00011</td>
<td>0.00062</td>
<td>0.00134</td>
</tr>
</tbody>
</table>

* 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 $\tilde{Z} \in \mathbb{R}^{M \times N}$ for the 2D input shapes $X \in \mathbb{R}^{2T \times N}$, which can be compared to normalised ground truth (GT) 3D shapes $Z$, 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 $z^i \in Z_{\text{align}}$, and $\tilde{z}^i \in Z_{\text{align}}$, defined as

$$\text{MSE}_{3D} = \frac{1}{3N} \sum_{i=1}^{T} \|z^i - \tilde{z}^i\|^2_{\text{Fro}}.$$  (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.
Figure 4: The six prototypical emotions, synthesised by the tensor model \((14)\) for the average rotation, and average person, with varying expression \(u_\text{var}\). First row: 2D. Second row: 3D. (a) shows the rigid 2D shape \(\hat{f}_0\) \((16)\), which looks the same with varying emotions (see Supplementary Material), (b)-(g) 2D shapes with the nonrigid part \(\hat{f} = f_0 + \Delta f \) \((17)\). 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.

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 \(x^i \in X\) and reprojected estimated 3D reconstruction \(\hat{x}^i \in \hat{X}\). We use the relative reprojection error defined by the Inverse Signal to Noise Ratio (iSNR) \((10)\) as

\[
iSNR = \frac{\|\epsilon\|_{\text{Fro}}^2}{\|X - \bar{X}\|_{\text{Fro}}^2},
\]

where \(\epsilon = \hat{X} - X\), and \(\bar{X}\) 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-
Figure 6: Illustration of the 12 rank-one basis shape $B_0 \pm \omega_k B_k$ retrieved by our proposed methods. (a) shows the identical rigid basis shapes $B_0$ 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 $k$th 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.)
References


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


